Прикладная эконометрика, 2020, т. 57, с. 119–139.
Applied Econometrics, 2020, v. 57, pp. 119–139.
DOI: 10.22394/1993-7601-2020-57-119-139

Е. В. Коссова, Л. А. Куприянова, Б. С. Потанин[1]

Сравнение точности оценок параметрических и полупараметрических методов коррекции многомерного смещения отбора

В статье рассматриваются параметрические и полупараметрические методы коррекции смещения отбора и проводится их сопоставление в случае двумерного механизма отбора наблюдений. Сравнение осуществляется на симулированных данных. Исследуется точность оценок параметрических и полупараметрических методов при распределениях случайных ошибок, существенно отличающихся от нормального несимметричностью, наличием «тяжелых хвостов» или бимодальностью. Делается вывод о высокой точности параметрических методов даже при серьезном нарушении предположений о распределении случайных ошибок.

Ключевые слова: смещение отбора; распределение ошибок с тяжелыми хвостами; скошенное распределение ошибок; бимодальное распределение ошибок; полупараметрические модели.

JEL classification: C34.

Введение

Проблема неслучайного отбора является, наряду с эндогенностью, одной из основных причин смещения и несостоятельности оценок параметров эконометрических моделей. Выборки с неслучайным механизмом отбора присутствуют во многих экономических исследованиях, в частности, посвященных экономике труда и оцениванию эффективности государственных программ.

Джеймс Хекман был одним из первых, кто предложил эконометрическую модель, позволяющую получать состоятельные оценки параметров в условиях неслучайного отбора наблюдений. Несмотря на популярность модели Хекмана, его подход неоднократно подвергался критике из-за наличия предпосылки о совместном нормальном распределении случайных ошибок, нарушение которой может привести к несостоятельности оценок предложенного им метода (Vella, 1998).

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

Куприянова Любовь Александровна — Сколковский институт науки и технологии, Москва; Lyubov.Kupriyanova@skoltech.ru.

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

В последние десятилетия все большую популярность приобретают полупараметрические статистические методы, позволяющие существенно ослабить предположения о виде совместного распределения случайных компонентов модели и функциональной формы зависимости между изучаемыми показателями, при этом нужная информация извлекается из самих данных (Racine, 2008). Подобные подходы были разработаны и для моделей, оцениваемых по селективным выборкам (Vella, 1998; Pigini, 2015). Однако реализация полупараметрического подхода требует существенно больше вычислительных затрат, чем при параметрическом подходе, а интерпретация получаемых оценок часто оказывается затруднительной (Racine, 2008).

Неочевидность предпочтения той или иной методологии послужила основным источником мотивации данной статьи. При каких условиях полупараметрические методы для моделей с неслучайным отбором работают лучше, чем параметрические? Стоит ли применять, в основном, полупараметрические методы для моделей со смещением отбора, или же полученные с помощью параметрического подхода оценки коэффициентов оказываются достаточно близкими к их истинным значениям даже при отсутствии нормальности распределения случайных ошибок? В последнем случае параметрические методы по-прежнему можно считать актуальными, поскольку они сохраняют привлекательную ясность интерпретации и, как правило, требуют меньше вычислительных мощностей.

В данной статье представлено сравнение оценок, полученных параметрическими и наиболее популярными полупараметрическими методами, для модели неслучайного отбора с двумя уравнениями отбора. Сопоставление моделей проводится на симулированных данных при различных предположениях о совместном распределении случайных ошибок.

Поскольку полупараметрические методы ослабляют предположения о распределении случайных ошибок, основная гипотеза исследования заключается в том, что полупараметрические методы будут работать лучше параметрических, когда распределение стохастических компонент сильно отличается от нормального. Не менее важный вопрос исследования — хорошо ли работает параметрический подход в случае не очень существенного отклонения распределения случайных ошибок от нормального. Что понимается под существенным и несущественным отклонением от нормального распределения, обсуждается далее.

Актуальность данной статьи обусловлена отсутствием работ, сравнивающих параметрический и непараметрический подходы коррекции смещения отбора в случае многомерного механизма отбора. Единственный пример двух уравнений отбора рассматривался лишь на одном наборе реальных данных (De Luca, Peracchi, 2012).

Настоящее исследование, конечно, имеет свои ограничения. Из-за вычислительной сложности число уравнений отбора ограничивается двумя. Этого зачастую бывает достаточно для большинства прикладных задач, в то время как одного уравнения может не хватить для корректного рассмотрения проблемы (Das et al., 2003; Potanin, 2019). При этом предлагаемые в данной работе модификации полупараметрических моделей нетрудно обобщить на случай произвольного числа уравнений отбора.

Работа имеет следующую структуру. В первом разделе приводится формальная постановка регрессионной модели, отвечающей ситуации, когда наблюдения зависимой переменной, описываемые основным (целевым) уравнением, доступны лишь при выполнении двух правил, задаваемых уравнениями отбора. Во втором разделе рассматриваются наиболее популярные параметрические и полупараметрические методы оценивания регрессионных уравнений, учитывающие смещение, возникающее при неслучайном отборе наблюдений,

и предлагаются их обобщения на случай двух правил отбора. Раздел 3 описывает дизайн проводимого численного эксперимента, где используются распределения случайных ошибок, отличающиеся от нормального наличием тяжелых хвостов, несимметричностью и бимодальностью. Четвертый раздел посвящен сопоставлению оценок параметров основного регрессионного уравнения, полученных описанными в разделе 2 параметрическими и полупараметрическими методами, по данным из разных распределений.

1. Двумерное смещение отбора

Предположим, что значение целевой переменной y_i^* наблюдается только в том случае, когда выполнены два условия отбора, т. е.

$$z_{1i}^* = w_{1i}'\gamma_1 + u_{1i}, \tag{1}$$

$$z_{2i}^* = w_{2i}'\gamma_2 + u_{2i}, \tag{2}$$

$$y_i^* = x_i'\beta + \varepsilon_i, \tag{3}$$

$$w_{1i}', \gamma_1 \in R^{n_{\gamma_1}}, w_{2i}', \gamma_2 \in R^{n_{\gamma_2}}, x_i', \beta \in R^{n_\beta},$$

$$z_{ji} = \begin{cases} 1, \text{ если } z_{ji}^* \geq 0, \\ 0, \text{ в противном случае,} \end{cases} \quad y_i = \begin{cases} y_i^*, & \text{если } z_{1i} z_{2i} = 1, \\ \text{не наблюдаем,} & \text{в противном случае,} \end{cases}$$

$$j \in \{1,2\}, \; i \in \{1, \ldots, n\},$$

$$\left(u_{1i}, u_{2i}, \varepsilon_i\right) \sim \Theta(\theta), \; \theta \in R^{n_\theta},$$

где (1) и (2) являются уравнениями отбора, а (3) — основным (целевым) уравнением. Векторы случайных ошибок независимы между наблюдениями, не зависят от векторов объясняющих переменных w_{1i}', w_{2i}' и x_i', и подчиняются некоторому совместному распределению Θ с параметрами θ. Детерминированные векторы коэффициентов γ_1, γ_2 и β отражают влияние независимых переменных на зависимые z_{1i}^*, z_{2i}^* и y_i^*, являющиеся ненаблюдаемыми: доступна лишь информация о значениях z_{1i}, z_{2i} и y_i. Линейные комбинации $w_{1i}'\gamma_1$ и $w_{2i}'\gamma_1$ именуются линейными индексами.

Для пояснения экономического смысла рассматриваемого процесса генерации данных кратко опишем пример, представленный в (De Luca, Peracchi, 2012). В качестве z_{1i} и z_{2i} (z_{1i}^* и z_{2i}^*) выступали факты ответа (склонности к ответу) домохозяйства на вопросы о его доходах и расходах на питание соответственно. При этом y_i^* отражал долю расходов на питание, наблюдаемую в форме переменной y_i, лишь для тех домохозяйств, которые предоставили ответы на оба вопроса. Случайные ошибки уравнений отражали вклады несистематических погрешностей измерений и ненаблюдаемых характеристик, влияющих как на вероятность ответа на соответствующие вопросы, так и на долю расходов домохозяйства на питание. При этом, если одни и те же или связанные друг с другом ненаблюдаемые факторы влияют на разные уравнения, то соответствующие случайные ошибки могут оказаться

зависимыми. В таком случае, без явного учета неслучайного отбора, соответствующая зависимость может приводить к несостоятельности оценок β, полученных при помощи классических методов эконометрического анализа, включая метод наименьших квадратов (Коссова, Потанин, 2018).

Поэтому исследователи, заинтересованные в оценивании влияния дохода домохозяйства на долю его расходов на питание, использовали различные параметрические и полупараметрические методы коррекции смещения отбора, которые, вместе с рядом других, рассматриваются в следующем разделе работы.

2. Методы коррекции двумерного смещения отбора

2.1. Параметрический подход Хекмана

В рамках параметрического подхода накладывается строгое ограничение на совместное распределение случайных ошибок. Обычно, в том числе и в данном исследовании, при использовании параметрических методов предполагается, что случайные ошибки подчиняются многомерному нормальному закону:

$$\begin{bmatrix} u_{1i} \\ u_{2i} \\ \varepsilon_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_0 & \sigma\rho_1 \\ \rho_0 & 1 & \sigma\rho_2 \\ \sigma\rho_1 & \sigma\rho_2 & \sigma^2 \end{bmatrix} \right).$$

При таком допущении состоятельные оценки параметров модели могут быть получены с помощью метода максимального правдоподобия или двухшаговой процедуры. В последнем случае на первом шаге оценивается двумерная пробит модель, а на втором — методом наименьших квадратов оценивается условное математическое ожидание целевой переменной:

$$\mathrm{E}(y_i) = \mathrm{E}\left(y_i^* \mid z_{1i}^* \geq 0,\, z_{2i}^* \geq 0\right) = x_i'\beta + \rho_1\sigma\lambda_{1i}(w_{1i}'\gamma_1, w_{2i}'\gamma_2, \rho_0) + \rho_2\sigma\lambda_{2i}(w_{1i}'\gamma_1, w_{2i}'\gamma_2, \rho_0),$$

$$\lambda_{ji}(w_{1i}'\gamma_1, w_{2i}'\gamma_2, \rho_0) = \frac{\varphi(w_{ji}'\gamma_j)}{\Phi_2(w_{1i}'\gamma_1, w_{2i}'\gamma_2; \rho_0)}\Phi\left(\frac{w_{ki}'\gamma_k - \rho_0 w_{ji}'\gamma_j}{\sqrt{1-\rho_0^2}}\right);\ \ j,k \in \{1,2\},\ k \neq j, \qquad (4)$$

где $\varphi(\cdot)$ и $\Phi(\cdot)$ — соответственно функции плотности и распределения стандартного нормального закона, $\Phi_2(\cdot,\cdot;\rho_0)$ — функция распределения двумерного стандартного нормального закона с корреляцией ρ_0. Вместо истинных значений параметров γ, ρ_0 используются их оценки, полученные на первом шаге. Особенности оценивания моделей такого вида обсуждаются в работе (Коссова, Потанин, 2018).

Данный подход позволяет получить состоятельные, а в случае использования метода максимального правдоподобия и асимптотически эффективные, оценки всех параметров модели, а также, за счет оценивания параметров совместного распределении случайных ошибок, определить характер связи между ними и оценить условное математическое ожидание зависимой переменной целевого уравнения при различных результатах отбора. Так, например, можно оценить условное математическое ожидание доли расходов на питание,

в том числе для тех домохозяйств, которые не готовы ответить ни на вопрос о своих доходах, ни на вопрос о тратах на еду.

Недостатком параметрического подхода является то обстоятельство, что свойства оценок максимального правдоподобия зависят от того, насколько верно специфицирована функция правдоподобия. Нарушение предпосылки о нормальном совместном распределении случайных ошибок может привести к потере состоятельности оценок. Альтернативным подходом является отказ от предположения о виде распределения случайных ошибок и использование полупараметрических и полу-непараметрических подходов, которые будут рассмотрены ниже.

2.2. Полупараметрический двухшаговый подход Newey

Полупараметрический подход подразумевает, что данные описываются уравнениями (1)–(3), но совместное распределение случайных ошибок неизвестно. Опишем принцип работы метода в случае одного уравнения отбора. Поскольку результат отбора зависит лишь от линейного индекса, условное математическое ожидание целевой переменной можно представить в следующем виде:

$$\mathrm{E}(y_i) = \mathrm{E}(y_i^* \mid z_i = 1) = \mathrm{E}\left(y_i^* \mid u_i > -w_i\gamma\right) = x_i\beta + \mathrm{E}(\varepsilon_i \mid u_i > -w_i\gamma) = x_i\beta + g^*(w_i\gamma). \quad (5)$$

Исходя из этого соображения, Newey (2009) предложил следующую двухшаговую процедуру, позволяющую получить состоятельные, асимптотически нормальные оценки β:

- на первом шаге при помощи полупараметрической или непараметрической модели бинарного выбора оценивается линейный индекс $w_i\gamma$;
- на втором шаге неизвестная функция g^* аппроксимируется с помощью сплайнов или полинома $g_k(w_i\gamma) = \sum_{m=1}^{k} \tau_m s(w_i\gamma)^m$, где $s(\cdot)$ — сглаживающая функция, в качестве которой обычно берется обратное отношение Миллса для стандартного нормального распределения $s(w_i\gamma) = \varphi(w_i\gamma)/\Phi(w_i\gamma)$. При этом рекомендуется предварительно стандартизировать линейный индекс в зависимости от оцененной или, по крайней мере, предполагаемой дисперсии и математического ожидания случайной ошибки уравнения отбора (Newey, 2009). Затем вместо $g^*(w_i\gamma)$ в (5) подставляется $\widehat{g_k(w_i\gamma)}$, и методом наименьших квадратов оцениваются β, $\tau_1, \ldots, \tau_k$. Параметр k подбирается при помощи кросс-валидации (leave-one-out), обычно предполагающей минимизацию квадратичной ошибки прогноза[2].

Newey также вывел выражение для состоятельной оценки ковариационной матрицы оценок регрессионных коэффициентов. Однако на практике для упрощения вычислений соответствующая ковариационная матрица обычно оценивается при помощи бутстрапа. Для тестирования гипотез можно использовать бутстрапированные доверительные интервалы

[2] Необходимо n раз повторить следующую процедуру. Сначала из выборки изымается одно, каждый раз новое, наблюдение. Затем на оставшихся $n - 1$ наблюдениях оценивается модель, после чего полученные оценки используются для расчета квадратичной ошибки прогноза значения изъятого из выборки наблюдения. Полученные по результатам n итераций квадратичные ошибки суммируются, и лучшей признается та модель, для которой соответствующая сумма оказывается наименьшей.

или асимптотическое распределение оценок, которое является нормальным. Отметим, что процедура бутстрапирования может оказаться достаточно ресурсоемкой, поскольку в каждой итерации необходимо переоценивать оба шага, а также осуществлять кросс-валидацию по k.

В работе (De Luca, Peracchi, 2012) предлагается обобщение метода Newey на случай двух уравнений отбора, наличие которых приводит к тому, что условное математическое ожидание зависимой переменной целевого уравнения принимает вид

$$\begin{aligned} \mathrm{E}(y_i) &= \mathrm{E}(y_i^* \mid z_{1i} = 1,\, z_{2i} = 1) = \mathrm{E}\left(y_i^* \mid u_{1i} > -w_{1i}\gamma_1,\, u_{2i} > -w_{2i}\gamma_2\right) = \\ &= x_i\beta + \mathrm{E}(\varepsilon_i \mid u_{1i} > -w_{1i}\gamma_1,\, u_{2i} > -w_{2i}\gamma_2) = x_i\beta + g^*(w_{1i}\gamma_1, w_{2i}\gamma_2). \end{aligned} \tag{6}$$

В соответствии с данным подходом, на первом шаге необходимо оценить линейные индексы $w_{1i}\gamma_1$ и $w_{2i}\gamma_2$ при помощи двумерной полупараметрической модели бинарного выбора. На втором шаге условное математическое ожидание случайной ошибки целевого уравнения аппроксимируется при помощи следующей функции:

$$g_k(w_{1i}\gamma_1, w_{2i}\gamma_2) = \sum_{m=1}^{k}\left(\tau_m^{(1)} s_1(w_{1i}\gamma_1, w_{2i}\gamma_2)^m + \tau_m^{(2)} s_2(w_{1i}\gamma_1, w_{2i}\gamma_2)^m\right). \tag{7}$$

При этом в качестве s_1 и s_2 используются обобщенные обратные отношения Миллса (4). Затем $g_k(\widehat{w_{1i}\gamma_1}, \widehat{w_{2i}\gamma_2})$ подставляется в уравнение (6), параметры которого оцениваются методом наименьших квадратов. Для определения степени полиномов k используется кросс-валидация.

Важно отметить, что ни в указанной работе, ни самим Newey не было представлено доказательство состоятельности оценок многомерной версии данного метода. Неизвестным также остается их асимптотическое распределение. В связи с этим приобретает актуальность проверка свойств соответствующих оценок, по крайней мере, на симулированных данных.

Соотношение (4) является достаточно сложным и требует предварительного нахождения оценки корреляции между случайными ошибками ρ_0, поэтому представляет интерес изучение свойств оценок β при использовании более простых функций от линейных индексов. В рамках данного исследования, помимо основного, рассматриваются два дополнительных подхода к спецификации (7). При этом в качестве функции s выступают обычные обратные отношения Миллса, что потенциально позволяет использовать на первом шаге полупараметрические системы бинарных уравнений, не оценивающие параметр ρ_0, а также оценивать параметры уравнений отбора по отдельности, т. е. в рамках одномерных моделей бинарного выбора, полагая $k = (k_1, k_2)$, имеем:

$$g_k(w_{1i}\gamma_1, w_{2i}\gamma_2) = \sum_{m=1}^{k_1}\tau_m^{(1)} s(w_{1i}\gamma_1)^m + \sum_{m=1}^{k_2}\tau_m^{(2)} s(w_{2i}\gamma_2)^m + \tau_0, \tag{8}$$

$$g_k(w_{1i}\gamma_1, w_{2i}\gamma_2) = \sum_{m=0}^{k_1}\sum_{l=0}^{k_2}\tau_{ml} s(w_{1i}\gamma_1)^m s(w_{2i}\gamma_2)^l. \tag{9}$$

Отметим, что выражение (8) является частным случаем (9), в котором фигурируют перекрестные коэффициенты, которые добавляются для того, чтобы учесть статистическую связь между переменными отбора.

В настоящей статье используется подход Newey с оцениванием первого шага методом, предложенным (Gallant, Nychka, 1987), суть которого будет описана ниже. Преимущество этого способа оценивания первого шага заключается в том, что оцениваются не только линейные индексы, но и корреляция между случайными ошибками в уравнениях отбора. Отметим также, что при оценивании моделей на симуляциях не использовались степени k_1 и k_2 выше трех, поскольку предварительный анализ показал, что степени более высокого порядка крайне редко обеспечивают бóльшую точность, и соответствующее преимущество, как правило, оказывается несущественным.

Достоинством рассмотренного метода, помимо более слабых предположений о совместном распределении случайных ошибок, является довольно невысокая сложность его реализации, особенно при спецификациях с обычными обратными отношениями Миллса, позволяющими вместо системы бинарных уравнений оценивать две одномерные модели. Однако данный подход обладает и рядом недостатков. Во-первых, должны быть соблюдены ограничения исключения (exclusion restrictions): каждое из уравнений отбора должно включать по крайней мере одну уникальную переменную, которой нет в целевом уравнении. Во-вторых, поскольку аппроксимируется лишь условное математическое ожидание зависимой переменной целевого уравнения для положительного результата отбора ($z_{1i}z_{2i}=1$), остаются неоцененными параметры совместного распределения случайных ошибок. Поэтому невозможно получить прогнозы условного математического ожидания целевой переменной для случаев, когда $z_{1i}=0$ или $z_{2i}=0$. В-третьих, так как в состав аппроксимирующей функции входит константа, то константа целевого уравнения не идентифицируется на втором шаге. Для ее идентификации необходимо использовать дополнительные (и довольно сложные) методы.

2.3. Полупараметрический двухшаговый подход Das–Newey–Vella

Обозначим через $F_X(x)$ функцию распределения случайной величины X в точке x. Тогда условное распределение случайной ошибки целевого уравнения в случае единственного уравнения отбора можно переписать в следующем виде:

$$\mathrm{E}\left(\varepsilon_i \mid z_i=1\right)=\mathrm{E}\left(\varepsilon_i \mid u_i \geq -w_i\gamma\right)=\mathrm{E}\left(\varepsilon_i \mid F_{u_i}(u_i) \geq F_{u_i}(-w_i\gamma)\right)=\mathrm{E}\left(\varepsilon_i \mid 1-F_{u_i}(u_i)<1-F_{u_i}(-w_i\gamma)\right)=$$
$$=g^*\left(1-F_{u_i}(-w_i\gamma)\right)=g^*\left(\mathrm{P}(u_i \geq -w_i\gamma)\right)=g^*\left(\mathrm{P}(z_i=1)\right).$$

Таким образом, условное математическое ожидание случайной ошибки можно представить в виде функции от вероятности отбора наблюдения. В работе (Das et al., 2003) был предложен подход, аналогичный подходу Newey, за исключением того, что вместо оценок линейного индекса $w_i\gamma$ в функцию g_k подставляются оценки вероятностей положительного отбора $p_{z_i}=\mathrm{P}(z_i=1)$, полученные при помощи непараметрических и полупараметрических подходов, а также метода наименьших квадратов. При этом сглаживающая функция принимает вид $s(p_{z_i})=p_{z_i}$ или $s(p_{z_i})=\varphi\left(\Phi^{-1}(p_{z_i})\right)\Big/p_{z_i}$, где $\Phi^{-1}(p_{z_i})$ — квантиль уровня p_{z_i} стандартного нормального распределения. В ходе предварительного анализа было выявлено, что несколько более точные прогнозы обеспечивает первый подход к сглаживанию функции, поэтому в настоящей работе именно он используется для оценивания моделей на симулированных данных.

Авторы также рассматривают обобщение данного подхода на многомерный случай, позволяющее получить асимптотически нормальные состоятельные оценки β при наличии произвольного числа уравнений отбора. Подход к обобщению аналогичен описанному для метода Newey, в рамках данной работы рассматриваются следующие конкурирующие спецификации аппроксимирующих функций:

$$g_k(p_{z_{1i}}, p_{z_{2i}}) = \sum_{m=1}^{k_1} \tau_m^{(1)} s(p_{z_{1i}})^m + \sum_{m=1}^{k_2} \tau_m^{(2)} s(p_{z_{2i}})^m + \tau_0, \quad (10)$$

$$g_k(p_{z_{1i}}, p_{z_{2i}}) = \sum_{m=0}^{k_1} \sum_{l=0}^{k_2} \tau_{(m,l)} s(p_{z_{1i}})^m s(p_{z_{2i}})^l, \quad (11)$$

где $p_{z_{1i}} = \mathrm{P}(z_{1i} = 1)$ и $p_{z_{2i}} = \mathrm{P}(z_{2i} = 1)$.

Недостатки и преимущества данного и предыдущего подходов довольно похожи, поэтому выбор между ними можно осуществлять на основе кросс-валидации или в зависимости от того, что проще или точнее можно рассчитать: линейные индексы или оценки вероятностей положительного результата отбора. Также следует отметить, что использование вероятностей положительного отбора $\mathrm{P}(z_{ji} = 1)$ вместо линейных индексов позволяет снять предположение о том, что функциональная форма уравнения отбора является линейной по параметрам.

Следуя примеру приложения данного метода к реальным данным, представленному в (Das et al., 2003) в разделе, посвященном анализу симулированных данных, на первом шаге вероятности оцениваются при помощи метода наименьших квадратов.

2.4. Полупараметрический трехшаговый разностный подход Robinson

В работе (Robinson, 1988) предлагается избежать смещения отбора с помощью вычитания из целевого уравнения равенства, полученного взятием условного математического ожидания (при фиксированном значении линейного индекса) от обеих частей целевого уравнения. Этот прием позволяет исключить из целевого уравнения смещение, являющееся функцией от линейного индекса. В случае двумерного неслучайного отбора новое оцениваемое уравнение принимает вид

$$y_i - \mathrm{E}\left(y_i \mid w_{1i}'\gamma_1, w_{2i}'\gamma_2\right) = \left[x_i - \mathrm{E}\left(x_i \mid w_{1i}'\gamma_1, w_{2i}'\gamma_2\right)\right]' \beta + \varepsilon_i. \quad (12)$$

На первом шаге, обычно, при помощи полупараметрической модели бинарного выбора, предложенной в (Klein, Spady, 1993), оцениваются линейные индексы[3]. На втором шаге, с помощью ядерных регрессий каждой из переменных на линейные индексы оцениваются условные математические ожидания. Как правило, при этом используется метод, предложенный

[3] Для получения оценок параметров модели осуществляется максимизация функции правдоподобия с заменой неизвестной функции распределения ее ядерной оценкой.

Nadaraya (1964) и Watson (1964). Наконец, на третьем шаге полученные оценки условных математических ожиданий подставляются в (12), и методом наименьших квадратов находятся состоятельные оценки β.

В настоящей работе для случая двух уравнений отбора рассматривается обобщение данного метода и его модификация, когда и на первом и на втором шагах используются случайные леса (Cook, Siddiqui, 2019).

Преимущества и недостатки подхода Robinson схожи с рассмотренными ранее. С точки зрения вычислительной реализации преимущество заключается в том, что нет необходимости применять кросс-валидацию для выбора степени полинома, а также отдавать предпочтение той или иной сглаживающей функции. В то же время эффективность итоговых оценок может зависеть от точности оценок, полученных на первых двух шагах.

2.5. Полу-непараметрический подход Gallant–Nychka

Данный метод применим практически для любых многомерных эконометрических моделей, включая модели бинарного выбора и модели с неслучайным отбором. Идея рассматриваемого подхода заключается в том, чтобы аппроксимировать неизвестную совместную функцию плотности при помощи полинома в форме Эрмита. В частности, учитывая, что для каждого наблюдения $i \in \{1,\ldots,n\}$ векторы случайных ошибок имеют одинаковое совместное распределение, следуя (Gallant, Nychka, 1987), получаем следующую аппроксимирующую функцию:

$$\tilde{f}_{(u_{1i},u_{2i},\varepsilon_i)'}(x_1,x_2,x_3)=\frac{\sum\limits_{i_1,i_2,i_3,j_1,j_2,j_3}\alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}\prod\limits_{k=1}^{3}x_k^{i_k+j_k}\varphi(x_k;\mu_k,\sigma_k)}{\sum\limits_{i_1,i_2,i_3,j_1,j_2,j_3}\alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}\prod\limits_{k=1}^{3}\mathcal{M}(i_k+j_k;\mu_k,\sigma_k)},$$

где $\varphi(x_k;\mu_k,\sigma_k)$ и $\mathcal{M}(i_k+j_k;\mu_k,\sigma_k)$, $k\in\{1,2,3\}$ — функция плотности в точке x_k и момент порядка i_k+j_k соответственно, относящиеся к нормальному распределению с математическим ожиданием μ_k и дисперсией σ_k^2, а $\sum\limits_{i_1,i_2,i_3,j_1,j_2,j_3}$ здесь и далее в разделе обозначает суммирование по $i_1,j_1\in\{0,\ldots,K_1\}$, $i_2,j_2\in\{0,\ldots,K_2\}$, $i_3,j_3\in\{0,\ldots,K_3\}$. Параметры K_1, K_2 и K_3 принимают неотрицательные целые значения и определяют порядок аппроксимирующего полинома, а $\alpha_{(i_1,i_2,i_3)}$ — коэффициенты при его степенях. В работе (Gallant, Nychka, 1987) было доказано, что если в функции правдоподобия вместо неизвестной функции плотности использовать аппроксимирующую функцию $\tilde{f}$ и увеличивать степени полинома по мере возрастания объема выборки, то можно получить состоятельные оценки параметров модели. При этом оценки могут оказаться несостоятельными, только если истинная функция плотности имеет очень высокую частоту колебаний (Gallant, Nychka, 1987). В своей работе Gallant и Nychka рассматривают случай с одномерным смещением отбора. В настоящем исследовании функция правдоподобия для модели с двойным неслучайным отбором выглядит следующим образом:

$$L(\beta,\gamma_1,\gamma_2,\alpha_{(0,0,0)},\ldots,\alpha_{(K_1,K_2,K_3)},\mu_1,\mu_2,\mu_3,\sigma_1,\sigma_2,\sigma_3) =$$
$$= \prod_{z_{1i}=1,z_{2i}=1} \tilde{F}_{u_{1i},u_{2i}|\epsilon_i=y_i-x_i\beta}(-w_{1i}\gamma_1,-w_{2i}\gamma_2,\infty,\infty)\tilde{f}_{\epsilon_i}(x) \prod_{z_{1i}=1,z_{2i}=0} \tilde{F}_{u_{1i},u_{2i}}(-w_{1i}\gamma_1,-\infty,\infty,-w_{2i}\gamma_2)\times$$
$$\times \prod_{z_{1i}=0,z_{2i}=1} \tilde{F}_{u_{1i},u_{2i}}(-\infty,-w_{2i}\gamma_2,-w_{1i}\gamma_1,\infty) \prod_{z_{1i}=0,z_{2i}=0} \tilde{F}_{u_{1i},u_{2i}}(-\infty,-\infty,-w_{1i}\gamma_1,-w_{2i}\gamma_2),$$

где[4]

$$\tilde{F}_{u_{1i},u_{2i}|\varepsilon_i=x}(\underline{x}_1,\underline{x}_2,\overline{x}_1,\overline{x}_2) = \mathrm{P}(\underline{x}_1 \le u_{1i} \le \overline{x}_1,\ \underline{x}_2 \le u_{2i} \le \overline{x}_2 \mid \varepsilon_i = x) =$$
$$= \frac{\sum_{i_1,i_2,i_3,j_1,j_2,j_3} \alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}x^{i_3+j_3}\prod_{k=1}^{2}\left(\Phi(\overline{x}_k;\mu_k,\sigma_k)-\Phi(\underline{x}_k;\mu_k,\sigma_k)\right)\mathcal{M}_{tr}(i_k+j_k;\mu_k,\sigma_k;\underline{x}_k,\overline{x}_k)}{\sum_{i_1,i_2,i_3,j_1,j_2,j_3} \alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}x^{i_3+j_3}\mathcal{M}(i_3+j_3;\mu_3,\sigma_3)},$$

$$\tilde{F}_{u_{1i},u_{2i}}(\underline{x}_1,\underline{x}_2,\overline{x}_1,\overline{x}_2) = \mathrm{P}(\underline{x}_1 \le u_{1i} \le \overline{x}_1,\underline{x}_2 \le u_{2i} \le \overline{x}_2) =$$
$$= \frac{\sum_{i_1,i_2,i_3,j_1,j_2,j_3} \alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}\mathcal{M}(i_3+j_3;\mu_3,\sigma_3)\prod_{k=1}^{2}\left(\Phi(\overline{x}_k;\mu_k,\sigma_k)-\Phi(\underline{x}_k;\mu_k,\sigma_k)\right)\mathcal{M}_{tr}(i_k+j_k;\mu_k,\sigma_k;\underline{x}_k,\overline{x}_k)}{\sum_{i_1,i_2,i_3,j_1,j_2,j_3} \alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}\mathcal{M}(i_1+j_1;\mu_1,\sigma_1)\mathcal{M}(i_2+j_2;\mu_2,\sigma_2)},$$

$$\tilde{f}_{\epsilon_i}(x) = \varphi(x;\mu_3,\sigma_3)\frac{\sum_{i_1,i_2,i_3,j_1,j_2,j_3} \alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}x^{i_3+j_3}\prod_{k=1}^{2}\mathcal{M}_{tr}(i_k+j_k;\mu_k,\sigma_k;\underline{x}_k,\overline{x}_k)}{\sum_{i_1,i_2,i_3,j_1,j_2,j_3} \alpha_{(i_1,i_2,i_3)}\alpha_{(j_1,j_2,j_3)}\mathcal{M}(i_3+j_3;\mu_3,\sigma_3)},$$

а $\mathcal{M}_{tr}(i_k+j_k;\mu_k,\sigma_k;\underline{x}_k,\overline{x}_k)$ есть (i_k+j_k)-й момент усеченной снизу в точке $\underline{x}_k$ и сверху в точке $\overline{x}_k$ нормальной случайной величины с математическим ожиданием μ_k и дисперсией σ_k^2.

Степени полинома K_1, K_2 и K_3 выбираются с помощью кросс-валидации при минимизации значения информационного критерия Акаике. Поскольку оценивание параметров при больших степенях полинома является весьма длительным, в рамках данного исследования при симуляциях полагается $K_1 = K_2 = K_3 = 1$.

К преимуществу данного подхода можно отнести возможность совмещать гибкость полупараметрических методов, позволяющую ослабить строгие предположения о распределении случайных ошибок, с простотой интерпретации параметрических методов, т. к. за счет аппроксимации совместного распределения можно рассчитать любую из его характеристик, включая условные математические ожидания и корреляционную матрицу. Поскольку $\tilde{f}$ при фиксированных значениях K_1, K_2 и K_3 удовлетворяет всем свойствам функции плотности, данный подход можно интерпретировать также как полностью параметрический — предполагается, что совместное распределение случайных ошибок относится к некоторому широкому семейству распределений.

[4] Вывод соответствующих формул предоставляется авторами по запросу.

Недостатком данного метода является необходимость оценивать большое число параметров. Это может привести, во-первых, к потере эффективности оценок, а во-вторых, к сложностям, связанным с нахождением максимума функции правдоподобия, поскольку она не обязательно вогнута. Также, при отказе от параметрической интерпретации для тестирования гипотез о параметрах необходимо использовать бутстрапированные доверительные интервалы, т. к. авторы метода не выводят асимптотическое распределение оценок, а ограничиваются лишь доказательством их состоятельности.

Отметим, что для идентифицируемости оцениваемых параметров при использовании данного метода следует стандартизировать один из коэффициентов полинома. Как правило, в том числе и в данной работе, полагается $\alpha_{(0,0,0)} = 1$. Также, поскольку математическое ожидание случайной ошибки не обязательно равняется нулю, при некоторых совместных распределениях (например, если истинные частные распределения случайных ошибок окажутся стандартными нормальными) неидентифицируемой может оказаться константа. Поэтому она исключается из всех трех уравнений и оценивается как математическое ожидание случайной ошибки целевого уравнения. Также полагается $\sigma_1 = \sigma_2 = 1$, и в бинарных уравнениях коэффициенты при первых объясняющих переменных фиксируются значениями 1, что является стандартной практикой при оценивании полупараметрических моделей.

3. Дизайн эксперимента

Для исследования поведения оценок параметров моделей в ситуациях различного типа отклонения совместного распределения случайных ошибок от нормального рассматриваются следующие распределения:

- распределение Стьюдента с 5 степенями свободы — тяжелые хвосты;
- бета-распределение — асимметрия;
- смесь двух нормальных распределений — бимодальность.

Число степеней свободы в распределении Стьюдента было выбрано так, чтобы рассмотреть распределение с более тяжелыми хвостами, чем у нормального. Параметры компонент случайного вектора, имеющего многомерное бета-распределение, были зафиксированы на уровне $a = 2$ и $b = 5$ для обеспечения асимметрии. Значения случайных ошибок из распределения Стьюдента и бета-распределения были умножены на $\sqrt{5.4}$ и 11.2 соответственно[5], для

[5] При генерации случайных ошибок было задано условие, чтобы дисперсия случайной ошибки основного уравнения была сопоставима с вариацией заданного линейного индекса, которая определилась после генерации значений объясняющих переменных. Таким образом, для распределения Стьюдента было получено значение дисперсии, равное 9. Особенности датчика случайных чисел, генерирующего многомерное распределение Стьюдента, состоят в том, что сначала генерируются одномерные случайные величины, распределенные по Стьюденту и имеющие дисперсию для пяти степеней свободы, равную 5/3, а затем их значения умножаются на нужную константу, чтобы выйти на заданную исследователем ковариационную матрицу. В нашем случае эта константа равна $\sqrt{5.4} = \sqrt{9 \cdot 3/5}$. Для того чтобы смоделировать бета-распределение случайной ошибки основного уравнения, использовался датчик случайных чисел, генерирующий многомерное распределение Дирихле, в котором ковариационная матрица определяется заданными параметрами распределения. Параметры выбирались таким образом, чтобы была очевидна несимметричность распределения случайной ошибки основного уравнения. Поскольку значение дисперсии оказалось в этих условиях достаточно малым (0.025), было принято решение умножить значения случайной ошибки на 11.2, что дает ковариационную матрицу, указанную далее в тексте.

того чтобы их дисперсии не оказались слишком маленькими: в противном случае оценки стали бы чрезвычайно близкими к истинным значениям параметров независимо от метода оценивания. Для генерации смеси из нормальных распределений с различными дисперсиями (4 и 1) и математическими ожиданиями (5 и 10) симулировались значения из обоих распределений[6], а затем с равной вероятностью выбиралось одно из них для каждого наблюдения. Наконец, при симуляции из всех случайных ошибок вычиталось их математическое ожидание.

В качестве независимых регрессоров генерировалось пять переменных (x_1, x_2, x_3, x_4, x_5) из многомерного нормального распределения с нулевым вектором математических ожиданий и случайной ковариационной матрицей, сгенерированной по методу[7], предложенному в (Joe, 2006). Во многих работах независимые переменные генерируются из совместного равномерного распределения. Однако в данном исследовании такой подход не рассматривался, т. к. при этом происходит искусственное устранение выбросов среди наблюдений и значения независимых переменных не выходят за пределы заданного отрезка, что не характерно для рассматриваемых на практике выборок.

Для каждого из изучаемых распределений случайных ошибок были осуществлены 100 симуляций в соответствии со следующим процессом генерации данных:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i ,$$

$$z_{1i} = 1 + x_3 + x_4 + x_5 + u_{1i},\ z_{2i} = 1 + x_1 + x_3 + x_5 + u_{2i},$$

$$\left(u_{1i}, u_{2i}, \varepsilon_i\right) \sim \Theta(\theta),\ i \in \{1, \ldots, n\},\ \theta \in R^{n_\theta},\ n = 5000,$$

$$\mathrm{E}(u_{1i}) = \mathrm{E}(u_{2i}) = \mathrm{E}(\varepsilon_i) = 0,$$

$$\mathrm{Cov}\left(u_{1i}, u_{2i}, \varepsilon_i\right) = \begin{cases} \begin{bmatrix} 1 & 0.5 & 1.5 \\ 0.5 & 1 & 1.5 \\ 1.5 & 1.5 & 9 \end{bmatrix}, & \text{для распределения Стьюдента,} \\ \begin{bmatrix} 3.2 & -1.6 & -1.6 \\ -1.6 & 3.2 & -1.6 \\ -1.6 & -1.6 & 3.2 \end{bmatrix}, & \text{для бета-распределения,} \\ \begin{bmatrix} 7.25 & 6.75 & 7 \\ 6.75 & 7.25 & 7 \\ 7 & 7 & 8.75 \end{bmatrix}, & \text{для смеси нормальных распределений,} \end{cases}$$

$$\beta_0 = 1,\ \beta_1 = -1.5,\ \beta_2 = 1.5,\ \beta_3 = -1.5.$$

[6] Параметры нормальных распределений выбирались таким образом, чтобы итоговая смесь имела дисперсию, близкую к 9, как в распределении Стьюдента.

[7] Для этого в пакете «clusterGeneration» в среде R используется метод onion.

Значения коэффициентов β, за исключением константы, были выбраны одинаковыми по модулю с целью сравнения точности оценок между собой. Если подбирать истинные коэффициенты таким образом, что один является больше других по абсолютному значению, модель может оценивать его лучше только потому, что его вклад в объяснение дисперсии целевой переменной выше. Нам же интересна относительная точность в оценивании коэффициентов в зависимости от их одновременного наличия или отсутствия в уравнениях. Так, например, предполагается, что наименьшую точность все методы продемонстрируют в отношении оценивания коэффициента β_3, поскольку переменная x_3 включена в оба уравнения отбора.

Ниже в таблице 1 представлена информация об используемых моделях. Обозначения моделей будут фигурировать в таблицах следующего раздела.

Таблица 1. Используемые параметрические и полупараметрические модели

Модель	*Первый шаг*	*Второй шаг*
Хекман 2 шага	Пробит	МНК
Хекман ММП	Метод максимального правдоподобия	
Newey 1	Полу-непараметрическая модель бинарного выбора с использованием метода (Gallant, Nychka, 1987)	МНК с аппроксимирующей функцией (7)
Newey 2		МНК с аппроксимирующей функцией (8)
Newey 3		МНК с аппроксимирующей функцией (9)
Das–Newey–Vella 1	Модель бинарного выбора, оцененная при помощи МНК	МНК с аппроксимирующей функцией (10)
Das–Newey–Vella 2		МНК с аппроксимирующей функцией (11)
Robinson 1	Модель бинарного выбора (Klein, Spady, 1993), а также ядерная регрессия (Nadaraya, 1964; Watson, 1964).	МНК
Robinson 2	Случайные леса (Cook, Siddiqui, 2019)	
Gallant–Nychka	Метод максимального правдоподобия с заменой неизвестной функции плотности на ее аппроксимацию полиномами Эрмита	
МНК	Отсутствует	МНК

Были использованы следующие меры качества оценок.

1. Среднее значение оценки каждого коэффициента по всем симуляциям:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_{ji},\ \ j = 0,1,2,3.$$

2. Среднеквадратичная ошибка:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{ji} - \beta_j)^2,\ \ j = 0,1,2,3.$$

3. 94%-ные доверительные интервалы (CI). Каждый интервал строится путем исключения из результатов 100 симуляций трех наименьших и трех наибольших оценок. В качестве нижней границы интервала (*CIL*) выбирается минимальное из оставшихся значений, а в качестве верхней (*CIU*) — максимальное.

4. Среднее относительное отклонение от истинного значения:

$$MD = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{\beta}_{ji} - \beta_j|}{|\beta_j|}, \quad j = 0,1,2,3.$$

4. Результаты симуляций

В центре наших интересов находятся коэффициенты при объясняющих переменных $\beta_1, \beta_2, \beta_3$. Оценивание константы β_0 непараметрическими методами затруднено, а при некоторых подходах и невозможно. Тем не менее, для полноты картины в таблицах ниже приводятся и оценки β_0, при этом пустые ячейки оставлены для тех случаев, когда оценивание константы невозможно.

4.1. Результаты симуляций со случайными ошибками из распределения Стьюдента

Согласно полученным результатам, представленным в табл. 2, наиболее точными, с точки зрения рассматриваемых критериев, оказались оценки при помощи параметрических методов и полупараметрического метода Newey. При этом как параметрические, так и полупараметрические методы, учитывающие смещение отбора, дали результаты, близкие к истинным значениям. Среднее относительное отклонение всех оценок не превышало, за редким исключением, 10%, в то время как обычные МНК оценки заметно отличаются от истинных значений параметров. Также следует отметить, что качество оценок коэффициентов не продемонстрировало зависимости от того, является ли соответствующая переменная уникальной для целевого уравнения, или входит также в одно или оба уравнения отбора.

Все спецификации метода Newey обеспечили приблизительно одинаковое качество оценок, в то время как второй подход к спецификации метода Robinson (с использованием случайного леса) оказался несколько точней, чем первый. Более низкая, чем у других полупараметрических методов, относительная точность оценок, полученных при помощи метода Gallant–Nychka, может быть связана с использованием в рамках данного исследования полинома достаточно малой степени. Также данный метод, наряду с подходами Das–Newey–Vella и Robinson, довольно часто выдавал далекие от истинных коэффициентов значения оценок, что привело к высоким значениям чувствительного к выбросам критерия MSE.

Наконец, отметим, что близкие к истинным величинам β_1, β_2 и β_3 средние значения оценок, полученных при помощи метода Newey, являются косвенным свидетельством их несмещенности.

Таблица 2. Статистические характеристики оценок с помощью различных моделей. Ошибки из распределения Стьюдента

Мера	Хекман		Newey			Das–Newey–Vella		Robinson		Gallant–Nychka	МНК
	2 шага	ММП	1	2	3	1	2	1	2		
						$\hat{\beta}_0$					
ME	0.945	0.941	1.041	1.013	1.031	10.45	6.242			0.996	1.720
MSE	0.017	0.017	0.051	0.039	0.051	92.70	322.7			0.040	0.525
MD	0.102	0.105	0.173	0.158	0.173	9.445	14.007			0.031	0.730
CIL	0.705	0.740	0.706	0.703	0.641	6.656	–29.10			0.922	1.565
CIU	1.159	1.141	1.539	1.373	1.476	14.08	43.33			1.092	1.870
						$\hat{\beta}_1$					
ME	–1.485	–1.489	–1.496	–1.498	–1.499	–1.462	–1.468	–1.513	–1.517	–1.611	–1.657
MSE	0.008	0.010	0.014	0.013	0.013	0.293	0.289	0.054	0.032	0.036	0.065
MD	0.050	0.050	0.063	0.056	0.058	0.129	0.130	0.111	0.091	0.102	0.135
CIL	–1.686	–1.735	–1.716	–1.703	–1.706	–1.967	–1.977	–1.855	–1.910	–1.944	–1.960
CIU	–1.347	–1.324	–1.280	–1.310	–1.310	–1.038	–1.035	–1.070	–1.230	–1.251	–1.290
						$\hat{\beta}_2$					
ME	1.495	1.495	1.497	1.496	1.498	1.504	1.502	1.495	1.443	1.499	1.484
MSE	0.011	0.012	0.014	0.014	0.013	0.022	0.021	0.025	0.018	0.033	0.055
MD	0.054	0.051	0.055	0.054	0.054	0.069	0.067	0.070	0.066	0.091	0.112
CIL	1.344	1.327	1.298	1.314	1.325	1.231	1.239	1.170	1.173	1.210	1.078
CIU	1.678	1.67	1.672	1.695	1.689	1.781	1.766	1.700	1.620	1.879	1.879
						$\hat{\beta}_3$					
ME	–1.466	–1.471	–1.499	–1.500	–1.501	–1.498	–1.500	–1.505	–1.499	–1.734	–1.859
MSE	0.014	0.012	0.016	0.013	0.014	0.036	0.035	0.048	0.032	0.082	0.160
MD	0.060	0.057	0.063	0.058	0.060	0.086	0.086	0.099	0.094	0.164	0.244
CIL	–1.646;	–1.648	–1.759	–1.756	–1.759	–1.826	–1.863	–2.005	–1.833	–2.086	–2.155
CIU	–1.267	–1.292	–1.270	–1.312	–1.342	–1.208	–1.212	–1.046	–1.197	–1.407	–1.535

4.2. Результаты симуляций со случайными ошибками из бета-распределения

При данном распределении случайных ошибок параметрический подход продемонстрировал явное преимущество в точности оценок по сравнению с полупараметрическими методами, о чем свидетельствуют результаты, представленные в табл. 3. Все спецификации метода Newey показывают примерно одинаковую точность, имея заметное превосходство над методом Robinson, среди спецификаций которого в данном случае сложно выбрать лучшую, поскольку некоторые коэффициенты лучше оценивались при одной спецификации, а другие — при иной. Хуже всех[8] показал себя метод Gallant и Nychka, для которого среднее относительное отклонение оценок от истинных значений превысило 10%.

[8] Лишь коэффициент β_1 хуже всего, согласно критериям *MD* и *MSE*, оценивается моделью Robinson 2.

Таблица 3. Статистические характеристики оценок с помощью различных моделей. Бета-распределение ошибок

Мера	Хекман		Newey			Das–Newey–Vella		Robinson		Gallant–Nychka	МНК
	2 шага	ММП	1	2	3	1	2	1	2		
						$\hat{\beta}_0$					
ME	1.022	1.093	1.059	1.131	1.095	–4.671	–6.538			0.925	0.191
MSE	0.011	0.013	0.139	0.110	0.139	33.72	200.1			0.011	0.665
MD	0.083	0.099	0.270	0.220	0.258	5.671	11.246			0.083	0.808
CIL	0.838	0.966	0.372	0.763	0.541	–6.723	–33.98			0.779	0.011
CIU	1.193	1.199	1.839	1.745	1.94	–1.970	20.74			1.042	0.385
						$\hat{\beta}_1$					
ME	–1.491	–1.514	–1.495	–1.503	–1.499	–1.473	–1.472	–1.490	–1.307	–1.340	–1.304
MSE	0.006	0.004	0.014	0.013	0.014	0.049	0.051	0.057	0.074	0.066	0.099
MD	0.041	0.031	0.056	0.051	0.054	0.073	0.072	0.099	0.141	0.133	0.164
CIL	–1.632	–1.602	–1.736	–1.706	–1.725	–1.705	–1.707	–1.888	–1.58	–1.696	–1.697
CIU	–1.367	–1.324	–1.261	–1.298	–1.303	–1.110	–1.121	–1.044	–0.814	–0.871	–0.71
						$\hat{\beta}_2$					
ME	1.502	1.495	1.495	1.495	1.495	1.494	1.495	1.446	1.430	1.502	1.501
MSE	0.004	0.003	0.008	0.007	0.008	0.015	0.014	0.039	0.029	0.038	0.067
MD	0.029	0.026	0.039	0.037	0.039	0.047	0.046	0.067	0.077	0.098	0.130
CIL	1.377	1.385	1.297	1.282	1.288	1.153	1.143	1.168	1.204	1.160	1.057
CIU	1.595	1.589	1.595	1.588	1.590	1.699	1.676	1.589	1.757	1.925	2.144
						$\hat{\beta}_3$					
ME	–1.488	–1.533	–1.502	–1.507	–1.505	–1.490	–1.490	–1.442	–1.317	–1.276	–1.199
MSE	0.005	0.004	0.010	0.008	0.009	0.041	0.041	0.040	0.068	0.082	0.144
MD	0.034	0.031	0.048	0.045	0.047	0.069	0.069	0.088	0.132	0.167	0.221
CIL	–1.617	–1.648	–1.694	–1.658	–1.689	–1.770	–1.776	–1.736	–1.559	–1.671	–1.677
CIU	–1.378	–1.292	–1.321	–1.328	–1.320	–1.140	–1.142	–0.976	–0.980	–0.845	–0.775

Отметим, что при данном «отклонении» распределения ошибок от нормального закона, а именно, при *скошенности* распределения самыми точными для всех методов оказались оценки коэффициента при уникальной переменной.

4.3. Результаты симуляций со случайными ошибками из бимодального распределения

Если результаты по симуляциям из предыдущих разделов были в целом примерно одинаковыми для всех методов, то при данном виде нарушения нормальности распределения случайных ошибок общая картина изменилась существенно, о чем говорят результаты, приведенные в табл. 4. Прежде всего, отметим, что более явным стало отличие между различными спецификациями метода Newey: третья спецификация заметно превосходит в точности

Таблица 4. Статистические характеристики оценок с помощью различных моделей. Бимодальное распределение ошибок

Мера	Хекман		Newey			Das–Newey–Vella		Robinson		Gallant–Nychka	МНК
	2 шага	ММП	1	2	3	1	2	1	2		
						$\hat{\beta}_0$					
ME	1.792	2.434	1.659	2.074	1.279	1.442	–6.034			1.867	3.191
MSE	0.998	2.220	0.567	1.307	0.156	5.290	1940.1			42.732	4.812
MD	0.821	1.434	0.663	1.074	0.321	1.473	21.26			0.984	2.191
CIL	0.917	1.773	1.025	1.359	0.804	–2.076	–146.9			0.653	3.024
CIU	3.129	3.190	2.480	2.876	1.813	5.581	48.21			1.121	3.384
						$\hat{\beta}_1$					
ME	–1.585	–1.656	–1.543	–1.562	–1.514	–1.505	1.503	–1.343	–1.486	–1.574	–1.715
MSE	0.052	0.061	0.024	0.035	0.013	0.070	0.062	0.941	0.027	0.015	0.114
MD	0.099	0.118	0.069	0.076	0.052	0.082	0.079	0.118	0.076	0.062	0.172
CIL	–1.972	–2.035	–1.854	–1.925	–1.701	–1.941	–1.933	–1.796	–1.754	–1.793	–2.230
CIU	–1.283	–1.406	–1.353	–1.352	–1.34	–1.213	–1.206	–0.99	–1.287	–1.421	–1.339
						$\hat{\beta}_2$					
ME	1.507	1.511	1.509	1.506	1.503	1.506	1.507	1.491	1.430	1.504	1.512
MSE	0.019	0.027	0.009	0.013	0.006	0.006	0.006	0.006	0.013	0.008	0.055
MD	0.058	0.080	0.042	0.047	0.034	0.035	0.034	0.037	0.058	0.047	0.127
CIL	1.303	1.268	1.367	1.338	1.359	1.365	1.374	1.345	1.255	1.336	1.051
CIU	1.752	1.753	1.658	1.702	1.614	1.663	1.687	1.629	1.563	1.677	1.899
						$\hat{\beta}_3$					
ME	–1.648	–1.763	–1.576	–1.616	–1.546	–1.473	–1.474	–1.547	–1.536	–1.636	–1.933
MSE	0.074	0.111	0.031	0.050	0.016	0.084	0.074	0.433	0.023	0.027	0.258
MD	0.115	0.179	0.070	0.090	0.056	0.073	0.070	0.126	0.069	0.094	0.294
CIL	–2.152	–2.169	–2.005	–2.044	–1.790	–1.803	–1.774	–1.786	–1.893	–1.840	–2.437
CIU	–1.388	–1.485	–1.405	–1.417	–1.397	–1.189	–1.214	–1.175	–1.303	–1.462	–1.486

оценок первые две. Возможно, это связано с тем, что стандартных отношений Миллса в данном случае недостаточно, а взаимозависимости между членами степенного ряда помогают точнее аппроксимировать условные математические ожидания случайных ошибок за счет учета связи между ними. Также следует отметить, что вторая спефикация метода Robinson оказалась существенно точнее первой по части оценивания коэффициентов β_1 и β_3.

В отличие от предыдущих видов распределения, параметрические модели с бимодальным распределением случайных ошибок уступают полупараметрическим как по средним значениям, так и по всем рассматриваемым критериям точности, для всех коэффициентов, кроме коэффициента при уникальной переменной β_2. Особенно сильные различия наблюдаются, как и ожидалось, для параметров, входящих в оба уравнения — константы и β_3. При этом двухшаговая параметрическая модель работает лучше ММП, что, вероятно, связано со схожестью двухшаговой модели Хекмана и полупараметрических двухшаговых методов по структуре. Все полупараметрические методы дают среднее относительное отклонение

от истинных значений, не превышающее 10%, в то время как отклонение оценок параметрического подхода для коэффициентов не уникальных переменных уравнения превышает 10%.

4.4. Выводы

Основной вывод, который можно сделать по результатам исследования, заключается в том, что степень различия оценок, полученных параметрическим и полупараметрическим способами, отражает то, насколько сильно совместное распределение случайных ошибок отличается от нормального. При нарушении нормальности в виде тяжелых хвостов распределения или присутствия асимметрии точность оценок, полученных при сохранении предположения о нормальном распределении, слабо отличается, а зачастую и вовсе выше, чем у методов, аппроксимирующих неизвестную функцию плотности или условное математическое ожидание различными способами. Если же отличие распределения случайных ошибок от нормального достаточно велико (как в случае бимодальности, которая фактически означает наличие в выборке двух отдельных совокупностей), различие в оценках становится уже более существенным, и в таком случае предпочтительнее использование полупараметрических методов.

Заключение

В данной работе сделана попытка дать ответ на вопрос, что происходит с оценками модели Хекмана при нарушении предположения о нормальном совместном распределении ошибок, и какое преимущество дает в этом случае применение непараметрических методов коррекции смещения отбора. Новизна работы состоит в рассмотрении двумерного механизма смещения отбора. Для устранения двумерного смещения получены модификации всех основных полупараметрических методов, используемых для оценивания моделей с одномерным смещением: Newey (использование полиномов), Das–Newey–Vella, Robinson (разностный подход с использованием ядерных оценок или случайных лесов), Gallant–Nychka (аппроксимация неизвестной совместной функции плотности случайных ошибок в форме Эрмита). Полученные обобщения описанных полупараметрических методов могут быть распространены на случай, когда число уравнений отбора больше двух.

Для оценивания точности параметрических и полупараметрических методов коррекции смещения отбора при отклонении совместного распределения случайных ошибок от нормального были проведены эксперименты на симулированных данных.

В качестве совместных распределений случайных ошибок уравнений использовались распределение Стьюдента (с пятью степенями свободы), бета-распределение и смесь двух нормальных распределений. В этих распределениях присутствуют основные признаки, демонстрирующие отсутствие нормальности: тяжелые хвосты, асимметрия и бимодальность. Как показали результаты симуляций, для унимодальных распределений с тяжелыми хвостами и асимметрией различные виды полупараметрических методов не дают преимущества (относительно параметрических методов) в близости среднего значения оценок коэффициентов к истинным. Более того, в большинстве случаев параметрические двухшаговая и ММП модели оказались точнее с точки зрения среднеквадратичной ошибки, процентного отклонения от истинных значений, а также доверительных интервалов. В ситуации с бимодальным распределением

картина меняется, и у полупараметрических методов появляется преимущество. Лучше всех из полупараметрических методов показала себя модификация метода Newey, использующая полиномы с перекрестными членами от отдельно оцененных индексов уравнений отбора.

Итак, несмотря на то что предположение о нормальном совместном распределении ошибок в модели Хекмана является достаточно жестким, его нарушение не приводит к потере качества оценок коэффициентов модели, за исключением случая бимодального распределения ошибок. Возможно, оценки ММП остаются состоятельными и для более широкого класса распределений случайных ошибок (включающего в себя нормальное распределение), но эта проблема требует отдельного исследования.

Список литературы

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

Cook J. A., Siddiqui S. (2019). Random forests and selected samples. https://ssrn.com/abstract=3068128.

Das M., Newey W. K., Vella F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 1 (1), 33–58.

De Luca G., Peracchi F. (2012). Estimating Engel curves under unit and item nonresponse. *Journal of Applied Econometrics*, 27 (7), 1076–1099.

Gallant A., Nychka D. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55 (2), 363–390.

Joe H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97, 2177–2189.

Klein R., Spady R. (1993). An efficient semiparametric estimator of the binary choice model. *Econometrica*, 61 (2), 387–421.

Nadaraya E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9 (1), 141–142.

Newey W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12, 217–229.

Pigini C. (2015). Bivariate non-normality in sample selection model. *Journal of Econometric Methods*, 4 (1), 123–144.

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

Racine J. S. (2008). Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics*, 3 (1), 1–88.

Robinson P. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56 (4), 931–954.

Vella F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, 33 (1), 127–169.

Watson G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26 (4), 359–372.

Поступила в редакцию 04.02.2020;
принята в печать 06.03.2020.

Kossova E. A., Kupriianova L., Potanin B. Parametric and semiparametric multivariate sample selection models estimators' accuracy: Comparative analysis on simulated data. *Applied Econometrics*, 2020, v. 57, pp. 119–139.

DOI: 10.22394/1993-7601-2020-57-119-139

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

Liubov Kupriianova
National Research University Higher School of Economics (NRU HSE), Moscow, Russian Federation; Lyubov.Kupriyanova@skoltech.ru

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

Parametric and semiparametric multivariate sample selection models estimators' accuracy: Comparative analysis on simulated data

This article is devoted to the comparative analysis of parametric and semiparametric sample selection models with two selection equations. Comparison has been conducted on simulated data under different random errors distributional assumptions: student, beta and mixture of normal. The results suggest that for student and beta distributions parametric models' estimates are more or equally accurate as semiparametric. However, former methods provide more accurate estimates under mixture distribution case. Therefore, parametric sample selection model estimators seem to be robust to violations of normality assumption in terms of tails thickness and asymmetry but fail to account for bimodality as good as their semiparametric counterparts.

Keywords: sample selection; heavy-tailed, asymmetric, bimodal random error distributions; semi-parametric models.

JEL classification: C34.

References

Kossova E., Potanin B. (2018). Heckman method and switching regression model multivariate generalization. *Applied Econometrics*, 50, 114–143 (in Russian).

Cook J. A., Siddiqui S. (2019). Random forests and selected samples. https://ssrn.com/abstract=3068128.

Das M., Newey K., Vella F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 1 (1), 33–58.

De Luca G., Peracchi F. (2012). Estimating Engel curves under unit and item nonresponse. *Journal of Applied Econometrics*, 27 (7), 1076–1099.

Gallant A., Nychka D. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55 (2), 363–390.

Joe H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97, 2177–2189.

Klein R., Spady R. (1993). An efficient semiparametric estimator of the binary choice model. *Econometrica*, 61 (2), 387–421.

Nadaraya E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9 (1), 141–142.

Newey W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12, 217–229.

Pigini C. (2015). Bivariate non-normality in sample selection model. *Journal of Econometric Methods*, 4 (1), 123–144.

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

Racine J. S. (2008). Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics*, 3 (1), 1–88.

Robinson P. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56 (4), 931–954.

Vella F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, 33 (1), 127–169.

Watson G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26 (4), 359–372.

Received 04.02.2020; accepted 06.03.2020.