Прикладная эконометрика, 2017, т. 45, с. 116–135.
Applied Econometrics, 2017, v. 45, pp. 116–135.

Е. В. Горбунова, В. В. Ульянов, К. К. Фурманов[1]

Построение модели выбытия студентов по данным университетов с разной периодичностью рубежного контроля

При регрессионном анализе выбытия студентов по данным нескольких университетов исследователь сталкивается с проблемой соединения данных с разной периодичностью отчислений в рамках одной модели. В статье предлагается подход к учету разной периодичности в рамках анализа наступления событий в дискретном времени и приводится пример оцененной модели выбытия.

Ключевые слова: анализ наступления событий; дискретный риск; выбытие студентов.
JEL classification: C41; I29.

1. Введение

В данной работе рассматривается проблема, возникшая в ходе изучения факторов выбытия студентов из американских вузов. Эмпирическая задача состояла в построении единой статистической модели с применением анализа наступления событий (далее — АНС) по данным восьми вузов. Однако ее практическая реализация осложнялась тем, что в рассматриваемых университетах различаются периоды, в которые фиксируется интересующее событие. В трех университетах учебный год состоит из трех триместров — осенний, зимний и весенний, тогда как в остальных пяти университетах учебный год состоит из двух семестров — осенний и весенний. Длительность каждого триместра приблизительно равна трем месяцам, тогда как длительность каждого семестра составляет четыре с половиной месяца. Поскольку предпосылкой использования моделей АНС является единая размерность временной шкалы, в которой событие фиксируется для каждого индивида, возникла необходимость разработки способа, позволяющего совместить данные с разной периодичностью. Насколько известно авторам, данная проблема впервые решается в рамках анализа дискретных процессов наступления событий. Отсутствие отечественных и зарубежных разработок по данной теме обусловлено тем, что в большинстве работ, применяющих анализ наступления событий, использовались данные с одинаковой периодичностью.

[1] **Горбунова Елена Васильевна** — Национальный исследовательский университет «Высшая школа экономики», Москва; e.gorbunova88@gmail.com.

Ульянов Владимир Васильевич — Национальный исследовательский университет «Высшая школа экономики», Москва; vulyanov@hse.ru.

Фурманов Кирилл Константинович — Национальный исследовательский университет «Высшая школа экономики», Москва; furmach@rbcmail.ru.

Стоит отметить, что методы анализа событий в дискретном времени хорошо разработаны и известны, а объединение данных с разной периодичностью может быть осуществлено в рамках дискретной модели пропорциональных рисков (Prentice, Gloeckler, 1978) или ее расширенного варианта с учетом ненаблюдаемой разнородности (Meyer, 1990). Однако эта модель выводится из предпосылки о непрерывном процессе, который лишь в наблюдениях дискретизуется из-за группировки данных (интервального цензурирования). При анализе процесса отчисления, дискретного по существу, дискретная модель пропорциональных рисков формально применима (и будет рассмотрена в настоящей статье), но ее параметры лишены интерпретации, т. к. относятся к несуществующему непрерывному процессу.

Решение задачи сведения данных с разной периодичностью рубежного контроля представляется актуальным, учитывая высокий потенциал административных данных для исследования образовательных результатов студентов и активное развитие межинституциональных исследований, предполагающих использование информации о нескольких вузах.

Следующий раздел настоящей статьи содержит обзор литературы, посвященной статистическому анализу выбытия студентов. Обзор подготовлен с упором на методологические особенности исследований. В разделе 3 приведено описание данных, используемых далее при построении модели выбытия. Раздел 4 содержит основную часть настоящей работы: описание основных дискретных моделей наступления событий и особенностей их применения к данным разной периодичности. Особое внимание уделяется модели пропорциональных шансов: рассматривается модификация модели для динамических регрессоров без ограничений на вид временной зависимости вместе с техническими особенностями оценивания (параметризация, выбор начальных значений). В разделе 5 приведена эмпирическая часть исследования — пример оценивания модели выбытия студентов, опирающейся на теорию академического импульса (Adelman, 1999). Раздел 6 содержит выводы и рекомендации по применению регрессионных моделей наступления событий для анализа выбытия студентов.

2. Обзор литературы

В этом разделе исследования рассматриваются в методологическом аспекте. Читатель, заинтересованный в содержательной части, может обратиться к краткому обзору (Груздев, 2011) или более обширной статье (Melguizo, 2011).

Методологическая база исследований факторов выбытия студентов весьма широка. Во многих работах используются дескриптивные методы анализа и простые регрессионные модели (Adelman, 1999, 2006; Knight, 1994; Knight, Arnold, 2000; Lam, 1999). Распространено и моделирование структурными уравнениями (так называемый путевой анализ — path analysis) для проверки валидности априорных теоретических моделей (Bean, 1980; Cabrera et al., 1992; Braxton et al., 2000). Однако использование перечисленных методов не позволяет проанализировать временной аспект изучаемого феномена, а именно, зависимость выбытия от времени, а также изменение значений регрессоров. В то же время большинство исследователей признавало, что выбытие студентов из вуза является динамическим процессом.

С 1990-х годов для изучения факторов выбытия студентов из вуза начинают применяться методы анализа наступления событий, позволяющие включить в анализ временной аспект (Willett, Singer, 1991). Поскольку история обучения студента фиксируется в дискретных

единицах (семестры, четверти, годы), широкое распространение получили дискретные модели АНС. В работе (DesJardins et al., 1999) была предложена спецификация дискретной модели АНС на основе модели пропорциональных рисков с ненаблюдаемой разнородностью и динамическими коэффициентами. С этого времени использование методов АНС для изучения факторов выбытия студентов становится популярным, поскольку эти модели являются удобными в применении и в наибольшей степени подходящими для анализа панельных данных, которые описывают историю обучения студентов. В дальнейшем модели АНС усложняются, в частности, они применяются в сочетании с многоуровневыми моделями (Bahr, 2009; Chen, 2012), структурными уравнениями (Voelkle, Sander, 2008).

В современных работах по анализу выбытия студентов, помимо моделей АНС, встречаются и другие современные методы анализа данных, в частности, квазиэкспериментальные методы (Melguizo et al., 2011; Agasisti, Murtinu, 2016; Hawley, Chiang 2016), многоуровневое моделирование (Lamote et al., 2013), а также интеллектуальные методы: нейронные сети (Chaplot et al., 2015), деревья решений и случайные леса (Chai, Gibson, 2015). Помимо количественных методов анализа данных, также используются качественные и смешанные методы.

В отечественной научной литературе статистический анализ выбытия представлен мало. Исследования (Чудиновских и др., 2004) и (Донец, 2011) опираются на описательный анализ, в том числе методами АНС (оценки функций дожития). В работе (Колотова, 2011), помимо дескриптивной статистики и сравнения средних, используется регрессия Кокса, в статье (Горбунова, 2013) — структурные уравнения. Во всех этих случаях не стояла проблема сопоставления данных с различной периодичностью рубежного контроля, т. к. выборка составлялась по данным одного вуза.

Среди работ, опирающихся на количественный анализ факторов выбытия из вуза, используются как опросные, так и административные данные. В то время как использование опросных данных позволяет существенно расширить набор изучаемых факторов и представить более полную концептуальную модель, их недостатком является существенное сужение исследуемой совокупности, а также смещения в результатах, вызванные как ошибками выборки, так и ошибками измерения (например, относительно успеваемости студента или времени его выбытия). Использование административных данных позволяет минимизировать ошибки измерения по важным для анализа признакам, а также исследовать полную совокупность студентов, однако существенно ограничивает набор изучаемых факторов. Встречаются статьи, в которых используются только опросные данные (например (Braxton et al., 2000)), в ряде работ используются только административные данные (DesJardins et al., 1999; Ishitani, 2003). Есть исследования, в которых совмещаются опросные и административные данные в пределах одного вуза (Cabrera et al., 1992) или данные национальных обследований (Ishitani, 2006).

В работах, использующих для анализа времени выбытия административные данные по нескольким вузам, не возникало проблемы совмещения разной периодичности учебного года, поскольку выбытие либо изначально фиксировалось в годовых интервалах (например (Chen, 2012)), либо совокупность изучаемых вузов имела одинаковую периодичность учебного года. Так, в статье (Bahr, 2009) использовались данные по 109 колледжам Калифорнии, но во всех из них история обучения фиксировалась в семестрах. Тем не менее, в одной из работ совокупность изучаемых вузов имела разную периодичность учебного года — семестры и триместры (Chiang, 2012). В этом исследовании на этапе анализа данных записи об истории обучения были приведены к годовым интервалам.

Приведение к годовым интервалам представляется простым и разумным методом решения задачи, когда анализируемый период продолжителен — в работе (Chiang, 2012) он составлял 8 лет. При этом, однако, происходит потеря потенциально полезной информации и возникает проблема с учетом изменчивости объясняющих переменных, которые могли принимать разные значения на протяжении года. Настоящее исследование опирается на данные с периодом менее трех лет, что делает агрегацию до года особенно нежелательной.

В работе (Горбунова, 2016) рассматриваются три варианта решения проблемы совмещения данных об истории обучения студентов в семестровой и триместровой периодичности: агрегирование до года, интерполяция до интервала в полтора месяца, сведение семестровой системы к триместровой с использованием распределений вероятностей наступления событий. Процедура агрегирования является наиболее удобной и интуитивно понятной, однако приводит к потере информации. Два других подхода позволяют сохранить детальность рассматриваемых признаков, но имеют свои ограничения. Во-первых, данные с измененной периодичностью являются условными, не соответствующими в точности описываемому объекту, которому навязывается несвойственная ему временная шкала, а во-вторых, остается проблема учета динамических объясняющих переменных.

Основная идея настоящей статьи состоит в том, что вместо приведения данных к единой периодичности перед оцениванием единой статистической модели целесообразно саму модель разрабатывать с учетом этой особенности данных.

3. Данные

Эмпирическая часть работы опирается на административные данные по восьми вузам одного из американских штатов, содержащие информацию о поступлении студентов в вуз, их демографических характеристиках, получении финансовой помощи, учебном плане, академических успехах, траекториях обучения и т. д. Университеты отбирались по следующим признакам: являются государственными, имеют селективную систему отбора (характеризуются наличием конкурса для поступления), не являются филиалом вуза, предоставляют образовательные услуги по освоению программ бакалавриата. Анализ проводится по студентам, поступившим на четырехлетние программы обучения в бакалавриате в 2007 году на «полный день», что составляет 25 339 человек. Выборку составляют студенты «традиционного типа», т. е. в возрасте от 18 до 24 лет, зачисленные на очную форму обучения, обучающиеся в университете впервые.

Анализируется выбытие студентов за период, равный двум учебным годам и осеннему периоду третьего года обучения. Особенностью исследуемого массива данных является то, что в нем отсутствуют данные о точном времени выбытия студента из вуза. Зависимая переменная — выбытие из вуза — конструируется на основе сведений о том, обучался ли студент в конкретном учебном периоде в данном вузе. Студент считается выбывшим из вуза, если он прервал свое обучение в данном вузе на срок более года (непрерывно). Время выбытия определяется как последний учебный период, после которого студент прекратил обучение.

Выбор регрессоров осуществлялся, в первую очередь, согласно теории академического импульса, разработанной К. Адельманом в работах (Adelman, 1999, 2006). «Академический импульс» определялся им как накопление студентом академических ресурсов и «скорость» освоения программы в старших классах школы и во время обучения в университете, особенно на первых курсах.

Опираясь на теорию «академического импульса», эмпирическая часть настоящего исследования изучает связь с выбытием студента из вуза следующих характеристик:

- отсутствие перерыва между окончанием школы и поступлением в вуз;
- отсутствие адаптационных курсов в учебном плане студента на первый период обучения в вузе (осенний семестр или триместр первого года обучения);
- объявление студентом своей специальности в первый период обучения;
- более высокое количество накопленных «кредитов» за первый период обучения (показатель отражает более высокую интенсивность обучения);
- высокий средний балл (GPA) за первый период обучения в вузе.

Кроме того, при оценивании учитываются контрольные переменные: получение студентом финансовой помощи, пол, этничность, возраст на момент поступления и др. Описательная статистика приведена в Приложении 1.

4. Методология

4.1. Основные понятия

Время отчисления («время жизни») студента описывается дискретной случайной величиной T, принимающей значения $t_1,...,t_p$ на интервале, соответствующем периоду обучения в вузе. Значение t_1 принимается в случае, когда студент был отчислен в первом периоде обучения, значение t_p — в случае, когда студент отчисляется в последнем периоде p. Значительная часть студентов доживает до выпуска из учебного заведения, так что $\mathrm{P}(T \le t_p) < 1$.

Распределение времени жизни, как правило, задается с помощью функции дожития (survivor function) или функции риска (hazard function). Функция дожития отражает вероятность того, что студент будет отчислен позднее некоторого срока t:

$$S(t) = \mathrm{P}(T > t).$$

Как и в случае с функцией распределения, есть две традиции: задавать функцию дожития через строгое ($T > t$) и через нестрогое ($T \ge t$) неравенство. Здесь используется первый способ — исключительно для удобства.

Функция риска дискретной случайной величины отражает вероятность отчисления в момент t для студента, который не был отчислен ранее:

$$h(t) = \mathrm{P}(T = t \mid T \ge t).$$

Обе функции однозначно задают закон распределения, так что модель времени отчисления может быть выражена с помощью любой из них. В дальнейшем будет использовано следующее выражение, связывающее функцию риска с функцией дожития:

$$h(t_j) = \frac{\mathrm{P}(T = t_j)}{\mathrm{P}(T \ge t_j)} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}. \quad (1)$$

Для всех значений аргумента, не входящих в множество возможных значений T, риск равен нулю, поэтому далее функция риска рассматривается только в точках $t_1,...,t_p$.

4.2. Регрессионные модели длительности с logit- и cloglog-связками

Стандартный способ учета регрессоров при анализе событий в дискретном времени состоит в построении модели, сводимой к одной из распространенных моделей бинарного выбора — чаще всего для этого используются cloglog- и logit-связки (cloglog and logit link functions — см. (Jenkins, 1995)). Модель с cloglog-связкой, далее — модель CL:

$$h(t_j; x, \beta, \alpha) = 1 - \exp(-\exp(x'\beta + g(t_j; \alpha))).$$

Здесь x — вектор регрессоров, β — вектор коэффициентов при этих регрессорах, $g(t;\alpha)$ — функция, отражающая временную зависимость (duration dependence — связь вероятности прекращения состояния с продолжительностью t пребывания в нем), α — вектор параметров этой функции.

Модель с logit-связкой (далее — LL):

$$\frac{h(t_j; x, \beta, \alpha)}{1 - h(t_j; x, \beta, \alpha)} = \exp(x'\beta + g(t_j; \alpha)).$$

Cloglog-связка удобна в тех случаях, когда дискретность времени жизни есть следствие группировки, агрегации первичных данных, а на самом деле время непрерывно, но фиксируется наблюдателем с точностью до принадлежности какому-либо интервалу. В такой ситуации можно предположить, что распределение ненаблюдаемого непрерывного времени описывается моделью пропорциональных рисков (Cox, 1972) с тем же вектором коэффициентов β, что и в модели cloglog, описывающей наблюдаемые дискретные данные (Prentice, Gloeckler, 1978).

Предпосылка о существовании непрерывной величины, стоящей за фиксируемыми в данных длительностями, очень удобна. Во-первых, коэффициенты модели приобретают интерпретацию — их потенцированные значения соответствуют отношениям риска (hazard ratios) в непрерывной модели. Во-вторых, решается проблема разной частотности. Различия в наблюдаемых длительностях связаны не с процессом, порождающим непрерывные величины, а со сбором информации. Используя специальную терминологию, можно сказать, что они вызваны особенностями цензурирования данных. Методы анализа данных с разными видами цензурирования хорошо известны и описаны в (Klein, Moeschberger, 2005).

Однако применительно к студентам предпосылка о непрерывности процесса нежелательна. Хотя отчисления могут происходить в любое время, большинство из них связано с рубежным контролем, поэтому разумнее моделировать процесс как дискретный по существу. К сожалению, это лишает коэффициенты cloglog-модели интерпретации и возвращает к проблеме сопоставления данных с разной частотностью.

Модель LL предпочтительнее с той точки зрения, что ее коэффициенты интерпретируются независимо от существования или несуществования непрерывного времени отчисления. Потенцированные коэффициенты модели с logit-связкой отражают отношения шансов[2] (odds ratios) отчисления: увеличение переменной x_i на единицу соответствует увеличению

[2] Шансы (odds) события — отношение вероятности того, что событие наступит, к вероятности того, что оно не наступит. Например, шансы отчисления в периоде t при условии дожития до этого периода равны $h(t) / (1 - h(t))$.

шансов отчисления в течение периода наблюдения в $\exp(\beta_i)$ раз при неизменных значениях остальных регрессоров. К сожалению, привязка к периоду наблюдения имеет нежелательное следствие: коэффициенты в моделях, оцененных по данным разной частотности, оказываются несопоставимыми. Одно и то же отношение шансов отчисления в течение года соответствует разным отношениям шансов отчисления в отдельный период обучения для семестровых и триместровых университетов. И наоборот, одно и то же значение коэффициента модели LL приводит к разным отношениям шансов отчисления за год для разных университетов. Для иллюстрации приведем пример.

Пример несопоставимости коэффициентов в модели с logit-связкой. Предположим, что процесс отчисления студентов описывается схемой Бернулли: риск выбытия одинаков во всех периодах обучения $h(t) = h$. Рассмотрим университет, в котором учебный год поделен на семестры, так что вероятность выбытия в течение года p связана с вероятностью выбытия в отдельном семестре h соотношением $p = h + (1-h)h$ — либо отчисление происходит в первом семестре (вероятность этого равна h), либо в первом семестре студент остается, а во втором выбывает (вероятность $(1-h)h$). То же соотношение можно записать иначе: $h = 1 - \sqrt{1-p}$. Здесь $(1-p)$ — вероятность «пережить» учебный год, а $\sqrt{1-p}$ — вероятность не быть отчисленным в течение семестра.

Пусть в университете обучаются две группы студентов: A и B. Среди группы A вероятность выбытия в течение года составляет 0.05, а среди группы B — 0.15. Риск выбытия за один семестр в первой группе равен $1 - \sqrt{1-0.05} = 0.025$, а во второй $1 - \sqrt{1-0.15} = 0.078$.

Логарифм шансов выбытия за семестр в группе A равен

$$\ln(0.025 / (1-0.025)) = -3.650,$$

в группе B:

$$\ln(0.078 / (1-0.078)) = -2.469.$$

Логарифм отношения шансов:

$$-2.469 - (-3.650) = 1.181.$$

Получаем выражение для риска выбытия с помощью logit-связки:

$$h/(1-h) = \exp(-3.650 + 1.181 \cdot B),$$

где B — индикатор принадлежности к группе B.

Теперь представим, что группы с теми же вероятностями выбытия в течение года обучаются в университете с обучением по триместрам. Риск выбытия в течение триместра связан с вероятностью выбытия за год соотношением $h = 1 - \sqrt[3]{1-p}$, так что вероятности выбытия за один период теперь будут равны $1 - \sqrt[3]{1-0.05} = 0.017$ в группе A и $1 - \sqrt[3]{1-0.15} = 0.053$ в группе B. Проведя те же вычисления, что и в предыдущем абзаце, получаем логарифм шансов выбытия за триместр в группе A, равный –4.060, и логарифм отношения шансов 1.172, так что риск выбытия описывается моделью $h/(1-h) = \exp(-4.060 + 1.172 \cdot B)$. То есть при одинаковой зависимости вероятности отчисления за год от индикатора B величина коэффициентов меняется — это значит, что оценки, полученные по данным разной периодичности, несопоставимы. Можно показать, что модель CL лишена этого недостатка.

Ее коэффициенты, кроме свободного члена, определяют характеристики непрерывного процесса и не связаны с периодичностью наблюдений.

На самом деле, если вероятности отчисления невелики, как обычно и бывает, то расхождение в коэффициентах logit при регрессорах пренебрежимо мало (в рассмотренном примере коэффициенты равны 1.181 и 1.172), а свободный член часто не представляет интереса для исследования, так что logit-связку можно считать практически применимой. Тем не менее, можно предложить модель, лишенную этого недостатка и имеющую преимущество с точки зрения интерпретации параметров.

4.3. Модель пропорциональных шансов

Эта модель (proportional odds model, далее — модель PO (McCullagh, 1980; Bennett, 1983)) опирается на предположение, что объясняющие переменные пропорционально связаны с шансами отчисления студента на временном отрезке любой длины:

$$\frac{1-S(t;x,\beta,\alpha)}{S(t;x,\beta,\alpha)}=\exp(x'\beta+g(t;\alpha)). \tag{2}$$

Отсюда выводится выражение для функции дожития:

$$S(t;x,\beta,\alpha)=\frac{1}{1+\exp(x'\beta+g(t;\alpha))}.$$

Функция $g(t;\alpha)$ задает логарифм опорных шансов (baseline odds) — шансов отчисления в случае равенства нулю всех регрессоров. Как и в рассмотренных ранее моделях, эта функция определяет характер временной зависимости, но в данной модели на нее накладывается ограничение: $g(t;\alpha)$ должна быть неубывающей по t, в противном случае убывающей окажется функция дожития, что невозможно. Обычно модель пропорциональных шансов рассматривается в непрерывном времени, но выражение (2) может задавать и дискретную модель — в этом случае функция $g(t;\alpha)$ будет кусочно-постоянной по t. Интерпретация коэффициентов такова: увеличение переменной x_i на единицу соответствует увеличению шансов отчисления в $\exp(\beta_i)$ раз при неизменных значениях остальных регрессоров. При этом период, в течение которого происходит либо не происходит отчисление, может быть любым — модель предполагает, что отношение шансов одинаково и для отчисления в первом семестре (триместре), и для отчисления за весь период обучения. В отличие от модели с logit-связкой, модель (2) опирается не на шансы отчисления в отдельный период времени $h(t)/(1-h(t))$, а на шансы отчисления вплоть до времени t, равные $(1-S(t))/S(t)$.

В случае неизменных во времени регрессорах коэффициенты β можно оценить, не накладывая ограничений на опорные шансы. При дискретном времени отчисления модель (2) — это обычная порядковая logit-регрессия. Чтобы позволить объясняющим переменным изменяться во времени, получим выражение для функции риска, опираясь на формулу (1):

$$h(t_j;x,\beta,\alpha)=1-\frac{S(t_j;x,\beta,\alpha)}{S(t_{j-1};x,\beta,\alpha)}=1-\frac{1+\exp(x'\beta+g(t_{j-1};\alpha))}{1+\exp(x'\beta+g(t_j;\alpha))},\quad j>1; \tag{3}$$

$$h(t_1;x,\beta,\alpha)=P(T=t_1)=1-S(t_1)=\frac{\exp(x'\beta+g(t_1;\alpha))}{1+\exp(x'\beta+g(t_1;\alpha))}. \tag{4}$$

В это выражение можно подставлять свой набор объясняющих переменных для каждого момента времени, в то время как подставлять их сразу в функцию дожития некорректно, потому что вероятность дожития до некоторого времени зависит от значений переменных в разные моменты.

Так как число возможных моментов отчисления конечно и невелико по сравнению с числом наблюдений, можно не накладывать ограничений на функцию $g(t;\alpha)$, задавая ее следующим образом:

$$g(t_1;\alpha) = g_1, \quad g(t_j;\alpha) = g_1 + \sum_{i=2}^{j} \exp(\theta_i).$$

Благодаря такой спецификации вектор параметров $\alpha = (g_1, \theta_2, ..., \theta_p)$ может принимать любые значения, при этом функция будет оставаться неубывающей (по *t*). Единственное накладываемое при этом ограничение — опорные шансы должны расти в каждый из моментов $t_1, ..., t_p$, т. к. $\exp(\theta_i) > 0$. Это означает, что вероятность отчисления в каждом периоде не равна нулю. В промежутках между возможными значениями t_{j-1} и t_j функция g постоянна. Как видно из выражения для функции риска, для идентифицируемости параметра g_1 линейная комбинация ковариат $x'\beta$ не должна содержать свободный член, либо свободный член можно оставить, а значение g_1 зафиксировать (естественно, положить $g_1 = 0$).

Векторы параметров β и $\alpha = (g_1, \theta_2, ..., \theta_p)$ можно оценить методом максимального правдоподобия. Задачу максимизации можно упростить, выбрав «хорошие» начальные условия. Так как значение функции риска в первый момент задается обычной logit-моделью, можно получить предварительную оценку вектора β из logit-регрессии, в которой объясняемая переменная — индикатор отчисления студента в первом периоде, при этом данные остальных периодов не используются. Полученная оценка свободного члена будет служить начальным значением для g_1.

Основным преимуществом использования модели PO является интерпретация параметров, не привязанная к периодичности.

4.4. Объединение данных с разной периодичностью

Основная идея, предлагаемая в настоящей статье, заключается в том, что принадлежность студентов к университетам с разной схемой обучения может быть учтена с помощью задания функции $g(t;\alpha)$ отдельно для семестровых и для триместровых университетов по аналогии со стратифицированной моделью Кокса (см., например, (Ata, Sözer, 2007)). Во всех рассмотренных моделях время учитывается только в указанной функции, поэтому параметры, отвечающие за связь риска с объясняющими переменными, можно считать непривязанными к периодичности рубежного контроля, если пренебречь проблемой несопоставимости коэффициентов модели LL, рассмотренной в п. 4.2.

Для удобства обозначения будем считать длительностью обучения число прошедших периодов: $t_j = j$. Таким образом, максимальное время обучения в университетах с разной периодичностью будет отличаться, а сопоставление моментов отчисления студентов будет требовать поправки, но на модели это никак не скажется, так как функции временной зависимости будут различными для университетов разных типов.

Пусть y_{it} — индикатор отчисления студента i в период обучения t ($y_{it}=1$, если студент был отчислен, 0 иначе), x_{it} — вектор объясняющих переменных, характеристик студента и университета, $trim_i$ — индикатор периодичности ($trim_i=1$, если студент i обучается в университете с триместровой системой, $trim_i=0$, если обучение разбито на семестры), $sem_i=1-trim_i$.

В моделях LL и CL будем учитывать временную зависимость с помощью фиктивных переменных. Пусть z_t^{trim} — вектор индикаторов временных периодов для университетов с обучением по триместрам, z_t^{sem} — для «семестровых» университетов. Модель с logit-связкой имеет следующий вид:

$$\frac{\mathrm{P}(y_{it}=1)}{\mathrm{P}(y_{it}=0)}=\exp(x_{it}'\beta+z_t^{trim}trim_i'\gamma+z_t^{sem}sem_i'\delta)\,. \tag{5}$$

Модель с cloglog-связкой:

$$\mathrm{P}(y_{it}=1)=1-\exp(-\exp(x_{it}'\beta+z_t^{trim}trim_i'\gamma+z_t^{sem}sem_i'\delta))\,. \tag{6}$$

В обоих случаях β,γ и δ — векторы оцениваемых коэффициентов, причем β не включает свободный член — он учитывается отдельно для разных типов университетов в векторах γ и δ.

Таким образом, при объединении данных с разной периодичностью предполагается, что коэффициенты при объясняющих переменных одинаковы, а временная зависимость может полностью различаться для двух типов университетов. С одной стороны, различный вид временной зависимости позволяет избежать проблем, связанных с разной периодичностью (не нужно подгонять данные к единому виду: агрегировать или разбивать на условные подпериоды). С другой стороны, совпадение коэффициентов при регрессорах позволяет оценивать единую модель для всех университетов, не разделяя данные на две части по несущественному признаку и способствуя разумной редукции: иметь меньший набор параметров удобнее для интерпретации.

Легко оценить модели CL и LL при рассмотренной организации данных (одно наблюдение соответствует одному периоду обучения одного студента, в данных присутствует индикатор отчисления): это делается стандартными командами оценки моделей бинарного выбора.

С учетом выбранной параметризации функции $g(t;\alpha)$ и различий в периодичности выражения (3) и (4) приобретают вид:

$$\mathrm{P}(y_{it}=1)=1-\frac{1+\exp\left[x_{it}'\beta+trim_i\left(g_1^{trim}+\sum_{j=2}^{t-1}\exp(\theta_j^{trim})\right)+sem_i\left(g_1^{sem}+\sum_{j=2}^{t-1}\exp(\theta_j^{sem})\right)\right]}{1+\exp\left[x_{it}'\beta+trim_i\left(g_1^{trim}+\sum_{j=2}^{t}\exp(\theta_j^{trim})\right)+sem_i\left(g_1^{sem}+\sum_{j=2}^{t}\exp(\theta_j^{sem})\right)\right]},\ t>1; \tag{7}$$

$$\mathrm{P}(y_{i1}=1)=\frac{\exp(x_{i1}'\beta+trim_i g_1^{trim}+sem_i g_1^{sem})}{1+\exp(x_{i1}'\beta+trim_i g_1^{trim}+sem_i g_1^{sem})}. \tag{8}$$

Как и ранее, модель для разных университетов отличается временной зависимостью. Параметры временной зависимости (функции опорных шансов) снабжены индексами *trim* и *sem* для триместровых и семестровых данных соответственно.

Оценивание параметров проводилось максимизацией функции правдоподобия в программе *Stata 11*, использовался метод Ньютона–Рафсона с численным расчетом производных. Начальные значения параметров β, g_1^{trim} и g_1^{sem} брались из logit-регрессии для данных первого периода наблюдения, оцениваемая регрессия соответствует формуле (8).

5. Пример оцененной модели выбытия

Все три рассмотренных типа моделей были оценены с помощью программы *Stata* (версии 11), для оценивания моделей CL и LL использовались встроенные команды, в случае PO использовался авторский модуль, опирающийся на реализованный в *Stata* алгоритм максимизации функции правдоподобия (при максимизации применялся алгоритм Ньютона–Рафсона с численным расчетом производных).

В таблице 1 приведены основные (с точки зрения теории академического импульса) статистически значимые параметры. Более полно результаты оценивания описаны в Приложении 2. Одна теоретически важная детерминанта — наличие перерыва между окончанием школы и поступлением в вуз — оказалась незначимой, несмотря на большое число наблюдений. Речь идет не только о статистической незначимости, т. е. о недостатке оснований для отвержения гипотезы об отсутствии связи. Оценка коэффициента при этой переменной близка к нулю, а стандартная ошибка довольно мала, так что причина незначимости состоит именно в отсутствии существенной связи между фактом отчисления и наличием временного зазора перед поступлением, а не в невозможности надежно установить эту связь исходя из имеющихся данных. То же самое касается и остальных незначимых коэффициентов — в данном случае можно считать, что соответствующие факторы незначимы не только статистически, но и практически.

Таблица 1. Оценки некоторых коэффициентов моделей LL, CL и PO

Переменная	LL		CL		PO	
	β	exp (β)	β	exp (β)	β	exp (β)
Наличие адаптационных курсов в первый период обучения	0.251	1.285	0.204	1.226	0.278	1.320
Число накопленных за первый период обучения кредитов (базовая категория — менее 15)						
≥17 кредитов	–0.268	0.765	–0.239	0.787	–0.299	0.742
[15; 17)	–0.136	0.872	–0.113	0.893	–0.170	0.844
Средний балл за первый период обучения (6 групп, базовая — 4)						
1 группа (низшие баллы)	2.108	8.231	1.778	5.918	2.825	16.861
2 группа	1.319	3.740	1.198	3.313	1.602	4.962
3 группа	0.641	1.898	0.596	1.815	0.749	2.115
5 группа	–0.484	0.616	–0.470	0.625	–0.539	0.583
6 группа (высшие баллы)	–0.755	0.470	–0.739	0.478	–0.832	0.435

Примечание. Все приведенные оценки значимы на уровне 0.1%.

Коэффициенты несопоставимы между моделями, но видно, что ранжировки коэффициентов по величине в разных моделях совпадают. Наличие адаптационных курсов оказывается наименее важным фактором, чуть в большей степени риск выбытия связан с числом кредитов и в намного большей мере — с успеваемостью в первом периоде.

Приведем пример интерпретации потенцированных коэффициентов. Из оценок модели LL следует, что шансы отчисления *в течение одного периода обучения* у студентов, проходивших адаптационные курсы, в 1.29 раз больше, чем у тех, кто не включал адаптационные курсы в индивидуальный учебный план (при прочих равных условиях). Из оценок модели PO следует, что шансы отчисления *в течение всего времени обучения* у студентов, выбравших адаптационные курсы, в 1.32 раза больше (при прочих равных и неизменных условиях). На самом деле, такое же соотношение будет выполняться и для любого другого срока, но именно возможность оценить отношение шансов для всего времени обучения делает модель PO особенно привлекательной. Впрочем, в настоящем случае речь идет лишь о студентах, окончивших первый период обучения: часть объясняющих переменных характеризует этот период, поэтому как выборка, так и генеральная совокупность включают только студентов после первого семестра или триместра.

Существенно иной будет интерпретация коэффициентов модели CL: при наличии адаптационных курсов риск отчисления в 1.23 раза больше, чем при их отсутствии, если остальные регрессоры не отличаются. При этом речь идет о функции риска для непрерывных случайных величин, которая определяется соотношением $h(t) = \lim_{\Delta \to 0} \mathrm{P}(t < T \le t + \Delta \,|\, T > t)/\Delta$, так что ее значение равно условной плотности величины T, а не условной вероятности, как в дискретном случае. С точки зрения авторов настоящей статьи, интерпретация в терминах функции плотности менее ясна, чем в терминах шансов, что можно считать дополнительным доводом в пользу модели PO, помимо отсутствия необходимости рассматривать процесс выбытия как непрерывный.

В пользу модели пропорциональных шансов говорит и качество подгонки, измеряемое значением функции правдоподобия, на втором месте стоит модель LL, так что в настоящем примере интерпретируемость и точность описания данных оказываются согласованными. Для сравнения подгонки в анализе событий обычно используется информационный критерий Акаике, но в случае одинакового числа оцениваемых параметров он приводит к тому же результату, что и значение функции правдоподобия. Это преимущество можно считать второстепенным и привязанным к конкретному набору данных: при изучении выбытия осмысленность результатов представляется более важной, чем качество подгонки. Превосходство правдоподобия для модели PO свидетельствует, видимо, о том, что различия в риске выбытия между разными категориями студентов со временем уменьшаются, что является одним из свойств, отличающих эту модель от LL и CL.

Полученные оценки коэффициентов при переменных, отражающих средний балл в первом периоде обучения, могут показаться неправдоподобными, т. к. отношение шансов отчисления между крайними группами очень велико: 16.861/0.435 = 38.76, что намного превосходит отношения, связанные с другими регрессорами. Однако полученные оценки вполне согласуются с частотой отчислений в группах с разной успеваемостью (табл. 2), построенной по учтенным при оценивании наблюдениям (здесь одному студенту соответствует набор наблюдений).

Как видно из таблицы, доля отчислений в первой группе в 22 раза выше, чем в шестой. Соответствующее отношение шансов, рассчитанное по данным этой таблицы, оказывается равным 33.6, что по порядку вполне соответствует оценкам регрессионной модели.

Таблица 2. Частота отчислений в группах по среднему баллу за первый период обучения

Группа	1	2	3	4	5	6
Доля отчислений, %	35.7	17.7	9.1	4.4	2.4	1.6

6. Заключение

В настоящей работе рассмотрена задача регрессионного анализа выбытия студентов по данным университетов с разной периодичностью рубежного контроля. Для решения этой задачи предложено использовать модели наступления событий со стратификацией по типу университета, выделяя страты с разным характером временной зависимости (в рассмотренном примере в первую страту входили университеты, где обучение поделено на семестры, во вторую — триместровые университеты). Преимущества такого подхода по сравнению с приведением данных к единой периодичности таковы:

- нет потери информации, которая происходит при агрегации данных до года;
- не создаются фиктивные наблюдения, которые возникают при искусственном разбиении относительно бóльших периодов на малые;
- не возникает трудностей при учете изменчивых во времени регрессоров.

Учет большего числа страт формально не составляет труда, хотя предполагает введение в модель дополнительных параметров по числу возможных значений продолжительности обучения в новой страте. Возможно, имеет смысл более подробная стратификация, при которой каждый университет (или даже факультет) выделяется в отдельную страту. Но это предполагает введение в модель дополнительных параметров по числу возможных значений продолжительности обучения в каждой дополнительной страте и может быть связано с вычислительными трудностями.

Популярные модели риска, опирающиеся на logit- и cloglog-связки, применимы к анализу выбытия студентов, но обладают недостатками:

- коэффициенты cloglog-модели не имеют интерпретации;
- интерпретация коэффициентов logit-модели привязана к периоду обучения, одно и то же значение коэффициента соответствует разной по величине связи вероятности (или шансов) выбытия с объясняющей переменной для разных периодичностей обучения.

Этих недостатков можно избежать при применении модели пропорциональных шансов, но процедуру ее оценивания при наличии изменяющихся во времени регрессоров и стратификации исследователю, возможно, придется реализовывать самостоятельно (напрямую применить стандартное программное обеспечение здесь невозможно). Если получение интерпретируемых коэффициентов не важно, эта модель не имеет существенных преимуществ по сравнению со стандартными, предусмотренными создателями статистических программ.

Хотя настоящее исследование опирается на американские данные, стоит отметить, что и среди российских университетов есть подобное различие: наряду с традиционной семестровой системой обучения существует модульная система с частым рубежным контролем и, соответственно, выбытием. По этой причине настоящая работа может быть интересна исследователям отечественного образования.

Список литературы

Горбунова Е. В. (2013). Влияние адаптации первокурсников к университету на вероятность их отчисления из вуза. *Universitas*, 1 (2), 59–84.

Горбунова Е. В. (2016) Сравнение подходов к совмещению данных с разной периодичностью в анализе наступления событий. В кн.: *Труды 7-й Международной научно-практической конференции студентов и аспирантов «Статистические методы анализа экономики и общества»* (17–20 мая 2016 г.). Национальный исследовательский университет «Высшая школа экономики», 88–89.

Груздев И. (2011). Зарубежный опыт исследований отчисленных студентов. *Мониторинг Университета*, 6, 7–10.

Донец Е. (2011). Опыт исследования студенческих отчислений на примере МГУ. *Мониторинг Университета*, 6, 33–38.

Колотова Е. (2011). Изучение отчислений студентов в бакалавриате/специалитете НИУ ВШЭ. *Мониторинг Университета*, 6, 22–32.

Чудиновских О. С., Телешова И. Г., Донец Е. В. (2004). *Возможности и ограничения завершения высшего образования в элитном вузе* (*на примере Московского государственного университета им. М. В. Ломоносова*). М.: МАКС Пресс.

Adelman C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U. S. Department of Education.

Adelman C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U. S. Department of Education.

Agasisti T., Murtinu S. (2016). Grants in Italian university: A look at the heterogeneity of their impact on students' performances. *Studies in Higher Education*, 41 (6), 1106–1132.

Ata N., Sözer M. T. (2007). Cox regression models with nonproportional hazards applied to lung-cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, 36 (2), 157–167.

Bahr P. R. (2009). Educational attainment as process: Using hierarchical discrete-time event history analysis to model rate of progress. *Research in Higher Education*, 50 (7), 691–714.

Bean J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12 (2), 155–187.

Bennett S. (1983). Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C*, 32 (2), 165–171.

Braxton J. M., Milem J. F., Sullivan A. S. (2000). The influence of active learning on the college student departure process toward a revision of Tinto's theory. *Journal of Higher Education*, 71 (5), 569–590.

Cabrera A. F., Nora A. L., Castaneda M. B. (1992). The role of finances in the persistence process: A structural model. *Research in Higher Education*, 33 (5), 571–593.

Chaplot D. S., Rhim E., Kim J. (2015). Predicting student attrition in MOOCs using sentiment analysis and neural networks. In: *Proceedings of AIED 2015 Fourth Workshop on Intelligent Support for Learning in Groups*. http://ceur-ws.org/Vol-1432/islg_proc.pdf.

Chai K. E. K., Gibson D. (2015). Predicting the risk of attrition for undergraduate students with time based modelling. In: *12th International Conference on Cognition and Exploratory Learning in Digital Age* (*CELDA 2015*). http://files.eric.ed.gov/fulltext/ED562154.pdf.

Chen R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher Education*, 53 (5), 487–505.

Chiang S. C. (2012). Applying event history analysis to investigate the impacts of developmental education on emerging adults' degree completion. Dissertation, The Ohio State University. https://etd.ohiolink.edu/rws_etd/document/get/osu1331061887/inline.

Cox D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, *Series B*, 34, 187–220.

DesJardins S. L., Ahlburg D. A., McCall B. P. (1999). An event history model of student departure. *Economics of Education Review*, 18, 375–390.

Ishitani T. T. (2003). A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college characteristics. *Research in Higher Education*, 44 (4), 433–449.

Ishitani T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *Journal of Higher Education*, 77 (5), 861–885.

Jenkins S. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics*, 57, 129–138.

Klein J. P., Moeschberger M. L. (2005). *Survival analysis. Techniques for censored and truncated data. Second Edition*. Springer.

Lamote C., van Damme J., van den Noortgate W., Speybroeck S., Boonen T., de Bilde J. (2013). Dropout in secondary education: An application of a multilevel discrete-time hazard model accounting for school changes. *Quality and Quantity*, 47 (5), 2425–2446.

McCullagh P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 42 (2), 109–142.

Melguizo T. (2011). A review of the theories developed to describe the process of college persistence and attainment. In: *Higher Education: Handbook of Theory and Research, Vol. 26*, 395–424. Springer.

Melguizo T., Kienzl G., Alfonso M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *The Journal of Higher Education*, 82 (3), 265–291.

Meyer B. D. (1990). Unemployment insurance and unemployment spells. *Econometrica*, 58 (4), 757–782.

Prentice R. L., Gloeckler L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34, 57–67.

Willett J. B., Singer J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research*, 61 (4), 407–450.

Voelkle M. C., Sander N. (2008). University dropout: A structural equation approach to discrete-time survival analysis. *Journal of Individual Differences*, 29 (3), 134–147.

Поступила в редакцию 01.12.2016;
принята в печать 18.02.2017.

Приложение 1. Описательные статистики

Переменная	Расшифровка	Доля (среднее), %
Dropout	Выбытие	4.0
Hs_gap	Перерыв между окончанием школы и поступлением в вуз от года и более	2.1
Took_rem_courses	Наличие в учебном плане студента за первый период обучения адаптационных курсов	13.1
Age_entry	Возраст на момент поступления (от 18 до 24 лет)	18.4
Female	Женский пол	52.0
Housing	Проживание в общежитии в первый период обучения	82.4
No_major_first	Студент не объявил специальность в первый период обучения	9.1
Need_grant	Получает финансовую помощь по причине низкого дохода	21.8
Merit_grant	Получает финансовую помощь за выдающиеся успехи (в обучении или других сферах)	38.4
Loan	Получает займ на образование	47.8
Work_study	Получает финансовую помощь взамен работы в вузе	3.1
	Этничность	
	Базовая категория	85.3
Black	Афроамериканцы	7.9
Hispanic	Латиноамериканцы	2.2
Asian	Азиаты	2.9
Other	Иностранные студенты	1.7
	Количество накопленных кредитов за первый период обучения	
	<15 (базовая категория)	20.5
Crhrs_groups_15_17	[15; 17)	58.6
Crhrs_groups_17_over	≥ 17	20.9
	Средний балл за первый период обучения в вузе	
Gpa_group_1	$= 0$	1.3
Gpa_group_2	(0, 1.5)	4.4
Gpa_group_3	[1.5; 2)	5.2
Gpa_group_4	[2; 3)	32.0
Gpa_group_5	[3; 3.5)	29.0
Gpa_group_6	[3.5; 4]	28.1

Примечание. Доли и средний возраст рассчитаны по наблюдениям, не по студентам.
Число наблюдений — 167 656, число студентов — 25 339.

Приложение 2. Оценки коэффициентов моделей выбытия

Переменная	LL	CL	PO
Hs_gap	–0.043	–0.081	0.016
	(0.081)	(0.072)	(0.095)
Took_rem_courses	0.251***	0.204***	0.278***
	(0.036)	(0.033)	(0.042)
Need_grant	0.173***	0.160***	0.153***
	(0.032)	(0.029)	(0.036)
Merit_grant	−0.265***	−0.244***	−0.295***
	(0.032)	(0.031)	(0.036)
Loan	0.159***	0.148***	0.156***
	(0.030)	(0.027)	(0.034)
Work_study	–0.070	–0.051	–0.058
	(0.077)	(0.072)	(0.087)
Chrhrs_group_15_17	−0.136***	−0.113***	−0.170***
	(0.031)	(0.029)	(0.036)
Chrhrs_group_17_over	−0.268***	−0.239***	−0.299***
	(0.046)	(0.043)	(0.052)
Age_entry	0.054**	0.041**	0.077***
	(0.021)	(0.019)	(0.025)
Female	0.174***	0.145***	0.238***
	(0.027)	(0.024)	(0.031)
Black	−0.165***	−0.179***	−0.200***
	(0.045)	(0.040)	(0.052)
Hispanic	0.166**	0.159**	0.177*
	(0.084)	(0.076)	(0.096)
Asian	–0.102	–0.097	–0.109
	(0.090)	(0.083)	(0.101)
Other	0.209**	0.177**	0.509***
	(0.086)	(0.076)	(0.108)
Housing	−0.161***	−0.137***	−0.154***
	(0.035)	(0.032)	(0.041)
No_major_first	0.107**	0.094**	0.122**
	(0.047)	(0.043)	(0.055)
Gpa_group_1	2.108***	1.778***	2.825***
	(0.059)	(0.04)	(0.078)
Gpa_group_2	1.319***	1.198***	1.602***
	(0.040)	(0.036)	(0.050)
Gpa_group_3	0.641***	0.596***	0.749***
	(0.045)	(0.041)	(0.053)
Gpa_group_5	−0.484***	−0.470***	−0.539***
	(0.038)	(0.036)	(0.041)
Gpa_group_6	−0.755***	−0.739***	−0.832***
	(0.045)	(0.043)	(0.048)
Логарифм правдоподобия	–22921.524	–22954.799	–22817.686

Примечание. LL — модель с logit-связкой, CL — с cloglog-связкой, PO — модель пропорциональных шансов. Число наблюдений — 167 656.

В скобках под оценками приведены стандартные ошибки. *, **, *** — значимость на уровне 10, 5 и 1% соответственно.
В каждую модель также включались фиктивные переменные для учета индивидуального эффекта учебного заведения и для временной зависимости согласно формулам (5)–(8).
Все модели значимы в целом, *p*-значение равно 0 с точностью до четвертого знака.

Gorbunova E. V., Ulyanov V. V., Furmanov K. K. Using data from universities with different structure of academic year to model student attrition. *Applied Econometrics*, 2017, v. 45, pp. 116–135.

Elena Gorbunova
National Research University Higher School of Economics, Moscow, Russian Federation;
e.gorbunova88@gmail.com

Vladimir Ulyanov
National Research University Higher School of Economics, Moscow, Russian Federation;
vulyanov@hse.ru

Kirill Furmanov
National Research University Higher School of Economics, Moscow, Russian Federation;
furmach@rbcmail.ru

Using data from universities with different structure of academic year to model student attrition

Pooling the data from a number of universities into a single sample poses a problem for researchers who are performing regression analysis of student attrition. Academic year can be divided into different academic terms in different universities, and this discrepancy has to be taken into account. This paper considers a problem of using data with different periodicity in the framework of discrete-time event history analysis and gives an example of an estimated attrition model.

Keywords: event-history analysis; discrete hazard; student attrition.

JEL classification: C41; I29.

References

Gorbunova E. V. (2013). Vlijanie adaptacii pervokursnikov k universitetu na verojatnost' ih otchislenija iz vuza. *Universitas*, 1 (2), 59–84 (in Russian).

Gorbunova E. V. (2016). Sravnenie podhodov k sovmeshheniju dannyh s raznoj periodichno-st'ju v analize nastuplenija sobytij. V kn: *Trudy 7-j Mezhdunarodnoj nauchno-prakticheskoj konferen-cii studentov i aspirantov «Statisticheskie metody analiza jekonomiki i obshhestva»* (*17–20 May 2016*). National Research University Higher School of Economics, 88–89 (in Russian).

Gruzdev I. (2011). Zarubezhnyj opyt issledovanij otchislennyh studentov. *Monitoring Universiteta*, 6, 7–10 (in Russian).

Donec E. (2011). Opyt issledovanija studencheskih otchislenij na primere MGU. *Monitoring Universiteta,* 6, 33–38 (in Russian).

Kolotova E. (2011). Izuchenie otchislenij studentov v bakalavriate/specialitete NIU VShJe. *Monitoring Universiteta*, 6, 22–32 (in Russian).

Chudinovskih O. S., Teleshova I. G., Donec E. V. (2004). Vozmozhnosti i ogranichenija zavershenija vysshego obrazovanija v jelitnom vuze (na primere Moskovskogo gosudarstvennogo universiteta im. M. V. Lomonosova). M.: MAKS Press (in Russian).

Adelman C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U. S. Department of Education.

Adelman C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U. S. Department of Education.

Agasisti T., Murtinu S. (2016). Grants in Italian university: A look at the heterogeneity of their impact on students' performances. *Studies in Higher Education*, 41 (6), 1106–1132.

Ata N., Sözer M. T. (2007). Cox regression models with nonproportional hazards applied to lung-cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, 36 (2), 157–167.

Bahr P. R. (2009). Educational attainment as process: Using hierarchical discrete-time event history analysis to model rate of progress. *Research in Higher Education*, 50 (7), 691–714.

Bean J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12 (2), 155–187.

Bennett S. (1983). Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C*, 32 (2), 165–171.

Braxton J. M., Milem J. F., Sullivan A. S. (2000). The influence of active learning on the college student departure process toward a revision of Tinto's theory. *Journal of Higher Education*, 71 (5), 569–590.

Cabrera A. F., Nora A. L., Castaneda M. B. (1992). The role of finances in the persistence process: A structural model. *Research in Higher Education*, 33 (5), 571–593.

Chaplot D. S., Rhim E., Kim J. (2015). Predicting student attrition in MOOCs using sentiment analysis and neural networks. In: *Proceedings of AIED 2015 Fourth Workshop on Intelligent Support for Learning in Groups*. http://ceur-ws.org/Vol-1432/islg_proc.pdf.

Chai K. E. K., Gibson D. (2015). Predicting the risk of attrition for undergraduate students with time based modelling. In: *12th International Conference on Cognition and Exploratory Learning in Digital Age* (*CELDA 2015*). http://files.eric.ed.gov/fulltext/ED562154.pdf.

Chen R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher Education*, 53 (5), 487–505.

Chiang S. C. (2012). Applying event history analysis to investigate the impacts of developmental education on emerging adults' degree completion. Dissertation, The Ohio State University. https://etd.ohiolink.edu/rws_etd/document/get/osu1331061887/inline.

Cox D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, *Series B*, 34, 187–220.

DesJardins S. L., Ahlburg D. A., McCall B. P. (1999). An event history model of student departure. *Economics of Education Review*, 18, 375–390.

Ishitani T. T. (2003). A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college characteristics. *Research in Higher Education*, 44 (4), 433–449.

Ishitani T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *Journal of Higher Education*, 77 (5), 861–885.

Jenkins S. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics*, 57, 129–138.

Klein J. P., Moeschberger M. L. (2005). *Survival analysis. Techniques for censored and truncated data. Second Edition*. Springer.

Lamote C., van Damme J., van den Noortgate W., Speybroeck S., Boonen T., de Bilde J. (2013). Dropout in secondary education: An application of a multilevel discrete-time hazard model accounting for school changes. *Quality and Quantity*, 47 (5), 2425–2446.

McCullagh P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 42 (2), 109–142.

Melguizo T. (2011). A review of the theories developed to describe the process of college persistence and attainment. In: *Higher Education: Handbook of Theory and Research, Vol. 26*, 395–424. Springer.

Melguizo T., Kienzl G., Alfonso M. (2011). Comparing the educational attainment of community college transfer students and four-year college rising juniors using propensity score matching methods. *The Journal of Higher Education*, 82 (3), 265–291.

Meyer B. D. (1990). Unemployment insurance and unemployment spells. *Econometrica*, 58 (4), 757–782.

Prentice R. L., Gloeckler L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34, 57–67.

Willett J. B., Singer J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research*, 61 (4), 407–450.

Voelkle M. C., Sander N. (2008). University dropout: A structural equation approach to discrete-time survival analysis. *Journal of Individual Differences*, 29 (3), 134–147.

Received 01.12.2016; accepted 18.02.2017.