УДК 311.2 + 616-036.22

Д. Зельтерман, А.Е. Пащенко, А.В. Суворова, В.Ф. Мусина,
Т.В. Тулупьева, А.Л. Тулупьев, Л.Е. Гро, Р. Хаймер

ДИАГНОСТИКА РЕГРЕССИОННЫХ УРАВНЕНИЙ В АНАЛИЗЕ ИНТЕНСИВНОСТИ РИСКОВАННОГО ПОВЕДЕНИЯ ПО ЕГО ПОСЛЕДНИМ ЭПИЗОДАМ

Зельтерман Д., Пащенко А.Е., Суворова А.В., Мусина В.Ф., Тулупьева Т.В., Тулупьев А.Л., Гро Л.Е., Хаймер Р. **Диагностика регрессионных уравнений в анализе интенсивности рискованного поведения по его последним эпизодам.**

Аннотация. В статье описан один из возможных подходов к развитию модели, предложенной ранее для обработки сведений о последних эпизодах рискованного поведения. Построены модели, позволяющие определить взаимосвязи между параметрами, определяющими интенсивность поведения, и некоторыми демографическими и психологическими характеристиками респондента. Рассмотрен ряд критериев качества для таких моделей. Кроме того, описан один из методов обработки неопределенности, возникающей при исследовании ответов вида «сегодня» на вопрос о времени последнего эпизода.

Ключевые слова: оценка интенсивности, модели поведения, последние эпизоды, неопределенность, регрессионные модели.

Zelterman D., Paschenko A.E., Suvorova A.V., Musina V.F., Tulupyeva T.V., Tulupyev A.L., Grau L.E., Heimer R. **Regression diagnostics in the rate analysis based on data about the last episodes.**

Abstract. We describe a technique that improves the beta-prime models fitted for the last episodes of risky behavior. Regression models show interconnections between rate parameters and respondents' demographic and psychological trades. We examine these models using such criteria as jackknife and test of overdispersion. Also we develop a method for uncertainty processing in case of a special type of respondents' answers ("today" answers) about the time of their last behavior episode.

Keywords: rate estimate, behavior models, last episodes, uncertainty, regression model.

1. Введение. Во многих отраслях социологических, психологических, маркетинговых исследований возникают задачи оценивания интенсивности социально-значимого поведения респондентов [1]. Например, в настоящее время наиболее острой эпидемиологической проблемой является оценка риска передачи и приобретения такой опасной и неизлечимой инфекции как вирус иммунодефицита человека (ВИЧ) в зависимости от особенностей инъекционного и сексуального поведения индивида.

Требуется предложить математические модели, позволяющие выполнить более дешевые косвенные измерения инсиденс-показателя на основе ответов респондентов, составляющих выборку из группы риска.

Данная статья описывает один из возможных подходов к развитию и применению моделей, предложенных в работе [2]. Подробнее описание разработанной ранее модели приведено в разделе 2. Целью работы является определение взаимозависимостей между характеристиками интенсивности рискованного поведения определенного типа, в частности употребления алкоголя, и демографическими показателями (пол, возраст), а также психологическими особенностями личности, такими как психологическая защита, используемые копинг-стратегии, склонность к риску.

2. Описание модели. Одними из наиболее доступных данных, связанных с поведением респондента, являются данные о последнем эпизоде поведения. Как отмечалось в [2], использование «прямых» вопросов о числе эпизодов поведения респондента в заданный длительный промежуток времени (т.е. вопросов вида «Сколько раз Вы *делали так* в течение последнего месяца (трех, шести, года)?»), применение Лайкерт-шкал (опросников, в которых используются качественные, а не количественные варианты: «Никогда», «Редко», «Иногда», «Часто», «Всегда») или категоризованных ответов является классическим приемом, но указанный опросный инструментарий разрабатывался без учета таких потребностей как, например, получение количественных оценок интенсивности рискованного поведения и риска, с ним связанного, в эпидемиологических исследованиях ВИЧ/СПИД.

Заметим, что ответы респондента на вопросы о последних эпизодах характеризуются стабильностью воспроизведения. Такая постановка вопроса (о ближайшем событии) позволяет уменьшить ошибку, возникающую при обращении респондента к отдаленным по времени событиям.

Интенсивность поведения предлагается оценивать по данным о последних эпизодах рассматриваемого поведения или, другими словами, по известным длинам интервалов между последовательными эпизодами поведения. Так, в случае опроса о последнем эпизоде известны значения длин интервалов между моментом интервью и последним эпизодом. Отметим, что момент интервью не является эпизодом поведения, его можно рассматривать как случайное событие в жизни респондента, причем чем длиннее интервал между эпизодами, тем более вероятно, что момент интервью попадет в этот (более длинный) интервал.

В качестве модели поведения респондента рассматривается обобщенный пуассоновский процесс с параметром λ, где λ — интенсивность поведения, которая также является случайной величиной, име-

ющей гамма-распределение, т.е. $\lambda \sim g(\lambda;\alpha,\sigma)$. С учетом всех перечисленных особенностей в [2] была предложена модель, согласно которой случайная величина T — длина интервала между последним эпизодом и моментом интервью — имеет плотность распределения следующего вида:

$$f(t|\alpha,\sigma) = Kt\int_0^\infty \lambda e^{-t\lambda} g(\lambda;\alpha,\sigma)d\lambda = \frac{\alpha(\alpha-1)}{\sigma}\frac{t/\sigma}{(1+t/\sigma)^{\alpha+1}},$$

где t — длина интервала между последним эпизодом и моментом интервью, λ — интенсивность рискованного поведения, K — нормирующая константа, α, σ — параметры, характеризующие эту интенсивность. Как отмечено в [2], это распределение может быть классифицировано как бета-простое (beta-prime) распределение, причем параметр $\sigma > 0$ характеризует масштаб, а $\alpha > 1$ определяет форму распределения. Например, на рис. 1 изображены графики распределения при $\sigma = 1$ и $\alpha = 1.2, 1.5$ и 2 соответственно (снизу вверх); рис. 2 иллюстрирует изменение параметра $\sigma = 4$, $\sigma = 1$, $\sigma = 0.5$ при постоянном $\alpha = 1.5$.

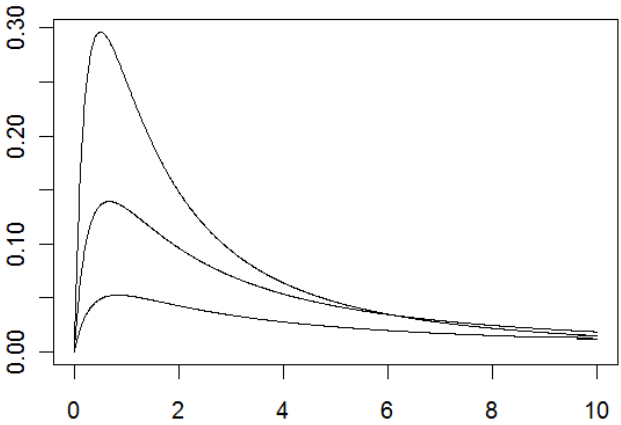

Рис.1. Изменение формы распределение при изменении параметра α.

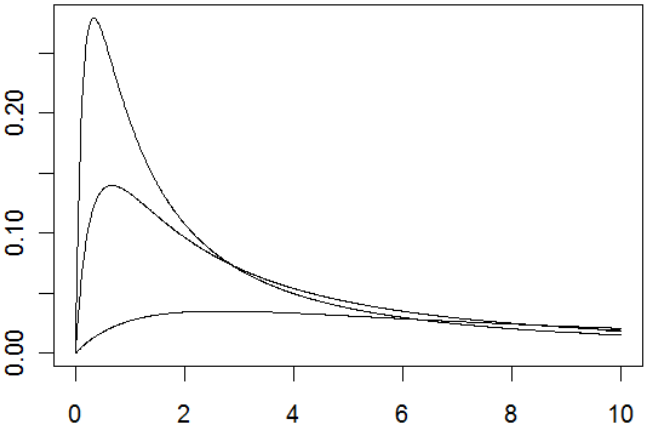

Рис. 2. Изменение масштаба при изменении параметра σ.

3. Регрессионные модели. Рассмотрим регрессионные модели для параметров α и σ, характеризующих интенсивность рискованного поведения, другими словами — модели, определяющие зависимости между предсказываемыми значениями этих параметров и некоторыми характеристиками респондента. В качестве этих характеристик (предикторов) рассмотрены как демографические (пол и возраст), так и психологические (психологическая защита Келлермана-Плутчика, копинг-тест Р.Лазаруса, изучение потребности в поиске новых ощущений, склонность к риску). Данные о рискованном поведении представляют собой сведения об употреблении алкоголя. Сбор данных осуществлялся Санкт-Петербургским институтом информатики и автоматизации РАН на базе СПб ГУЗ «Центр по профилактике и борьбе со СПИДом и инфекционными заболеваниями» (СПИД-Центр) с помощью опросника [3-5].

Пусть t_i — длина временного интервала между последним эпизодом и моментом интервью для i-ого респондента, $\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)$ — вектор характеристик респондента (предикторов). Возможны несколько вариантов построения регрессионных моделей [6]. Рассмотрим первый из них. Предполагая, что значение t_i

определяется характеристиками $\mathbf{x_i}$, получим, что длина t_i имеет плотность распределения $f(t_i|\alpha_i,\sigma)$, где α_i определяется следующим образом:

$$\log(\alpha_i - 1) = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}_i . \quad (1)$$

Коэффициенты регрессии $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ оцениваются методом максимального правдоподобия.

Преобразуя (1), получим $\alpha_i = 1 + \exp(\boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}_i)$, таким образом, $\alpha_i > 1$ для любых значений $\boldsymbol{\beta}$ и $\mathbf{x}_i$, что соответствует ограничению на значение параметра, полученному в A*. Оценка коэффициентов методом максимального правдоподобия будет проводиться с помощью процедуры nlm среды R.

Основным и основополагающим отличием R от остальных программ для статистической обработки данных является то, что R является свободной программной средой вычислений с открытым исходным кодом при сохранении большинства возможностей, необходимых для успешной работы. R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей применения.

Другая регрессионная модель строится в предположении, что длина t_i имеет плотность распределения $f(t_i|\alpha,\sigma_i)$, где σ_i определяется следующим образом:

$$\log\sigma_i = \boldsymbol{\gamma}^{\mathbf{T}}\mathbf{x}_i , \quad (2)$$

другими словами $\sigma_i = \exp(\boldsymbol{\gamma}^{\mathbf{T}}\mathbf{x}_i)$, где $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$.

Применим построенные модели к данным описанного в начале этого раздела исследования СПИИРАН о потреблении алкоголя. Использование метода nlm среды R показывает, что только по таким параметрам как копинг-стратегии и возраст и только для модели (1) обнаружена статистическая значимость зависимости рассматриваемых параметров и длины интервала между последним эпизодом употребления алкоголя и интервью.

Для характеристики качества полученных регрессионных моделей используются два критерия: jackknife и тест на overdispersion.

Первый критерий — jackknife — позволяет обнаружить выбросы, являющиеся постоянным источником проблем при анализе данных. Несколько сомнительных точек могут сильно исказить распределение, скрыть значимость (или, наоборот, незначимость) результатов. Такие неудобные данные нельзя просто выбросить, если они не поддерживают гипотезу, но метод jackknife позволяет выявить небольшое подмножество данных, влияющих на статистические показатели. Один из способов построения такого критерия описан в работе [7]. Пусть $y_1, y_2, \ldots, y_n$ — независимые наблюдения из некоторой совокупности с плотностью распределения $f(y|\theta)$, зависящей от вещественного параметра θ. Если при удалении i-ого наблюдения y_i исследуемая модель, построенная по этим наблюдениям, существенно изменяется (т.е. y_i является выбросом), то изменение оценки, полученной методом максимального правдоподобия, выражается следующим образом:

$$D_i(\theta) = \frac{\partial/\partial\theta \log f(y_i|\theta)}{-\sum_j (\partial/\partial\theta)^2 \log f(y_j|\theta)}, \qquad (3)$$

где в качестве значения θ берется оценка максимального правдоподобия этого параметра, вычисленная по полному набору наблюдений. Отметим, что знаменателем в полученном выражении (3) является наблюдаемое количество информации для параметра θ, и это значение одинаково для любого $i = 1,...,n$. Таким образом, при сравнительном анализе влияния того или иного наблюдения на модель можно ограничиться рассмотрением только числителя выражения (3).

Применяя этот критерий к предложенным регрессионным моделям, получим числители следующего вида:

$$D_i(\alpha) = \frac{1}{\alpha} + \frac{1}{\alpha - 1} - \log\left(1 + \frac{t_i}{\sigma}\right);$$

$$D_i(\sigma) = \frac{1}{\sigma}\left(\frac{(\alpha+1)t_i}{\sigma + t_i} - 2\right);$$

$$D_i(\beta_j) = x_{ij}(\alpha_i - 1)D_i(\alpha_i) = x_{ij}\left\{2 - 1/\alpha_i - (\alpha_i - 1)\log\left(1 + t_i/\sigma\right)\right\}.$$

Также для изучения полученных моделей используется тест на overdispersion, предложенный Д.Зельтерманом и Ченом [8]. Данный критерий позволяет определить, является ли параметр распределения случайной величиной или же имеет некоторое постоянное значение. В качестве гипотезы берется утверждение, что θ — некоторая константа, альтернативой является то, что θ — случайная величина. Рассматривается тестовая статистика следующего вида [8]:

$$U(\mathbf{x}\,|\,\theta) = \sum_i U_i(x_i\,|\,\theta) = \sum_i (\partial/\partial\theta)^2 \log f(x_i\,|\,\theta) + \left\{(\partial/\partial\theta)\log f(x_i\,|\,\theta)\right\}^2,$$

где в качестве значения θ берется $\hat{\theta}$ — оценка максимального правдоподобия этого параметра. При больших значениях U гипотеза отклоняется.

Тестовая статистика U выражает разницу между двумя оценками наблюдаемого количества информации для параметра θ. В случае, если гипотеза верна, U имеет нормальное распределение с нулевым математическим ожиданием. Мы будем использовать i-ое слагаемое из выражения для тестовой статистики $U_i = U_i(x_i\,|\,\theta)$ в качестве критерия того влияния, которое оказывает i-ое наблюдение.

Для рассматриваемых регрессионных моделей:

$$U_i(\alpha_i) = x_{ij}^2\left[4 - 3/\alpha_i - (5 - 2/\alpha_i)\log(1 + t_i/\sigma) + (\alpha_i - 1)^2\left\{\log(1 + t_i/\sigma)\right\}^2\right];$$

$$U_i(\sigma) = 6\sigma^{-2} - \frac{6(\alpha+1)}{\sigma^2}\left(\frac{t_i}{\sigma + t_i}\right) + \frac{\alpha(\alpha+1)}{\sigma^2}\left(\frac{t_i}{\sigma + t_i}\right)^2.$$

4. Обработка неопределенности. Один из самых сложно анализируемых ответов о времени последнего эпизода рискованного поведения — ответ вида «сегодня». И вариант, когда такому ответу соответствует значение 0 дней, и вариант 1 день оказывают сильное влияние на регрессионную модель. В качестве альтернативы описанным подходам рассмотрим следующий. Пусть фактическое значение — случайная величина, имеющая равномерное распределение на отрезке $\left[\frac{1}{24}; \frac{1}{2}\right]$, т.е. от одного до двенадцати часов. С шагом, например, один час, т.е. $\frac{1}{24}$ дня моделируем значения параметров, а затем вычисляем их среднее арифметическое, которое и считается итоговым.

Предложенный подход можно развить, применяя другие вероятностные распределения в зависимости от предположений о характере

ответа «сегодня», например, треугольное, трапециевидное, биномиальное и т.д.

5. Заключение. Необходимость оценки интенсивности поведения и нахождения ее связей с другими параметрами исследуемых объектов возникает в современных науках о человеке и обществе при решении многих задач. Одним из способов такого оценивания является рассмотрение данных об интервалах между последним эпизодом поведения и моментом интервью. Применение методики регрессионных уравнений позволяет делать выводы о связи таких данных с характеристиками респондента. Так, в результате проведенного исследования были обнаружены взаимозависимости между параметрами, определяющими интенсивность, и такими характеристиками, как возраст и копинг-стратегии. Кроме того, рассмотрены различные подходы к обработке такого случая, как получение ответа «сегодня» на вопрос о последнем эпизоде поведения.

Результатов, полученных в рамках классических операций с распределениями вероятности, моментами случайных величин и функциями правдоподобия, оказывается недостаточно при переходе к неточным данным, доступным в результате проведения интервьюирования или опроса респондентов. То есть, для дальнейшей обработки неопределенности, связанной с гранулярностью исходных данных, требуется использовать метод сводных показателей [9-11]; кроме того, полученные результаты предполагается развить с помощью методов анализа нечетких временных рядов, опираясь, в частности, на работы [12–14].

Поддержка исследований. В публикации представлены результаты исследований, поддержанных грантом для молодых ученых и кандидатов наук от Правительства Санкт-Петербурга в 2009 №25.05/027/27 «Разработка математических моделей, вычислительных алгоритмов и комплекса программ для оценки интенсивности рискованного поведения в условиях дефицита информации». Руководитель — А.Е. Пащенко. Также исследования поддержаны грантом для молодых ученых и кандидатов наук от Правительства Санкт-Петербурга в 2010 «Разработка математических моделей, алгоритмов и распределенного комплекса программ для косвенной оценки рисков, связанных с угрозообразующим поведением». Руководитель — А.Е. Пащенко. Также работа поддержана грантом AIDS International Training and Research Program “Training and Research in HIV Prevention in Russia” (2 D43 TW001028) от Fogarty International Center, National Institutes of Health, USA.

Литература

1. *Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В., Красносельских Т.В.* Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. №4. 2010. С. 30–38.

2. *Зельтерман Д., Тулупьев А.Л., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьева Т.В., Красносельских Т.В., Гро Л., Хаймер Р.* Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пауссоновской модели поведения // Труды СПИИРАН. 2011. Вып. 16. С. 160–185.
3. *Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С.* Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
4. *Тулупьева Т. В., Тулупьев А. Л., Пащенко А. Е., Сироткин А. В., Столярова Е. В., Ламанова Е. Б., Бадосова Н. В., Никитин П. В.* Психологическая защита и копинг-стратегии ВИЧ-инфицированных с позиции опасности для общественного здоровья: автоматизация сбора данных и итоги исследования // Труды СПИИРАН. 2007. СПб.: Наука, 2007. Вып. 4. С. 357–387.
5. *Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* Оценка интенсивности поведения респондента в условиях информационного дефицита // Труды СПИИРАН. Вып. 7. СПб.: Наука, 2008. С. 239–254.
6. *Zelterman D.* Models for Discrete Data: Revised Edition. New York: Oxford University Press, 2006. 285 p.
7. *Pregibon D.* Logistic regression diagnostics // Annals of Statistics. 1981. No. 9. P. 705-24.
8. *Zelterman D., Chen Ch.* Homogeneity test against central-mixture alternatives // Journal of the American Statistical Association. 1988. Vol. 83. No. 401. P. 179–182.
9. *Hovanov N., Yudaeva M., Hovanov K.* Multicriteria estimation of probabilities on basis of expert non-numeric, non-exact and non-complete knowledge // European Journal of Operational Research. 2009. Vol. 195. Issue 3. P. 857–863.
10. *Хованов Н.В.* Анализ и синтез показателей при информационном дефиците. СПб.: Изд-во СПбГУ, 1996. 196 с.
11. *Хованов Н.В.* Метод рандомизированных траекторий в задачах оценки функциональной зависимости // Труды СПИИРАН. 2009. Вып. 9. С. 262–279.
12. *Ярушкина Н. Г.* Современный интеллектуальный анализ нечетких временных рядов // Интегрированные модели и мягкие вычисления в искусственном интеллекте. V-я Международная научно-практическая конференция. Сборник научных трудов. В 2-х т. Т. 1. 2009. С. 19–29.
13. *Ковалев С.М.* Гибридные коннекционистские модели извлечения темпоральных знаний // Интегрированные модели и мягкие вычисления в искусственном интеллекте. V-я Международная научно-практическая конференция. Сборник научных трудов. В 2-х т. Т. 1. 2009. С. 30–40.
14. Нечеткие гибридные системы. Теория и практика / под ред. Ярушкиной Н.Г. М.: Физматлит, 2007. 208 с.

Зельтерман Даниэл — Ph.D., Full Professor; профессор отделения биостатистики, факультет эпидемиологии и общественного здоровья, медицинский факультет, Йельский университет. Область научных интересов: разработка статистических методов обработки категориальных данных о выживаемости, прикладная биостатистика. Число научных публикаций — 150. daniel.zelterman@yale.edu; 60 College St, LEPH 204, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-5574, fax: +1 203 785-6912.

Zelterman Daniel — Ph.D., Full Professor; Professor, Division of Biostatistics, Yale School of Epidemiology and Public Health, Yale University School of Medicine, Yale University. Research area: statistical methodology developments for categorical and survival data, applied

biostatistics. Number of publications — 150. daniel.zelterman@yale.edu; 60 College St, LEPH 204, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-5574, fax: +1 203 785-6912.

Суворова Алена Владимировна — младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), аспирант математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности, применение методов математического моделирования в эпидемиологии. Число научных публикаций — 21. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — А.Л. Тулупьев.

Suvorova Alena Vladimirovna — junior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), PhD student, Faculty of Mathematics and Mechanics of St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory, application of mathematical modeling in epidemiology. The number of publications — 21. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450. Scientific advisor — A.L. Tulupiev.

Пащенко Антон Евгеньевич — младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН). Область научных интересов: математическая статистика, статистическое моделирование, применение методов биостатистики и математического моделирования в эпидемиологии. Число научных публикаций — 45. AEP@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Paschenko Anton Evgen'evich — junior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: mathematical statistics, statistical modeling, application of biostatistics and mathematical modeling in epidemiology. The number of publications — 45. AEP@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Мусина Валерия Фуатовна — программист лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), студент математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности, биостатистика, обработка данных. Число научных публикаций — 2. valery.musina@gmail.com, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — А.Л. Тулупьев.

Musina Valery Fuatovna — programmer, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), student, Faculty of Mathematics and Mechanics of St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory, biostatistics, data processing. The number of publications — 2. valery.musina@gmail.com, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450. Scientific advisor — A.L. Tulupiev.

Тулупьев Александр Львович — д.ф.-м.н., доцент; заведующий лабораторией теоретических и междисциплинарных проблем информатики СПИИРАН, доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (СПбГУ). Область научных интересов: представление и обработка данных и знаний с неопределенностью, применение методов математики и информатики в социокультурных исследованиях, применение методов биостатистики и математического моделирования в эпидемиологии, технология разработки программных комплексов с СУБД. Число научных публикаций — 220. ALT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Tulupyev Alexander Lvovich — PhD in Computer Science, Dr. of Sc.. Associate Professor; Head of Theoretical and Interdisciplinary Computer Science Laboratory, SPIIRAS, Associate Professor of Computer Science Department, SPbSU. Research area: uncertain data and knowledge representation and processing, mathematics and computer science applications in socio-cultural studies, biostatistics, simulation, and mathematical modeling applications in epidemiology, data intensive software systems development technology. Number of publications — 220. ALT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 14-th line V.O., 39, St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Тулупьева Татьяна Валентиновна — канд. психол. наук, доцент; старший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук С.-Петербургский институт информатики и автоматизации РАН (СПИИРАН), доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (СПбГУ), доцент кафедры психологии управления и педагогики Северо-Западной академии государственной службы (СЗАГС). Область научных интересов: применение методов математики и информатики в гуманитарных исследованиях, информатизация организации и проведения психологических исследований, применение методов биостатистики в эпидемиологии, психология личности, психология управления. Число научных публикаций — 70. TVT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

Tulupyeva Tatiana Valentinovna — PhD in Psychology, associate professor; senior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), associate professor, Computer Science Department, Faculty of Mathematics and Mechanics, St. Petersburg State University (SPbSU), associate professor, Management Psychology and Pedagogic Department, North-West Academy of Public Administration (NWAPA). Research interests: application of mathematics and computer science in humanities, informatizsation of psychological studies, application of biostatistics in epidemiology, psychology of personality,

management psychology. Number of publications — 70. TVT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Гро Лоретта Е. — PhD; научный сотрудник, медицинский факультет, Йельский университет. Область научных интересов: исследования среди потребителей инъекционных наркотиков, обнаружение и предупреждение передозировки наркотиками, количественные и качественные методы анализа данных, разработка и проверка методик тестирования, качественные методы интервьюирования, когнитивные и эмоциональные соотношения между рискованным поведением и профилактическими мероприятиями. Число научных публикаций — 26. lauretta.grau@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-2904, fax: +1 203 785-3260.

Lauretta E. Grau — PhD; Associate Research Scientist, Yale School of Public Health, Yale University. Research area: health promotion among injection drug users and opioid overdose recognition and prevention, quantitative and qualitative data analysis, the development and validation of quantitative instruments, qualitative interviewing skills, and the cognitive and emotional correlates of risk and preventive health behaviors. Number of publications — 26. lauretta.grau@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-2904, fax: +1 203 785-3260.

Хаймер Роберт — Ph.D., Full Professor; профессор, медицинский факультет, Йельский университет. Область научных интересов: эпидемиология инфекционных заболеваний (в особенности ВИЧ, гепатит, ИППП). Число научных публикаций — 114. robert.heimer@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-6732, fax: +1 203 785-3260.

Heimer Robert — Ph.D., Full Professor; Professor, Yale School of Public Health, Yale University. Research area: epidemiology of infectious diseases (with the focus on HIV, hepatitis, STD). Number of publications — 114. robert.heimer@yale.edu; 60 College St, LEPH 504, EPH, Yale University, New Haven, CT, 06510-3210, USA; ph.: +1 203 785-6732, fax: +1 203 785-3260.

Рекомендовано ТиМПИ СПИИРАН, зав. лаб. А.Л. Тулупьев, д.ф.-м.н., доцент.
Статья поступила в редакцию 25.06.2011.

РЕФЕРАТ

Зельтерман Д., Пащенко А.Е., Суворова А.В., Мусина В.Ф., Тулупьева Т.В., Тулупьев А.Л., Гро Л.Е., Хаймер Р. **Диагностика регрессионных уравнений в анализе интенсивности рискованного поведения по его последним эпизодам**

Задачи оценивания интенсивности и производных характеристик поведения респондентов по их самоотчетам об эпизодах поведения возникают во многих отраслях социологических, психологических, маркетинговых исследований.

Заметим, что ответы респондента на вопросы о последних эпизодах характеризуются стабильностью воспроизведения. Однако ограниченное число и неточность, недоопределенность, нечеткость естественно-языковых формулировок ответов не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения, поэтому возникает необходимость в предложении новых математических моделей.

В статье описан один из возможных подходов к развитию модели, предложенной ранее для обработки сведений о последних эпизодах рискованного поведения. Построены регрессионные модели, позволяющие определить взаимосвязи между параметрами, определяющими интенсивность поведения, и некоторыми демографическими и психологическими характеристиками респондента, такими как пол, возраст, психологическая защита, копинг-стратегии, склонность к риску и потребность в поиске новых ощущений. Рассмотрен ряд критериев качества для таких моделей.

Все вычисления произведены на данных, полученных в результате исследования Санкт-Петербургского института информатики и автоматизации РАН на базе СПб ГУЗ «Центр по профилактике и борьбе со СПИ-Дом и инфекционными заболеваниями» (СПИД-Центр). Данные о рискованном поведении представляют собой сведения об употреблении алкоголя.

Кроме того, описан один из методов обработки неопределенности, возникающей при исследовании ответов вида «сегодня» на вопрос о времени последнего эпизода, предложены подходы по его дальнейшему усовершенствованию.

SUMMARY

Zelterman D., Paschenko A.E., Suvorova A.V., Musina V.F., Tulupyeva T.V., Tulupyev A.L., Grau L.E., Heimer R. **Regression diagnostics in the rate analysis based on data about the last episodes.**

In many fields of sociological, psychological and marketing research, we face the problem of socially significant behavior rate or frequency estimated on the base of respondents' self-reports about their behavior. The traditional approaches of asking respondents about their behavior frequency fall into two categories. The first category relies upon the question about the number of episodes of the behavior that have happened during month, 3 months, 6 months or another period of time; and it is highly implausible that respondents are able to recall all the episodes. The second category allows for collecting answers in Likert scale (e.g. "always", "very often", "often", "sometimes", "rarely", "never"); this type of answers does not provide information enough for making quantitative estimates of the behavior rate or frequency.

Our earlier studies have shown that respondents can stably provide their answers about last episodes of their behavior. In this paper, we describe a technique that improves the beta-prime model that allows for making quantitative estimates of behavior rate or frequency based on the respondents' responses about the last episodes of their behavior.

We concentrate on the data on alcohol abuse extracted from the studies conducted by the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences in collaboration with the St. Petersburg Municipal AIDS Center. The covariates we examined included the sex of the respondent, age in years, coping with stressful experiences, welcoming new experiences, defense mechanisms and the tendency to engage in risky behaviors. We construct regression models and examine them using such criteria as jackknife and test of overdispersion.

Finally, we develop a method for uncertainty processing in case of a special type of respondents' answers ("today" answers) about the time of their last behavior episode and discuss ways to improve this technique.