330.43

08.00.13

ЭКОНОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ КОЛИЧЕСТВА ЧЕЛОВЕК С РЕДКИМ ДИАГНОЗОМ НА ОСНОВЕ ПУАССОНОВСКОЙ РЕГРЕССИИ

Егорова Е. В. Пермский государственный национальный исследовательский университет (614990, г. Пермь, ул. Букирева 15), egorovaev97@gmail.com

Радионова М.В. Пермский государственный национальный исследовательский университет (614990, г. Пермь, ул. Букирева 15), m.radionova@rambler.ru

Аннотация. Достаточно часто в эконометрических исследованиях приходится работать с зависимыми переменными, которые являются результатами подсчетов, а именно представляют собой количество событий, появившихся за некоторый промежуток времени, причем, эти события появляются с постоянной интенсивностью и независимы. Важно отметить, что поскольку одним из базовых этапов эконометрического исследования является правильная спецификация модели, к выбору вида модели нужно подходить с большим вниманием и осторожностью, учитывая все особенности моделируемой величины и факторов, на нее влияющих. Для моделирования процессов, связанных со «счетными переменными», используются модели счетных данных, такие как модель преодоления препятствий, Пуассоновская регрессия и ее модификация– модель с отрицательным биномиальным распределением. Такие процессы активно используются в исследованиях у маркетологов, экономистов, социологов, работников здравоохранения и не только.

В данной статье исследуется вопрос прогнозирования количества человек на 1000 населения с редким диагнозом «Врожденные аномалии (пороки развития), деформации и хромосомные нарушения» по регионам Российской Федерации за 2016 год на основе пуассоновской регрессии. По данным 85 регионов РФ была построена эконометрическая модель, на основании которой было показано, что на количество больных человек с врожденными аномалиями и пороками среди 1000 людей, проживающих в регионе, влияют такие факторы, как процент здоровых во время беременности женщин, общая заболеваемость населения региона, уровень жизни населения и показатель загрязнения воздуха, взятый за 2005 год, и показатель эффективности здравоохранения.

Ключевые слова: эконометрическое моделирование, пуассоновская регрессия

ECONOMETRIC MODELLING OF A AMOUNT OF PEOPLE WITH THE RARE DIAGNOSIS BASED ON POISSON REGRESSION

Egorova K.V. Perm State National Research University

Radionova M.V. Perm State National Research University

Annotation. Quite often in econometric researches authors should use dependent variables, where every dependent variable is the result of the calculations, more specifically these are some number of events, that had occurred during the period of time which we have chosen for consideration, with condition, that these events occur independently, with a fixed average intensity. It is important that the one of the basic stages of econometric research is the correct specification of the model, the choice of the model type requires a lot of attention and caution, and to take into account all the properties of the simulated value and the factors influence to it. In the simulation processes associated with the "counting variables", scientists use the models of data counting, such as the model of overcoming obstacles, Poisson's regression and its modification - a model with a negative binomial distribution. These processes are used in research by marketers, economists, sociologists, medical staff and not only.

In this article carry out the research of forecasting the quantity of people per 1000 of population with a rare diagnosis, it is “Congenital anomaly (hereditary deformity), deflection and chromosomal abnormalities”. This is regional information and it had taken for 2016 in the Russian Federation. The research had based on Poisson regression. According to the data of 85 regions of the Russian Federation, an econometric model was built, on the basis of which was shown that the quantity of patients with congenital anomalies and defects among 1000 people, which living in the region, influence factors such as the percentage of healthy women during pregnancy, the overall morbidity rate of the population in the region, the standard of living of the population and the index of air pollution and the indicator of health efficiency.

Key words: econometric modeling, Poisson regression.

Введение и обзор литературы

Достаточно часто в эконометрических исследованиях приходится работать с зависимыми переменными, которые являются результатами подсчетов. Важно отметить, что поскольку одним из базовых этапов эконометрического исследования является правильная спецификация модели, к выбору вида модели нужно подходить с большим вниманием и осторожностью, учитывая все особенности моделируемой величины и факторов, на нее влияющих. Для моделирования процессов, связанных со «счетными переменными», используются модели счетных данных, такие как модель преодоления препятствий [9], пуассоновская регрессия [10] и ее модификация – модель с отрицательным биномиальным распределением [12]. Такие процессы активно используются в исследованиях у маркетологов, экономистов, социологов, работников здравоохранения и не только.

В статье [8] анализируется зависимость количества смертей населения в г. Красноярске от метеорологических параметров и концентрации загрязнителей в атмосферном воздухе (температура, влажность, бензол, фенол и т.д.) с использованием пуассоновской регрессионной модели.

В статье [13] пуассоновская регрессия используется для изучения детерминант интенсивности принятия методов управления фермами по производству какао в Гане. В качестве зависимой переменной было выбрано количество методов управления фермами какао, которое принято фермерами, а объясняющими переменными выступают «возраст», «пол», «семейное положение», «уровень образования», «размер домохозяйства», «опыт фермеров», «расширение посещений» и «членство в FBO».

Еще одним из примеров применения пуассоновской регрессии является статья [5], в которой прогнозируется результативность футбольных матчей. В качестве зависимой переменной выступает количество голов, забитых участниками футбольных матчей.

В настоящее время существует тесная взаимосвязь между уровнем здоровья населения региона и состоянием экономики. Изучением такой проблемы занимается экономика здравоохранения. В свою очередь экономика здравоохранения является экономической основой деятельности системы здравоохранения. Врожденные аномалии (пороки развития), деформации и хромосомные нарушения являются достаточно редким диагнозом среди всех основных групп болезней. Эта проблема очень значима для здравоохранения, так как напрямую связана с генофондом нации и здоровьем всех людей на планете в целом, что играет большую роль для нашего будущего. Пороки развития и хромосомные нарушения в большинстве случаев возникают в результате генетических и средовых влияний. К факторам риска, способствующим их возникновению, относят биологические, химические и физические факторы.

Методология, данные и основные гипотезы

Для проведения исследования были выбраны кросс-секционные данные по 85 регионам РФ за 2016 год, взятые с сайта Федеральной служба государственной статистики [7].

В качестве зависимой переменной выступает количество больных человек на 1000 населения с редким диагнозом «Врожденные аномалии (пороки развития), деформации и хромосомные нарушения», проживающих в регионе.

Объясняющие переменные представлены в таблице 1.

Таблица 1

Факторные признаки

Показатель	Единицы измерения	Обозначения	Источник
Количество абортов в регионе, приходившееся на 100 родов	единицы	*abortion*	сайт Федеральной служба государственной статистики [7]
Заболеваемость населения на 1000 человек по регионам	Чел.	*ilnesses*	сайт Федеральной служба государственной статистики [7]
Выбросы загрязняющих веществ в атмосферный воздух» за 2005 год в каждом регионе	тысячи тонн	*old*	сайт Федеральной служба государственной статистики [7]
Рейтинг регионов по качеству жизни	Балл (от 1 до 100)	*lifelevel*	сайт агентства «Риарейтинг» [6]
Выбросы в атмосферу загрязняющих веществ от стационарных и передвижных источников на единицу площади населенных пунктов	тысячи тонн	*atmosphere*	сайт агентства «Риарейтинг» [6]
Значения эффективности региональных систем здравоохранения в регионах	единицы	*efficiency*	журнал "ОРГЗДРАВ: новости, мнения, обучение. Вестник ВШОУЗ" [3]
Процент беременностей, протекающих без отклонений и осложнений, в каждом из регионов	%	*womhealthy*	сборник, подготовленный ФГБУ "Центральный научно-исследовательский институт организации и информатизации здравоохранения" Минздрава Российской Федерации [4]
Экологический индекс по регионам	единицы	*ecologyindex*	интернет- газета «Гласнарода» [2]

В исследовании были выделены следующие гипотезы:

H1: Взаимосвязь между эффективностью региональных систем здравоохранения и количеством больных людей в регионе носит отрицательный характер.

Оценка эффективности региональных систем здравоохранения была составлена с учетом ожидаемой продолжительности жизни, подушевыми государственными доходами на здравоохранение, валовым региональным продуктом и показателем потребления крепких алкогольных напитков на душу населения. Все эти аспекты тесно связаны с состоянием здоровья людей, что дает основание предполагать о наличии обратной связи между эффективностью региональных систем здравоохранения и количеством больных людей в регионе.

H2: Между заболеваемостью беременных женщин во время вынашивания плода и количеством людей с врожденными аномалиями существует отрицательная связь.

Действительно, для уменьшения риска врожденных аномалий и пороков развития у родившегося человека, нужно, чтобы его мать при вынашивании плода была здорова, соблюдала все рекомендации врача, как можно раньше встала на учет по беременности, больше времени проводила на свежем воздухе и не подвергала себя стрессам. В связи с этим имеет место предположение об отрицательной связи между заболеваемостью беременных женщин во время вынашивания плода и количеством людей с врожденными аномалиями. Если подходить к данному вопросу более глобально, то важно отметить, что женщина должна понимать, что она, продолжает не только свой род, но и несет ответственность за генофонд всех людей на планете.

H3: Чем выше уровень жизни в регионе, тем меньшая вероятность того, что родится человек с рассматриваемым диагнозом.

Показатель «Уровень жизни» по регионам был рассчитан на основе многих признаков, таких как уровень доходов населения, жилищные условия населения, безопасность проживания, обеспеченность объектами социальной инфраструктуры, уровень экономического развития и т.д. Из различных медицинских исследований известно, что улучшение условий жизни влечет повышение уровня здоровья населения. На основании этого можно предполагать, что между уровнем жизни и количеством больных людей существует обратная связь.

H4: Показатель «выбросы загрязняющих веществ в атмосферный воздух» за 2005 год имеет влияние на моделируемое количество человек с рассматриваемым диагнозом в 2016 году.

Если человек всю жизнь живет в среде с определенным уровнем загазованности атмосферного воздуха, то его организм адаптируется к этим внешним воздействиям и вырабатывает иммунитет. Здесь имеет место понятие приспособленности организмов к среде обитания. Существует вероятность, что, последующие поколения будут иметь более крепкий иммунитет и в меньшей степени реагировать на загрязненность воздуха, чем предки.

H5: Округ, к которому относится регион имеет влияние на количество людей с врожденными аномалиями

Каждый федеральный округ имеет климатические особенности в связи со своим экономико-географическим положением, большим или малым количеством промышленных объектов, уровнем развития инфраструктуры. Все эти критерии отражаются на здоровье населения, что сказывается на уровне заболеваемости, следовательно, округ может влиять на количество человек с редким заболеванием.

Поскольку зависимая переменная (количество человек с врожденными аномалиями и пороками) является дискретной переменной, то решено было использовать модель счетных данных, а именно Пуассоновскую регрессию. С учетом этого вероятность возникновения числа событий Y_t определяется [1, 10]:

$$P(Y_i = y_i) = \frac{e^{-\lambda} \cdot \lambda^{y_i}}{y_i!}, \quad y_i = 0, 1, 2..$$

где $\lambda = e^{x'\beta}$ - параметр распределения Пуассона, зависящий от неизвестных параметров $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ и независимых переменных $x = (x_0, ..., x_k)^T$.

Если математическое ожидание ошибки модели будет равно нулю, то условное математическое ожидание числа событий определяется как: $E[y|x_t] = \lambda = e^{x'\beta}$, а условное математическое ожидание и условная дисперсия равны [10]: $E[y|x_t] = \lambda = e^{x'\beta} = D[y|x_t]$.

Поскольку Пуассоновская регрессия нелинейна по параметрам, то для нахождения параметров модели применяют метод максимального правдоподобия [14], а найденные параметры не имеют стандартной интерпретации. В данной модели рассматриваются предельные эффекты переменных, которые вычисляются как: $\frac{\partial E[y|x_t]}{\partial x_i} = \frac{\partial e^{x'\cdot\beta}}{\partial x_i} = e^{x'\beta} \cdot \beta_i$, где скаляр x_i обозначает *i*-й регрессор. Предельный эффект показывает насколько в среднем увеличится значение зависимой переменной, если соответствующий факторный признак изменится на единицу при условии, что эта единица достаточно мала. В свою очередь если x_j измеряется в логарифмическом масштабе, то β_i является эластичностью[11].

Результаты

На основе выбранных переменных был проведен разведочный анализ данных. В таблице 2 представлены описательные статистики исследуемых переменных.

Таблица 2

Описательные статистики переменных

Переменная	Среднее	Минимум	Максимум	Ст. откл.	Вариация
Y	2,306	1,000	9,000	1,448	0,628
old	296,13	1,000	4179,0	64,531	0,218
ecologyindex	0,887	0,538	1,9412	0,215	0,242
lifelevel	43,957	12,530	76,540	11,210	0,255
efficiency	52,252	17,620	88,690	16,245	0,311
atmosphere	42,776	1,000	85,000	24,824	0,580

womhealthy	1,764	0,010	7,170	1,4513	0,823
abortion	49,612	14,000	100,00	15,656	0,316
ilnesses	802,40	447,30	1380,7	174,44	0,217

Из таблицы 2 видно, что уровень заболеваемости населения с редким диагнозом «Врожденные аномалии (пороки развития), деформации и хромосомные нарушения» принимает значения от 1 до 9 человек на каждую 1000 человек населения, среднее значение составляет 2,306. Данные неоднородны, так как вариация составляет 63%. Распределение имеет сильную правостороннюю асимметрию.

В таблице 3 представлены коэффициенты корреляции между всеми переменными.

Таблица 3

Корреляционная матрица

Показатели	*efficiency*	*atmosphere*	*womhealthy*	*ilnesses*	*lifelevel*	*old*	*Y*	*abortion*	*number*
ecology index	0,218	-0,087	-0,116	0,050	0,078	-0,082	-0,023	-0,155	-0,164
efficiency	1	-0,049	-0,203	-0,307	0,126	-0,125	-0,231	-0,518	-0,255
atmosphere		1	-0,133	0,287	0,233	0,381	0,151	-0,068	0,122
womhealthy			1	-0,080	-0,220	0,091	-0,116	0,156	0,128
ilnesses				1	0,020	0,157	0,450	0,340	0,140
lifelevel					1	0,201	-0,145	-0,110	-0,345
old						1	0,042	0,142	0,228
Y							1	0,195	0,124
abortion								1	0,411

На основе анализа таблицы 4 можно сказать, что на уровень заболеваемости наибольшее влияние оказывает заболеваемость населения на 1000 человек. Мультиколлинеарности нет.

В ходе работы были рассмотрены разные спецификации модели: линейные и логарифмические модели, в которых в качестве объясняющих переменных были рассмотрены натуральные логарифмы переменных. На основе метода пошаговой регрессии логарифмическая спецификация является лучшей с точки зрения информационных критериев Г. Шварца и Х. Акайке и скорректированного коэффициента детерминации [1].

В таблице 4 приведены результаты эконометрического моделирования, представлена наилучшая модель.

Таблица 4

Результаты моделирования

	Переменные	Оценки параметров модели	Ст. ошибки коэффициентов	
const		-5,27	2,834	*
efficiency	Эффективность региональных систем здравоохранения	-0,01	0,005	
womhealthy	Процент беременностей, протекающих без отклонений и осложнений, в каждом из регионов	-0,07	0,039	*
L_ilnesses	Логарифм заболеваемости населения	1,16	0,342	***

	на 1000 человек по регионам			
L_lifelevel	Логарифм уровня жизни	-0,34	0,23	**
old	Выбросы загрязняющих веществ в атмосферный воздух» за 2005 год в каждом регионе	-0,0001	0,00005	**

*- коэффициент значим на 10% уровне, ** - коэффициент значим на 5% уровне, *** - коэффициент значим на 1% уровне

Оценка функции регрессии зависимости количества человек с врожденными аномалиями и пороками от различных факторов имеет следующий вид:

$$\hat{Y} = e^{-5,27-0,01efficiency-0,07womhealthy+1,16\ln(ilnesses)-0,34\ln(lifelevel)-0,0001old}, \; R^2_{adj} = 0,57.$$

Статистические показатели качества моделей подтверждают ее приемлемое качество: скорректированный коэффициент детерминации равен 0,57, критерий Шварца – 284,38, критерий Акайке – 269,95 и нет мультиколлинеарности.

Для интерпретации параметров используются предельные эффекты, которые были описаны выше, и получены следующие результаты:

- если оценка эффективности региональных систем здравоохранения увеличится на единицу, то вероятность ожидаемого количества больных исследуемым диагнозом людей уменьшится на 0,49%.

- при изменении показателя уровня жизни на единицу вероятность моделируемой величины уменьшится примерно на 1,7%.

- если процент беременностей, протекающих без отклонений и осложнений увеличится на 1, то вероятность ожидаемого количества человек с врожденными аномалиями может уменьшится на 2,1%.

- в случае изменения показателя заболеваемости населения на 1000 человек на единицу вероятность ожидаемого количества больных исследуемым диагнозом людей может вырасти на 6,9%.

Связь показателя «выбросы загрязняющих веществ в атмосферный воздух» за 2005 год и количества больных людей редким заболеванием в 2016 году подтверждает, что неблагоприятные факторы окружающей среды влияют на здоровье человека, уровень и характер изменений функционального состояния организма, а также порождают возможности развития нарушений. Более того известно, что условия окружающей среды по-разному влияют на жизнеспособность и репродуктивность организмов с разными генотипами. Генофонд человечества постепенно изменяется в результате естественного отбора, и более приспособленные генотипы имеют меньшую вероятность иметь врожденные аномалии и пороки.

Заключение

На основании построенной модели можно сказать, что не были отвергнуты гипотезы

H1, H2, H3, H4. Гипотеза H5 была отвергнута, поскольку округ, в котором расположен регион не оказывает влияния на уровень заболеваемости аномалиями и пороками. Наибольшее влияние на изменение вероятности моделируемой величины оказывает фактор «Заболеваемость населения на 1000 человек». Так же хочется отметить, что человек, в процессе преобразования биосферы и улучшения качества собственной жизни, невольно способствует развитию неконтролируемых факторов, влияющих на ход генетических процессов. В их число входят мутационные эффекты, вызванные влиянием окружающей среды, и эта закономерность имеет тенденцию к увеличению. Этот факт подтверждается выявленной связью между показателями загрязнения воздуха за 2005 год и количеством больных в 2016.

Список литературы

1. Вербик М. Путеводитель по современной эконометрике. – Леувенский (Бельгия) и Тилбургский (Голландия) университеты, 2005. – 696 с.
2. Итоговый экологический рейтинг субъектов Российской Федерации за 2016 г. [электронный ресурс] // Интернет-газета «ГласНарода» URL: https://glasnarod.ru/novosti/8-sreda/62510-itogovyj-ekologicheskij-rejting-subektov-rossijskoj-federaczii-za-2016-g от 11.01.2017 [дата обращения 09.06.2018].
3. Научно-практический рецензируемый журнал "ОРГЗДРАВ: новости, мнения, обучение. Вестник ВШОУЗ" // URL: http://orgzdrav.vshouz.ru/pages/about.html [дата обращения 10.06.2018].
4. Основные показатели здоровья матери и ребенка. Деятельность службы охраны детства и родовспоможения в Российской Федерации [электронный ресурс] // URL: http://www.demoscope.ru/weekly/2017/0743/biblio05.php [дата обращения 09.06.2018].
5. Понарин Э.Д., Лисовский А.В., Зеликова Ю.А. Модели для Пуассоновских зависимых переменных: можно ли прогнозировать результативность футбольных матчей? // Социология: 4М. – 2013. – № 36. – С. 36–64.
6. Рейтинговое агентства «РИА Рейтинг», медиа группа «Россия сегодня». Качество жизни в российских регионах – рейтинг [электронный ресурс] // URL: http://riarating.ru/regions/20170220/630056099.html [дата обращения 10.06.2018].
7. Статистический сборник «Регионы России. Социально-экономические показатели» [электронный ресурс] // URL: http://www.gks.ru/bgd/regl/b17_14p/Main.htm [дата обращения 09.06.2018].

8. Тасейко О. В., Бельская Е. Н., Сугак Е. В. Прогноз смертности населения г. Красноярска в условияхповышенных температур с учетом качества атмосферного // Техносферная безопасность. Решетневские чтения. – 2015. – С. 317–319.
9. Тихомиров Н.П., Дорохина Е.Ю. Эконометрика. – М.: Изд-во Рос. экон. акад., 2002. – 640 с.
10. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия, алгоритмы Data Mining с использованием R (электронная книга). URL: https://ranalytics.github.io/data-mining/ [дата обращения 03.05.2018].
11. Cameron A.C., Trivedi P. Regression Analysis of Count Data. – Cambridge University Press, 2013. – 566 c.
12. Cameron A. C., Trivedi Pravin K. Microeconometrics, methods and applications. – Cambridge University Press, 2005. –1034 p.
13. Dennis Sedem Ehiakpor, Gideon Danso-Abbeam, Judidia Zutah, Alhassan Hamdiyah. Adoption of farm management practices by smallholder cocoa farmers in prestea Huni-valley district, Ghana // RJOAS. – 2016. – №5 (53). – С. 117–124.
14. Rainer Winkelmann Econometric Analysis of Count Data, firth edition. – Springer - Verlag Berlin Heidelberg, 2008. – 342 p.