112. Wanner, C. 2005. Money, Morality and New Forms of Exchange in Postsocialist Ukraine, *Ethnos* 70(4): 515-537.
113. Werner, C. 2002. Gifts, Bribes and Development in Post-Soviet Kazakhstan, in *Economic Development: an Anthropological Approach* edited by J. H. Cohen and N. Dannhaeuser Walnut Creek, Lanham, New York, Oxford: Altamira Press: 183-208.
114. White, R. and C. Williams (2014) Anarchist economic practices in a 'capitalist' society: Some implications for organisation and the future of work, Ephemera 14(1): 951-975
115. Williams, C.C. 2005. *A Commodified World? mapping the limits of capitalism*, Zed: London.
116. Williams, C.C., Nadin, S. and Rodgers, P. (2011) 'Beyond a "varieties of capitalism" approach in Central and Eastern Europe: Some lessons from Ukraine' Employee Relations 33(4): 413–27.
117. Williams, C., & Martinez, A. (2014). Explaining cross-national variations in tax morality in the European Union: an exploratory analysis. Studies of Transition States and Societies 6(1), 5-18.
118. Williamson, Claudia R. "Informal institutions rule: institutional arrangements and economic performance ." *Public Choice, 139*, 2009 : 371–387.
119. Yalçın-Heckmann, L. 2014. Informal Economy Writ Large and Small: From Azerbaijani Herb Traders to Moscow Shop Owners. In Morris, J. and Polese, A. (eds.). 2014. *The Informal Post-Socialist Economy: Embedded Practices and Livelihoods,* London and New York Routledge.

УДК 303.7:159.942

ЛИНЕЙНЫЙ И НЕЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ (НА ПРИМЕРЕ ИЗУЧЕНИЯ ВОСПРИЯТИЯ ЛИЧНОЙ БЕЗОПАСНОСТИ В СТРАНАХ МИРА)

Бова А. А.
кандидат социологических наук, старший научный сотрудник,
начальник научно-исследовательского отдела,
Государственный научно-исследовательский институт МВД Украины

У статті використовуються методи виявлення залежності сприйняття особистої безпеки від рівня вбивств, Індексу людського розвитку, довіри до уряду і людям на основі різних предикативних моделей: лінійної регресії, поліноміальної регресії, регресії з нелінійними компонентами, лінійні регресійні моделі на побудованому регресійному дереві. Докладний

аналіз дає змогу виявити найбільш інформативні незалежні змінні, домогтися збільшення точності моделі, виявити закономірності, характерні як для вибірки в цілому, так і окремих підгрупах, поліпшити інтерпретацію змістовних висновків. Емпіричною базою дослідження служать узагальнені результати Всесвітнього опитування Геллапа 2011 р. і матеріали Доповіді про людський розвиток 2013 ООН за 115 країнами світу.

Ключові слова: сприйняття особистої безпеки, крос-національні дослідження, регресійний аналіз, нелінійний регресійний аналіз, регресійні дерева.

В статье используются методы обнаружения зависимости восприятия личной безопасности от уровня убийств, Индекса человеческого развития, доверия к правительству и людям на основе различных предиктивных моделей: линейной регрессии, полиномиальной регрессии, регрессии с нелинейными компонентами, линейных регрессий на построенном регрессионном дереве. Подробный анализ позволяет выявить наиболее информативные независимые переменные, добиться увеличения точности модели, выявить закономерности, характерные как для выборки в целом, так и отдельных подгруппах, улучшить интерпретацию содержательных выводов. Эмпирической базой исследования служат обобщенные результаты Всемирного опроса Гэллапа 2011 г. и материалы Доклада о человеческом развитии 2013 ООН по 115 странам мира.

Ключевые слова: восприятие личной безопасности, кросс-национальные исследования, регрессионный анализ, нелинейный регрессионный анализ, регрессионные деревья.

In the article it is used methods of detection of dependence of perception of a personal safety from level of homicides, human development and trust to the government on the basis of various predictive models: linear regression analysis, polynomial regression, regression with nonlinear components, linear regression models on regression tree induction. The detailed analysis allows to reveal the most informative independent variables, to achieve increase in accuracy of model, to reveal regularities, characteristic as for selection as a whole, and separate subgroups, to improve interpretation of substantial conclusions. As empirical base of research the generalised results of Gallup World Poll 2011 and materials of the Human Development Report 2013 United Nations on 115 countries of the world has been served.

Key words: perception of personal safety, cross-national research, linear regression analysis, nonlinear regression analysis, regression trees.

Используемый в современной социологии для подтверждения теоретических положений регрессионный анализ, как правило, фиксирует линейные связи и обладает относительно небольшой точностью. Не всегда эффекты, достигнутые в одном исследовании, можно повторить на других выборках. Наряду с классической регрессией по методу наименьших квадратов, а также структурным моделированием массовому пользователю доступны разнообразные дружественные реализации машинного обучения (*Machine Learning*), позволяющие работать в условиях теоретической неопределенности, нелинейности взаимосвязей, большого числа наблюдений и переменных.

В научной литературе отмечается, что предсказательные модели по-разному используют эмпирические данные. В общем виде *глобальные модели* (нелинейный регрессионный анализ, искусственные нейронные сети) строятся для всей выборки, *частично-глобальные* (деревья решений, адаптивная регрессия) интегрируют частные закономерности в единое целое, *локальные модели* (метод ближайших соседей, локально-взвешенная регрессия) – используют часть выборки. В случае нелинейного многопараметрического решения содержательное объяснение связи затруднительно, однако можно указать на значимость переменных, представить на графике изменение значений отклика от значений регрессоров.

Улучшить информативность модели и точность прогноза может, в частности, устранение аномалий в данных, учет значимых эффектов взаимодействия, селекция нелинейных моделей, использование кусочных функций – построение регрессионного уравнения, релевантного для определенного интервала (сегмента), границы которого определяются точками разрыва, где меняется его вид. Ансамбли моделей придают еще большую точность и устойчивость результатам машинного обучения.

Целью статьи является сравнение некоторых видов регрессионных зависимостей с точки зрения точности прогнозирования и информативности для социального исследователя (на примере изучения восприятия личной безопасности). При этом использовались различные реализации алгоритмов и программы, в том числе пробные версии пакетов *SPSS*, *AMOS*, *STATISTICA.*

Автор предлагает в общем виде такие этапы регрессионного анализа, которые могут быть применимы к большому классу исследовательских задач:

1. Проверка базовых предположений регрессионного анализа, выявление аномалий в данных, их нивелирование, а при достаточно большом объёме выборки – устранение необычных наблюдений. При наличии аномальных значений наблюдений в зависимой переменной, регрессорах или всей системы переменных, включенных в анализ, рекомендуется использовать робастные методы оценки параметров.

2. Графическое отражение схемы влияний, тестирование переменных в роли модераторов и медиаторов, моделирование структурными уравнениями для уточнения или подтверждения теоретических положений, оценка полного, прямого и косвенного влияния переменных.

3. Выбор и последующая интерпретация обобщенной регрессионной модели, символьной регрессии (перебор суперпозиций функций) и нейросетевых алгоритмов, выявляющих форму связи и определяющих неизвестные параметры уравнения, что позволяет проверять бóльшее количество статистических моделей и увеличивает точность прогнозирования.

4. Построение регрессионного уравнения для одной или одновременно нескольких зависимых переменных с проецированием данных в пространство меньшей размерности (метод частичных наименьших квадратов), что целесообразно использовать при большом количестве регрессоров в условиях их мультиколлинеарности. Для решения последней проблемы также применяется метод лассо, метод эластичной сети и гребневая регрессия.

5. Построение регрессионного уравнения на интервалах независимой переменной, кластерах, латентных классах или набора регрессионных уравнений в листьях деревьев решений. Подмножества могут выбираться исходя из содержательных соображений, например, по квантилям, или формально-математических критериев.

6. Сравнение точности полученных моделей (детерминационная составляющая, ошибка прогноза), их содержательная интерпретация.

Эмпирическими данными служили результаты Всемирного опроса Гэллапа (за 2007–2011 гг.), значения коэффициента убийств (за 2004–2011 гг.) и интегрального Индекса человеческого развития (2012 г.), содержащиеся в Докладе о человеческом развитии 2013 (*N*=115 стран) [1, с. 144–147, с. 174–177]. Модели валидизированы на частично обновленных для 116 стран данных, опубликованных в рамках Программы развития ООН в 2014 г. [2, p. 160–163, p. 204–207, p. 220–223]. Показатели определялись таким образом:

Восприятие безопасности (*personal safety – PS*) – процент респондентов, ответивших «да» на вопрос: «Чувствуете ли вы себя в безопасности, когда прогуливаетесь в одиночестве ночью в городе или районе, в котором вы живете?».

Индекс человеческого развития (*Human development index – HDI*) – комбинированный индекс, измеряющий среднюю величину достижений в трех основных измерениях человеческого развития: здоровье и долголетие, знания и достойные условия жизни.

Доверие к правительству страны (*trust in national government – TNG*) – процент респондентов, ответивших «да» на вопрос: «Доверяете ли вы правительству вашей страны?».

Доверие к людям (*trust in other people – TP*) – процент респондентов, ответивших «да» на вопрос: «В целом, считаете ли вы, что большинству людей можно доверять, или вы считаете, что следует проявлять осторожность, имея дело с людьми?».

Коэффициент убийств (*homicide rate – HR*) – число умышленных убийств, т. е. смертей, незаконно причиненных человеку другим человеком, в пересчете на 100 тыс. чел.

Согласно усредненным данным доклада показатель ощущения безопасности человека, прогуливающегося в одиночестве ночью в местности проживания, выше в странах со средним и очень высоким уровнем человеческого развития – соответственно 73,4 % и 68,4 %, по сравнению со странами с высоким (47,6 %) и низким (57,4 %) уровнем человеческого развития. Показатель доверия к национальному правительству выше в странах с низким уровнем человеческого развития (50,8 %), чем в странах с очень высоким уровнем человеческого развития (38,1 %). В странах с низкими [0,304; 0,534] и высокими [0,712; 0,796] значениями Индекса человеческого развития коэффициенты убийств составляют 14,6 и 13,0 случаев на 100 тыс. населения, а в странах со средним [0,536; 0,710] и очень высоким [0,805; 0,955] уровнем человеческого развития – соответственно 3,9, и 2,1 случаев.

Доля тех, кто чувствует себя ночью в безопасности на улице, в среднем по выборке составляет 60 %, а в Украине – 48 %. Дескриптивные статистики свидетельствуют о значительной неоднородности информации. Так, коэффициент вариации для *HR* составляет 138 %, *TP* – 49 %, *TNG* – 39 %, *HDI – 27 %, PS* – 26 %. Наиболее неравномерное распределение имеет коэффициент убийств с минимальным значением 0,2 случая на 100 тыс., максимальным – 91,6, средним – 10,5, медианой – 4,2, значением первого квартиля – 1,6 и третьего – 14,7. Значительный разброс данных может быть обусловлен, с одной

стороны, как разными подходами к учету этого вида преступления, так и массовыми смертями в период войн и гражданских конфликтов, с другой – недооценкой этого вида преступлений в развивающихся странах.

Существуют разнообразные способы выявления аномальных значений в данных. На основе *Z*-критерия определены резко выделяющиеся значения переменной *HR*: Гондурас (91,6), Сальвадор (69,2), Кот-д'Ивуар (56,9), Венесуэла (45,1); переменной *TP* – Дания (60 %), Финляндия (58 %), Марокко (58 %), Швеция (55 %), Джибути (55 %); переменной *TNG* – Руанда (96 %), Камбоджа (90 %), Таджикистан (89 %), Катар (89 %), Румыния (12 %), Латвия (11 %); переменной *HDI* – Нигер (0,304) и Конго (0,304); переменной *PS* – Руанда (92 %) и Грузия (91 %).

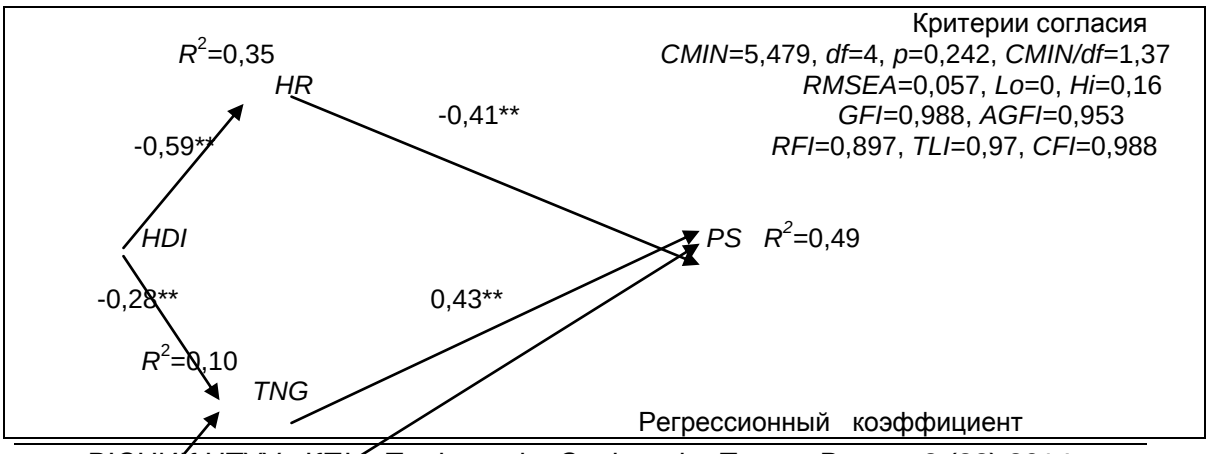
Корреляционный анализ обнаруживает положительную статистически значимую связь между долей людей, которые чувствуют себя в безопасности в ночное время с долей респондентов, которые доверяют правительству (0,4) и людям (0,43), Индексом человеческого развития (0,2) и отрицательную связь с коэффициентом убийств (-0,4).

При сравнении адекватности модели данным будем использовать квадрат коэффициента корреляции Пирсона наблюдаемых значений с предсказанными по модели, возведенный в квадрат (R^2). Для *линейной регрессионной модели* с оценкой параметров по методу наименьших квадратов коэффициент множественной детерминации составляет 0,44 (на проверочной выборке – 0,14).

$$\hat{PS} = 27,4 - 0,31HR + 0,32TNG + 0,41TP + 14,8HDI \qquad (1)$$

Исходя из значений коэффициентов в стандартизованном масштабе, наибольший вклад в модель вносит колебания уровней доверия населения к национальному правительству (0,39) и людям (0,30), коэффициент убийств (-0,29) и, наконец, Индекс человеческого развития (0,17). Интерпретация направления связи коэффициентов регрессии аналогична коэффициентам корреляции. После удаления из выборки 6 объектов с необычными остатками (Грузия, Словения, Уругвай, Гондурас, Ботсвана, Афганистан) отмечается, при сохранении характера связи, увеличение значения коэффициента детерминации (R^2=0,52).

Проверка и уточнение теоретических предположений с помощью системы *линейных уравнений с путевыми коэффициентами*, безусловно, является стандартом в количественных социологических исследованиях. Для нашей выборки стран значение многомерного эксцесса составляет 10,2 при критическом значении 6,5, что свидетельствует о большой крутизне распределения по сравнению с нормальным. Оценка модели по асимптотическому непараметрическому критерию и оценки адекватности, позволяют принять нулевую гипотезу о соответствии данных теоретической схеме



0,15*	0,31**	статистически значимо отличается от нуля на уровне статистической значимости: **α=0,01, *α=0,05
TP		

Схема влияний факторов на восприятие личной безопасности в виде стандартизированных регрессионных коэффициентов

Примечание. *CMIN* – минимальное значение расхождения между моделью и данными (критерий хи-квадрат), *df* – число степеней свободы, *p* – вероятность ошибки при отклонения нулевой гипотезы, *RMSEA* – квадратный корень среднеквадратической ошибки аппроксимации, *Lo*, *Hi* – нижняя и верхняя граница 90% доверительного интервала *RMSEA*, *GFI* – Индекс Джорескога, *AGFI* – скорректированный индекс Джорескога, *RFI* – относительный индекс адекватности, *TLI* – коэффициент Такера – Льюиса, *CFI* – сравнительный индекс адекватности.

На восприятие личной безопасности влияют как объективные, так и субъективные факторы. В системе регрессионных уравнений вырисовываются две ключевых переменные, обусловливающие влияние других независимых переменных – коэффициент убийств и уровень доверия к правительству.

Результаты локально взвешенной регрессии (с последовательным включением от 10% до 50% наблюдений с шагом в 10%) хотя и не увеличивают прогностическую точность по сравнению с линейным регрессионным анализом, однако *LOESS*-сглаживание дает наглядное представление о нелинейном характере связи каждого из регрессоров с откликом. Парная регрессия может натолкнуть на необходимость различных преобразований переменных.

Значение коэффициента детерминации *регрессионного анализа с нелинейными компонентами* R^2 = 0,53. В уравнении сохраняется и первоначальная логика взаимосвязи переменных.

$$\hat{PS} = 42{,}6 + 3{,}58\sqrt{TP} + 0{,}003\,TNG^2 - 12{,}9\,LgHR \qquad (2)$$

Исходя из значений бета-коэффициентов наибольший вклад в модель вносит десятичный логарифм коэффициента убийств (-0,47), квадрат доли, доверяющих правительству (0,41), корень квадратный из доли, доверяющих людям (0,26). Модель была валидизирована на данных, опубликованных в 2014 г. (R^2 = 0,52).

Имеются и другие многомерные статистические методы идентификации нелинейной связи, формирования околооптимального решения, например, метод группового учета аргументов (МГУА) и искусственные нейронные сети [3]. МГУА, основанный на принципах самоорганизации индуктивный метод, находит зависимость в виде полиномов низкой степени и произведений переменных. С помощью его обрабатываются как небольшие, так и значительные наборы данных с выбором наиболее информативных признаков. Различные реализации МГУА (нейросетевые, гибридные), при большей сложности по сравнению с регрессией с нелинейными компонентами, показывают значение R^2 на данных 2013 г. от 0,52 до 0,6, и на проверочных данных за 2014 г. – от 0,49 до 0,54. При обучении многослойного персептрона с четырьмя входными переменными, двумя нейронами внутреннего слоя и одним выходным нейроном (*MLP* 4-2-1) выборка была разделена на обучающую (104 объекта) и экзаменационную (11 объект). Функция активации для внутреннего слоя – гиперболический тангенс, внешнего – линейная. Рассчитанный на общем наборе данных за 2013 г. R^2 = 0,57, а за 2014 г. R^2 = 0,53.

Современным методом поиска количественных закономерностей является символьная регрессия, суть которой сводится к отбору математического выражения посредством скрещивания и мутаций наборов грамматик, принимающих участие в моделировании (констант, переменных, арифметических, тригонометрических, логических и других функций). Сложность уравнения регулируется показателями желаемой точности модели, набором грамматик, максимальной длиной и вложенностью математического или логического выражения. Нелинейная связь меньшего количества независимых переменных может иметь практическую ценность в случае существенного увеличения расходов на сбор информации по какому-либо показателю. Генетические алгоритмы также используются при оптимизации параметров нейронных сетей (количество входных переменных и нейронов внутреннего слоя, набор весовых коэффициентов), генерации деревьев решений с листьями, которым соответствуют регрессионные уравнения.

Существуют различные методы построения кусочно-линейных и кусочно-полиномиальных регрессионных моделей, генерирования терминальных вершин с уравнениями регрессии, в частности: выявление наличия специфических групп посредством распознавания образов без учителя (разнообразные иерархические и неиерархические методы) с последующим проведением регрессионного анализа в кластерах; регрессионный кластеринг, учитывающий дополнительную информацию о зависимой переменной с минимизацией числа кластеров и максимизацией коэффициентов детерминации в них; регрессионный анализ на латентных классах; индукция деревьев решений и логических правил с построением регрессионных уравнений; многомерные адаптивные регрессионные сплайны. Также возможно применение структурного моделирования на подгруппах, образованных деревьями решений.

Специализированные алгоритмы кусочно-линейной регрессии могут работать с разнотипными данными, особым образом учитывать пропущенные значения или заменять их средними по выборке, пошагово вводить переменные в анализ и автоматически сокращать количество логических закономерностей, что повышает устойчивость модели и упрощает её интерпретацию, выявлять значимые взаимодействия, использовать для построения прогноза различные регрессионные или классификационные функции, принимать решение «большинством голосов» и усреднением по ряду моделей, построенных на одной выборке.

Описательный анализ таблиц сопряженности, содержащихся в Докладе о человеческом развитии 2013 свидетельствует о наличие группы стран, не вписывающихся в общую линейную тенденцию, что предполагает обращение к возможностям типологической регрессии. Построим модель неоднородных данных путем предварительной направленной сегментации объектов. На основе *F-критерия* важность влияния независимых переменных имеет такой порядок (α = 0,001): *HR* (5,7), *TP* (4,4), *TNG* (3,8), *HDI* (3,4). Используя алгоритм *CHAID* с выделением одноуровневого дерева, коэффициент убийств был разбит на три интервала по изменению доли людей, чувствующих себя в безопасности. Первая группа включает 11 стран (*HR* < 0,9); вторая – 58 стран с распределением *HR* в интервале от 0,9 до 7,5; третья – 46 стран (*HR* > 7,5). Для обучающей (*N*=115) и проверочной (*N*=116) выборок R^2 общей кусочно-линейной модели составляет 0,53. Ниже для каждой группы стран приведены линейные регрессионные уравнения после шагового исключения менее важных переменных.

$$\hat{PS}_1 = 63{,}8 + 0{,}29TNG_1 \qquad (3)$$

($\overline{PS}_1$ =77,4%, $\overline{TP}_1$ = 31%, $\overline{TNG}_1$ = 47%, $\overline{HR}_1$ = 0,6, $\overline{HDI}_1$ = 0,89, $R_1^2 = 0{,}39$)

$$\hat{PS}_2 = 46{,}8 + 0{,}2TP_2 + 0{,}3TNG_2 - 1{,}9HR_2 \qquad (4)$$

($\overline{PS}_2$ =62,5%, $\overline{TP}_2$ = 24%, $\overline{TNG}_2$ = 49% , $\overline{HR}_2$ = 2,8, $\overline{HDI}_2$ = 0,74, $R_2^2 = 0{,}38$)

$$\hat{PS}_3 = 20{,}9 + 0{,}7TP_3 + 0{,}3TNG_3 \qquad (5)$$

($\overline{PS}_3$ =51,8%, $\overline{TP}_3$ = 21%, $\overline{TNG}_3$ = 52% $\overline{HR}_3$ = 22,6, $\overline{HDI}_3$ = 0,55, $R_3^2 = 0{,}4$)

Первая группа стран (включающая, например, такие высокоразвитые страны как Австрию, Данию, Германию, Швейцарию, Гонконг, Японию, Сингапур), характеризуется высоким уровнем человеческого и экономического развития (ежегодный валовой внутренний продукт с учетом паритета покупательной способности составляет в среднем 40 023 доллара на человека) и низким коэффициентом убийств. Доверие к национальному правительству в этих странах ниже, нежели по выборке в целом. На фоне высокого уровня межличностного доверия фактором, определяющим изменение восприятия личной безопасности, выступает эффективность правительства (бета-коэффициент 0,12).

Вторая группа стран отличается средним уровнем доверия к правительству, несколько более высоким коэффициентом убийств по сравнению с первой группой, меньшим уровнем доверия к людям и меньшим уровнем человеческого развития (ВВП составил в среднем 14 563 доллара). Как следует из уравнения регрессии, субъективное восприятие личной безопасности зависит как от показателей доверия к правительству (бета-коэффициент 0,48) и окружающим людям (0,23), так и от распространения преступности в обществе (-0,24). В процессе классификации в эту группу стран попала и Украина, точечное расчетное значения субъективного уровня безопасности для которой составило 51 %.

Третья группа стран (в неё входят, прежде всего, страны Африки и Латинской Америки – Конго, Бурунди, Боливия, Никарагуа, Перу) характеризуется низким межличностным доверием и высоким доверием к правительству на фоне более низкого человеческого и экономического развития (ВВП равняется 4 496). В этих обществах коэффициент убийств