

КАЧЕСТВЕННЫЙ АНАЛИЗ ИНФОРМАЦИОННОЙ МОДЕЛИ СВЯЗИ ЧИСЛЕННОСТИ БЕЗРАБОТНЫХ И ЧИСЛА ПРЕСТУПЛЕНИЙ

С. В. ПОЛЯКОВ,

к. т. н., доцент кафедры информационных технологий Владимирского филиала ФГБОУ ВПО «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации»

Качественный анализ информационной модели, описывающей состояние системы, позволяет снизить негативные последствия принимаемых решений, а также избавиться от установившихся стереотипов, ставших с течением времени ложными.

В качестве информационной базы для проведения качественного анализа возьмем статистические данные для Центрального федерального округа «Численность официально зарегистрированных безработных граждан (тысяча человек)» и «Число зарегистрированных преступлений в расчете на 100 тыс. чел. населения (единица)» (табл. 1).

Таблица 1

Год	Численность официально зарегистрированных безработных граждан (тыс. чел.)	Число зарегистрированных преступлений в расчете на 100 тыс. чел. населения (единица)
2001	210	137
2002	215,5	150
2003	234,4	139
2004	226,7	130
2005	248,2	14
2006	228,8	161
2007	214,2	179
2008	177,2	162
2009	236,1	157
2010	363,8	136
2011	260,6	116
2012	217,7	113

Примем, что «Численность официально зарегистрированных безработных граждан (тыс. чел.)» влияющий фактор x , а «Число зарегистрированных преступлений в расчете на 100 тыс. чел. населения (единица)» результативный признак y . Установим, существует ли статистически значимая связь между влияющим фактором x и результативным признаком y .

Прежде чем выполнять оценку связи факторов исходных данных, необходимо выделить «выбросы» точек наблюдения, т.е. выявить те наблюдения, которые вносят по тем или иным причинам искажения в общую оценку. С этой целью строится диаграмма рассеяния (рис. 1).

В диаграмме рассеяния можно выделить два возможных «выброса»: точки с координатами (248,2; 14) для 2005 года и (363,8; 136) для 2010 года. О причинах «выбросов» можно только догадываться. Так, по всей вероятности, наблюдение с координатами (248,2; 14) имеет ошибку внесения данных в базу и, возможно, должна быть (248,2; 140). Для второй точки наблюдения (363,8; 136) не исключено, что в 2010 году имеем аномальное увеличение численности официально зарегистрированных безработных граждан, связанное, по всей вероятности, с кризисными процессами в стране. Точки наблюдений предполагаемых выбросов необходимо оценить на предмет выбора одного из решений: оставить эти наблюдения в исходных данных или исключить их из исходных данных, как не отражающих общую тенденцию

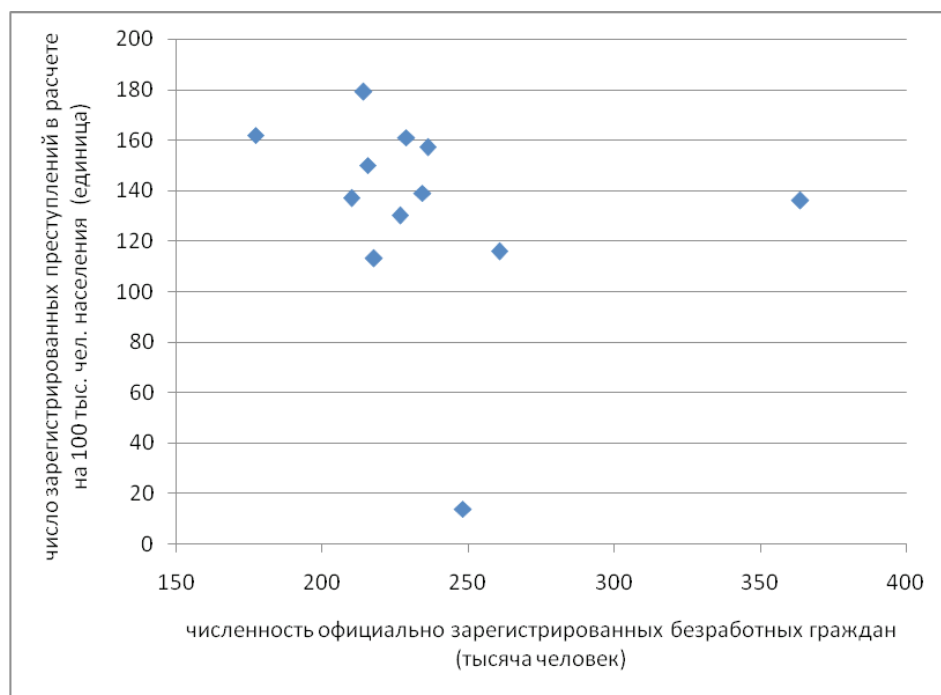


Рис.1.

несмотря на возможную ошибку, допущенную при внесении этой координаты в базу данных. Аналогичный расчет выполнен по отношению возможного «выброса» наблюдения с координатами (363,8; 136), который показал результата $(r_{xy})_2 = -0,50$. Получено существенное изменение характера зависимости факторов y и x от слабого до заметного. Следовательно, наблюдение с координатами (363,8;

136) можно считать «выбросом», искажающим общий характер зависимости между численностью официально зарегистрированных безработных граждан и числом зарегистрированных преступлений для Центрального федерального округа. Поэтому наблюдение с координатами (363,8; 136) исключаем из дальнейшего анализа, а окончательное значение линейного коэффициента корреляции зависимости между факторами y и x примем $(r_{xy})_2 = -0,50$.

Практическую значимость в качественном анализе исходных данных имеет коэффициент детерминации, который показывает, какой удельный вес в общей дисперсии ряда y занимает дисперсия, вызываемая вариацией фактора x . Коэффициент детерминации можно вычислить как квадрат линейного коэффициента корреляции, т.е. $[(r_{xy})_2]^2 = 0,25$. Это позволяет делать вывод, что для линейной модели доля дисперсии числа зарегистрированных преступлений в расчете на 100 тыс. чел. населения (y), вызванная влиянием численности официально зарегистрированных безработных граждан (x) составляет 25 %, а остальная доля дисперсии фактора y размером 75 % вызвана влиянием неучтенных в модели факторов. Иными словами, для линейной модели число зарегистрированных преступлений в расчете на 100 тыс. чел. населения в Центральном федеральном округе объясняется только на 25 % численностью официально зарегистрированных

изучаемого процесса. Для этого воспользуемся оценкой тесноты зависимости между факторами y и x линейным коэффициентом корреляции,

$$r_{xy} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y},$$

где $\overline{y \cdot x}, \bar{y}, \bar{x}$ средние значения;

σ_x, σ_y средние квадратические отклонения.

Линейный коэффициент корреляции, вычисленный по приведённой формуле для всех точек наблюдения таблицы 1, составил $(r_{xy})_1 = -0,21$. Следовательно, теснота зависимости между численностью официально зарегистрированных безработных граждан и числом зарегистрированных преступлений в расчете на 100 тыс. чел. населения для Центрального федерального округа слабая и отрицательная. Без предполагаемой точки «выброса» с координатами (248,2; 14) линейный коэффициент корреляции для оставшихся точек наблюдения составляет $(r_{xy})_2 = -0,30$, т.е. квалификация тесноты зависимости исследуемых факторов осталась неизменной теснота зависимости слабая и отрицательная. Это говорит о том, что предполагаемый «выброс» с координатами (248,2; 14) не оказал существенное влияние на характер зависимости факторов y и x . Следовательно, наблюдение с координатами (248,2; 14) можно оставить в составе исходной базы наблюдений,

безработных граждан, а 75 % дисперсии числа зарегистрированных преступлений вызвано влиянием иных не учтенных в модели факторов.

Обязательным условием оценки качества связи исследуемых факторов является проверка статистической значимости коэффициента корреляции. При количестве наблюдений n менее 30 значимость коэффициента корреляции проверяется на основе t -критерия Стьюдента. Для этого рассчитывается фактическое значение критерия по формуле

$$t_r = |r| \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}},$$

которое сопоставляется с табличным $t_{\text{табл}}$. Табличное значение t -критерия Стьюдента определяется по таблицам (например, таблица приложения 9) для числа степеней свободы $m = n - 2$ и заданного уровня значимости, который обычно принимается $\alpha = 0,05$ (или 5 %).

Если $t_r > t_{\text{табл}}$, коэффициент корреляции r_{xy} считается значимым, а связь факторов y и x – реальной. Если $t_r < t_{\text{табл}}$, то считается, что связь между y и x отсутствует, а полученное значение r_{xy} носит случайный характер.

Для рассматриваемого примера имеем $t_r = 1,76$. При числе степеней свободы $m = 9$ и уровне значимости $\alpha = 0,05$ табличное (критическое, пороговое) значение t -критерия Стьюдента составляет $t_{\text{табл}} = 2,26$.

Поскольку фактическое значение t -критерия меньше табличного ($t_r < t_{\text{табл}}$), то линейный коэффициент корреляции $(r_{xy})_2 = -0,50$ считается незначимым, т.е. коэффициент обладает случайным характером, а следовательно, связь между y (число зарегистрированных преступлений) и x (численность официально зарегистрированных безработных граждан) отсутствует. Ошибка такого вывода не превышает 5 %.

Определим, для какой ошибки α можно получить вывод о наличии статистически значимой зависимости числа зарегистрированных преступлений от численности официально зарегистрированных безработных граждан. Для этого используем инструмент «Подбор параметра»

Excel и функцию СТЬЮДРАСПОБР. Несложно получить, что для $\alpha = 11,2$ % имеем равенство $t_r = t_{\text{табл}}$. Таким образом, с ошибкой не менее 11,2 % можно утверждать, что в Центральном федеральном округе число зарегистрированных преступлений определяется в том числе численностью официально зарегистрированных безработных граждан.

ВЫВОДЫ

Анализ диаграммы рассеяния влияющего фактора x «Численность официально зарегистрированных безработных граждан (тыс. чел.)» и результативного признака y «Число зарегистрированных преступлений в расчете на 100 тыс. чел. населения (единица)» для Центрального федерального округа за период 2001-2012 годов позволил определить точку наблюдения с координатами (363,8; 136) для 2010 года как «выброс», поэтому данная координата была исключена из дальнейшего анализа.

Значение линейного коэффициента корреляции численности официально зарегистрированных безработных граждан и числа зарегистрированных преступлений составило $(r_{xy})_2 = -0,50$. Знак минус показывает, что с увеличением зарегистрированных безработных граждан число зарегистрированных преступлений снижается.

Для линейной модели число зарегистрированных преступлений в расчете на 100 тыс. чел. населения в Центральном федеральном округе объясняется только на 25 % численностью официально зарегистрированных безработных граждан, а 75 % дисперсии числа зарегистрированных преступлений вызвано влиянием неучтенных в модели факторов.

В Центральном федеральном округе число зарегистрированных преступлений определяется численностью официально зарегистрированных безработных граждан с ошибкой не менее 11,2 %. Т.е. надёжность значимости связи рассматриваемых факторов составляет не более 88,8 %.

¹ Центральная база статистических данных <http://www.gks.ru/dbscripts/Cbsd/DBInet.cgi> на 25 марта 2013 года.

² Громыко Г. Л. Теория статистики: Практикум. – 4-е изд., доп. и перераб. – М.: ИНФРА-М, 2009. С. 117.

³ Громыко Г. Л. Теория статистики: Практикум. – 4-е изд., доп. и перераб. – М.: ИНФРА-М, 2009. С. 236.