



МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Национальный исследовательский
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ВСЕ ГРАНИ МАТЕМАТИКИ И МЕХАНИКИ

**Сборник статей
Всероссийской молодежной
научной конференции студентов**

Томск, 27 мая – 1 июня 2021 г.



ТОМСК
«Издательство НТЛ»
2021

Определение статистически значимых факторов, влияющих на величину страховых взносов

П.А. Сомова, Е.А. Пчелинцев

*Национальный исследовательский
Томский государственный университет, г. Томск, Россия*

В работе рассматривается влияние на величину страхования жизни (без пенсионного) по субъектам Российской Федерации таких факторов, как количество договоров, сумма страхования, средняя стоимость договоров, численность населения, ожидаемая продолжительность жизни (ОПЖ) и экологический уровень. Установлены зависимости между признаками, проведены факторный и кластерный анализ, построена регрессионная модель.

Ключевые слова: *страхование жизни, корреляционно-регрессионный анализ, критерий независимости хи-квадрат, факторный анализ, кластерный анализ, регрессионная модель.*

Методы математической статистики широко распространены в экономике, в частности в сфере страхования.

Основой исследования послужили данные страхования жизни (без пенсионного) за 2016–2020 гг. по субъектам Российской Федерации [1–3]. В работе устанавливается влияние таких факторов, как количество договоров, сумма страхования, средняя стоимость договоров, численность населения, ожидаемая продолжительность жизни и экологический уровень, включающий в себя природоохранный, промышленно-экологический и социально-экономические индексы, на величину страхования взносов [4].

Для отобранных данных проведена первичная обработка, а именно построены гистограммы относительных частот по каждой выборке отдельно за каждый год, в совокупности за 5 лет и найдены выборочные средние, выборочные дисперсии и медианы (табл. 1, X1 – количество договоров, X2 – сумма страхования, X3 – средняя стоимость договоров, X4 – численность населения, X5 – ОПЖ, X6 – экологический уровень).

Таблица 1

Результаты первичной обработки данных

2016 год	X1	X2	X3	X4	X5	X6
Выборочное среднее	25435,21	12128050,60	1251,11	1724055,41	71,17	45,81
Выборочная дисперсия	23875679920,81	1836695542175510,00	4662213,65	3120872893094,41	5,83	26,13
Медиана	6054,00	5620194,00	842,53	1187685,00	70,94	46,00

Построены гистограммы, исходя из которых можно предположить, что:

- гистограммы количества договоров, суммы страхования и численности населения имеют распределение из семейства гамма-распределений (рис. 1):

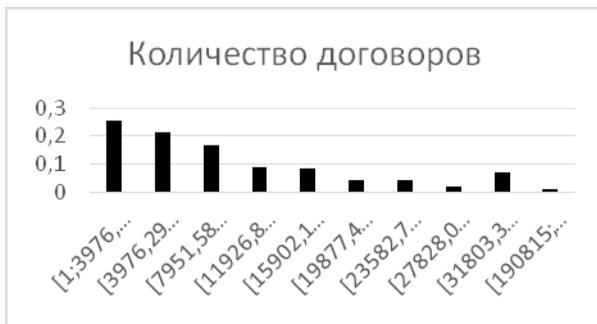


Рис. 1. Сводная гистограмма количества договоров

- гистограмма экологического уровня и ОПЖ – нормальное распределение (рис. 2):



Рис. 2. Сводная гистограмма экологического уровня

Средние выборочные значения и выборочные дисперсии с каждым годом увеличиваются, что говорит о возрастании спроса на страхование жизни.

Для определения степени связи между факторами построены корреляционные матрицы (табл. 2). Получено пять таблиц с одинаковым результатом в каждом году: наибольшая теснота линейной зависимости

наблюдается между суммой страхования и количеством договоров, численностью населения и количеством договоров, а также численностью населения и суммой страхования. Во всех остальных случаях связь либо слабая, либо вовсе отсутствует.

Таблица 2

Корреляционная матрица

2020 год	X1	X2	X3	X4	X5	X6
X1	1,000					
X2	0,998	1,000				
X3	-0,003	0,045	1,000			
X4	0,720	0,738	0,184	1,000		
X5	0,252	0,263	0,061	0,324	1,000	
X6	0,047	0,046	0,085	-0,103	0,169	1,000

Для факторов, между которыми слабая связь, проверили гипотезу о независимости с помощью критерия хи-квадрат. Уровень значимости $\alpha = 0,05$, 25 степеней свободы. Предполагаем, что:

H_0 : связи между признаками нет, они независимы;

H_1 : связь между признаками есть, они не независимы.

Затем построили по две таблицы сопряженности на каждый признак с наблюдаемыми и ожидаемыми частотами и вычислили критические и наблюдаемые значения (табл. 3).

Таблица 3

Критические и наблюдаемые значения

Хи- квадрат выборочное	35117,00
Хи-квадрат табличное	37,65

Выборочное значение статистики хи-квадрат больше, чем табличное, следовательно, принимается гипотеза H_0 об отсутствии зависимости. Все признаки со слабой связью по критерию хи-квадрат тоже являются независимыми, как и по корреляционной матрице.

С помощью факторного анализа пробуем выявить скрытые переменные факторы, отвечающие за наличие линейной статистической корреляции между наблюдаемыми переменными [5] (табл. 4).

Таблица 4

Результаты факторного анализа

Количество договоров	0,944243	0,068189
Сумма страхования	0,956162	0,080475
Ср. стоимость договора	-0,134242	0,724811
Численность населения	0,860052	-0,155623
ОПЖ	0,462086	0,071663
Экологический уровень	0,074156	0,764536
Expl.Var	2,782572	1,150346
Prp.Totl	0,463762	0,191724

Из проведенного анализа видно влияние значимых факторов на переменные. На сумму страхования и количество договоров оказывает влияние спрос, который порождается численностью населения, а экологический уровень оказывает влияние на среднюю стоимость договоров страхования.

Кластерным анализом свяжем наши субъекты и агрегируем в кластеры, состоящие из различающихся элементов. Используем евклидово расстояние. Воспользуемся алгоритмом иерархического кластерного анализа для создания группы сходных объектов по субъектам РФ по всем шести факторам (рис. 3).

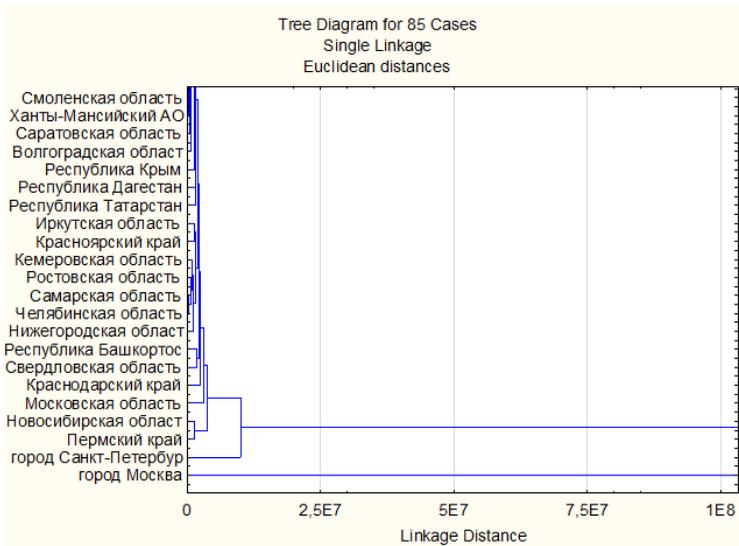


Рис. 3. Результаты кластерного анализа

По кластерам видно, что централизация отрасли происходит в Москве и Санкт-Петербурге, в остальных регионах достаточно равномерно. Также можем сделать вывод, что фактор времени никак не влияет на группировку, так как факторы за все 5 лет не меняют свой вид.

Проведём регрессионный анализ, будем рассматривать множественную регрессию между зависимой переменной Y и несколькими причинно-обусловленными предсказывающими X_1, X_2, \dots, X_k :

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k,$$

где X_1, X_2, \dots, X_k – факторы, Y – отклик, b_0, b_1, \dots, b_k – параметры (коэффициенты).

Регрессионным анализом выявлено три мультиколлинеарных фактора, проанализированы остатки, на основе этого построена частотная гистограмма остатков, гипотеза о нормальности не отклонилась. Далее с помощью нормально-вероятностного графика (рис. 4) подтверждено нормальное распределение остатков, выявлена зависимость, определена приемлемость модели, как хорошая и оценен высокий показатель коэффициента детерминации – 98 % (рис. 5).

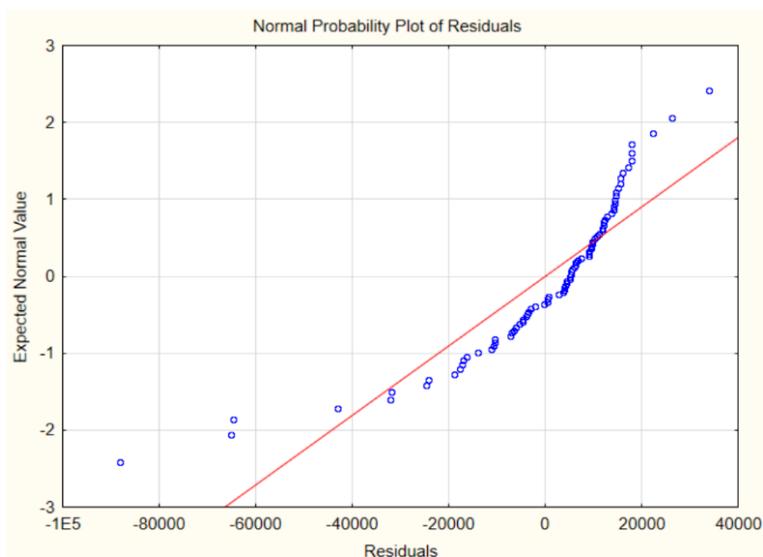


Рис. 4. Нормально-вероятностный график остатков

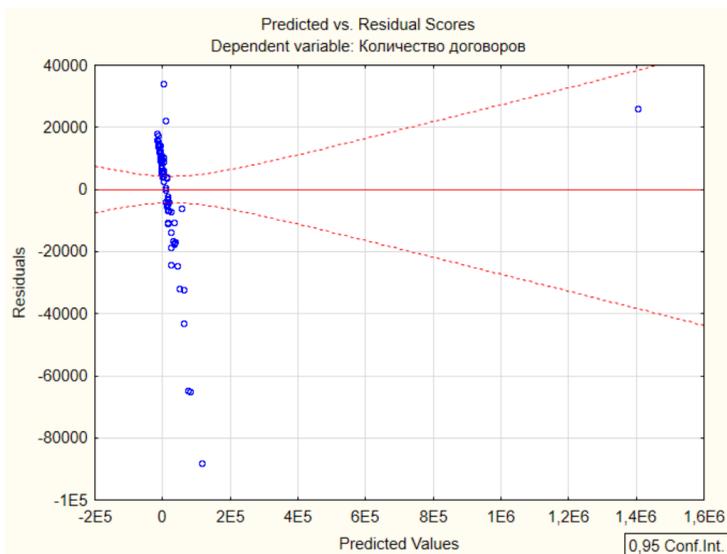


Рис. 5. Зависимость остатков от предсказанных значений

СПИСОК ЛИТЕРАТУРЫ

1. Банк России. URL: <https://www.cbr.ru/>
2. Федеральная служба государственной статистики: URL: <https://rosstat.gov.ru/>
3. Статистика по России: URL: <https://russia.duck.consulting/>
4. Национальный экологический рейтинг: URL: <http://greenpatrol.ru/ru>
5. Ким Дж.-О., Мьюллер Ч. У. Факторный анализ: статистические методы и практические вопросы // Факторный, дискриминантный и кластерный анализ: сб. работ: пер. с англ. / под ред. И.С. Енюкова. М.: Финансы и статистика, 1989. 215 с.

Сомова Полина Анатольевна, студентка ММФ ТГУ; p.a.somova@gmail.com

Пчелинцев Евгений Анатольевич, к.ф.-м.н., доцент кафедры математического анализа и теории функций; evgen-pch@yandex.ru