УДК 004

Исследование зависимости количества автомобилей от различных факторов в регионах с помощью среды RStudio

Шайдуров Александр Александрович
Приамурский государственный университет имени Шолом-Алейхема
Студент

Баженов Руслан Иванович
Приамурский государственный университет имени Шолом-Алейхема
к.п.н., доцент, зав. кафедрой информационных систем, математики и правовой информатики

Аннотация
В статье проведено исследование зависимости количества населения от различных факторов. Для исследования была выбрана среда RStudio. С помощью этой среды был выявлен параметр, который больше всего влияет на количество машин в регионе.
Ключевые слова: RStudio, статистика, корреляция, Пирсон.

A study of the dependence of the number of cars on various factors in regions using the RStudio environment

Shaidurov Aleksandr Aleksandrovich
Sholom-Aleichem Priamursky State University
Student

Bazhenov Ruslan Ivanovich
Sholom-Aleichem Priamursky State University
Candidate of pedagogical sciences, associate professor, Head of the Department of Information Systems, Mathematics and Law Informatics

Abstract

The article investigates the dependence of the population on various factors. The RStudio environment was chosen for the study. With the help of this environment, a parameter was identified that most affects the number of machines in the region.
Keyword: RStudio, statistics, correlation, Pearson.

В наше время актуальны различные подсчёты и статистики в различных сферах, это может пригодится, для вычисления определённых показателей и коэффициентов. Все показатели могут пригодится для составления статистик по определённым регионам.

Целью исследования является анализ зависимости количества машин в регионе от различных факторов. Были рассмотрены такие факторы, как; количество населения, средняя зарплата, прирост населения за последний год и площадь региона. В качестве инструмента для вычисления была использована RStudio.

RStudio - свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R, который предназначен для статистической обработки данных и работы с графикой. RStudio написана на языке программирования C++ и использует фреймворк Qt для графического интерфейса пользователя. Большим преимуществом среды RStudio является её кроссплатформенность для операционных систем.

От других продуктов для статистической обработки данных, таких как Stata, SAS, SPSS Statistics или STATISTICA R выгодно отличается лицензией GNU GPL, подразумевающей свободное распространение, кроссплатформенностью и гибкостью — помимо осуществления стандартных вычислений существует возможность строить картограммы, создавать интерактивные веб-приложения и проводить тестирования. Кроме этого, R позволяет максимально эффективно использовать вычислительные мощности ЭВМ — вычислительная среда адаптирована для работы на высокопроизводительных кластерах и в многоядерных системах. Эти преимущества способствуют популяризации R в научной среде, а сформировавшееся сообщество позволяет оперативно получать техническую поддержку и находить ответы на возникающие вопросы.

О применении и возможностях среды RStudio можно найти в статье М.И. Калугина и С.В. Бегичева Современные возможности визуализации результатов исследований в среде R [1]. О.В. Прокофьев и И.Ю. Семочкина в статье Применение языка R и среды RStudio для математической обработки данных провели анализ возможностей применения языка программирования R и среды RStudio для математической обработки данных [2]. Провели установление географического распространения организмов в целях мониторинга окружающей среды и исследования их экологии с помощью RStudio О.В. Синчук и С.В. Буга в своей статье Картирование распространения инвазивных видов животных фауны Беларуси средствами RStudio [3]. А.А. Шарий провёл оценку факторов, влияющих на дефолт заёмщика автокредита с помощью RStudio в статье «Моделирование дефолта на рынке автокредитования» [4].

Для исследования были найдены данные 30 регионов по таким параметрам, как количество машин, количество населения, средняя зарплата, прирост населения за последний год и площадь региона. Все данные были взяты из сети Интернет среди нескольких источников.

Исследование было проведено с помощью метода корреляции Пирсона. Коэффициент корреляции Пирсона применяется для исследования взаимосвязи двух переменных на одной и той же выборке. Он позволяет определить, насколько пропорциональная изменчивость двух переменных. Данный коэффициент разработали Карл Пирсон, Фрэнсис Эджуорт и

Рафаэль Уэлдон в 90-х годах XIX века. Корреляция - статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Коэффициент корреляции изменяется в пределах от минус единицы до плюс единицы. Коэффициент корреляции Пирсона характеризует существование линейной связи между двумя величинами.

Значение коэффициента корреляции k

- $0 < k <= 0,2$ - очень слабая корреляция;
- $0,2 < k <= 0,5$ - слабая корреляция;
- $0,5 < k <= 0,7$ - средняя корреляция;
- $0,7 < k <= 0,9$ - сильная корреляция;
- $0,9 < k <= 1$ - очень сильная корреляция.

Если коэффициент отрицательный, то это означает что корреляция обратная (чем выше один параметр, тем ниже другой).

Была построена таблица данных, под параметром «cities» указаны название городов, данные которых были рассмотрены, «car» означает количество автомобилей в этом регионе, «population» – население в регионе, «salary» параметр, который означает среднюю зарплату в рублях в регионе, «growth» – прирост населения за год, «area» – параметр, означающий площадь региона в км^2.

	A	B	C	D	E	F
1	Cities	Car	Population	Salary	Growth	Area
2	Moscow	5600000	12380664	67899	50538	2561,5
3	St. Petersburg	2400000	5281579	44026	55889	1403
4	Omsk	776568	1178391	29478	291	573
5	Ekaterinburg	650104	1455904	43910	11465	468
6	Nizhny Novgorc	444106	1261666	27500	-5205	410,68
7	Kazan	433621	1231878	39300	14913	425,3
8	Chelyabinsk	421998	1198858	31500	6864	530
9	Samara	411741	1169719	29212	-1191	382
10	Krasnoyarsk	411000	1082933	38361	15999	348
11	Rostov-na-Don	396105	1125299	27500	5424	348,5
12	Ufa	392677	1115560	33700	4584	707,9
13	Khabarovsk	390777	616242	42767	5082	386
14	Krasnodar	385205	881476	26000	27628	339,3
15	Novosibirsk	379674	1602915	32484	18777	505,62
16	Permian	368898	1048005	31420	6129	803
17	Voronezh	366010	1039801	28300	7419	596,5
18	Volgograd	357486	1015586	32270	-551	859,4
19	Vladivostok	313400	606589	50082	-64	331,16
20	Saratov	253900	845300	32000	1840	394
21	Tyumen	263572	744554	39542	23979	698
22	Tolyatti	237100	710567	32000	-2052	314,8
23	Irkutsk	189104	623736	36100	312	277
24	Izhevsk	178100	646277	28218	2781	316,6
25	Barnaul	167500	633301	20918	-2284	321
26	Ulyanovsk	165300	624518	30237	3004	622,5
27	Yaroslavl	137426	608079	29505	1376	205
28	Tomsk	147600	572740	35487	3447	297,2
29	Orenburg	173848	564443	27225	1874	259
30	Kemerovo	135319	556920	30471	3844	294,8
31	Chita	99900	347088	31100	3577	534

Рис.1. Таблица найденных данных для исследования

Были введены команды для нахождения коэффициентов корреляции методом Пирсона, проводится корреляция между параметрами «car» и «population», «salary», «growth», «area».

```
cor(a$Car, a$Population, method = "pearson")
cor(a$Car, a$Salary, method = "pearson")
cor(a$Car, a$Growth, method = "pearson")
cor(a$Car, a$Area, method = "pearson")
```

Рис.2. Команды для нахождения коэффициента корреляции методом Пирсона в системе RStudio

Из полученных данных следует, что коэффициент корреляции между параметрами «car» и «population» составляет 0,9957135, между «car» и «salary» 0,7575602, между «car» и «growth» 0,7551809 и между «car» и «area» 0,9361552. Вычисленные коэффициенты корреляции указывают на то, что найденные данные для исследования являются подходящими.

Для нахождения функции линейной регрессии были введены следующие команды.

```
M=lm(Car ~ Population, a )
summary(M)

M=lm(Car ~ Salary, a )
summary(M)

M=lm(Car ~ Growth, a )
summary(M)

M=lm(Car ~ Area, a )
summary(M)
```

Рис.3. Команды для получения функции линейной регрессии

После ввода команд были получены следующие результаты.

```
Call:
lm(formula = Car ~ Population, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-270101  -35893  -25233    6831  322030

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.740e+04  2.118e+04  -4.127 0.000298 ***
Population   4.599e-01  8.073e-03  56.966  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 97360 on 28 degrees of freedom
Multiple R-squared:  0.9914,    Adjusted R-squared:  0.9911
F-statistic:  3245 on 1 and 28 DF,  p-value: < 2.2e-16
```

Рис.4. Функция линейной регрессии параметров Car и Population

```
Call:
lm(formula = Car ~ Salary, data = a)

Residuals:
     Min       1Q   Median       3Q      Max
-1638339  -433130   -32713   313211  2088006

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.434e+06  5.047e+05  -4.822 4.51e-05 ***
Salary       8.757e+01  1.426e+01   6.141 1.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687100 on 28 degrees of freedom
Multiple R-squared:  0.5739,    Adjusted R-squared:  0.5587
F-statistic: 37.71 on 1 and 28 DF,  p-value: 1.253e-06
```

Рис 5. Функция линейной регрессии параметров Car и Salary

```
Call:
lm(formula = Car ~ Growth, data = a)

Residuals:
     Min       1Q   Median       3Q      Max
-1207154  -151018   -30167   176986  2757785

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.511e+04  1.488e+05   0.572    0.572
Growth      5.455e+01  8.949e+00   6.096 1.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 690000 on 28 degrees of freedom
Multiple R-squared:  0.5703,    Adjusted R-squared:  0.555
F-statistic: 37.16 on 1 and 28 DF,  p-value: 1.414e-06
```

Рис 6. Функция линейной регрессии параметров Car и Growth

```
Call:
lm(formula = Car ~ Area, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-876591  -99702  129996  215730  698071

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -617842.6   107946.7  -5.724 3.86e-06 ***
Area           2154.9      152.9  14.089 3.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 370100 on 28 degrees of freedom
Multiple R-squared:  0.8764,    Adjusted R-squared:  0.872
F-statistic: 198.5 on 1 and 28 DF,  p-value: 3.091e-14
```

Рис 7. Функция линейной регрессии параметров Car и Area

Для построения нужных графиков была загружена библиотека "ggplot2".

```
install.packages("ggplot2", dependencies = TRUE)
library(ggplot2)
setwd("~/Shaidurov")
```

Рис.8. Загрузка библиотеки "ggplot2"

Для построения графиков зависимости были использованы следующие команды.

```
ggplot ( a, aes(Car, Population))+geom_line()+geom_point(size=2)+geom_smooth(method="lm")
ggplot ( a, aes(Car, Salary))+geom_line()+geom_point(size=2)+geom_smooth(method="lm")
ggplot ( a, aes(Car, Growth))+geom_line()+geom_point(size=2)+geom_smooth(method="lm")
ggplot ( a, aes(Car, Area))+geom_line()+geom_point(size=2)+geom_smooth(method="lm")
```

Рис.9. Команды для получения графиков зависимости

Были получены следущие графики.

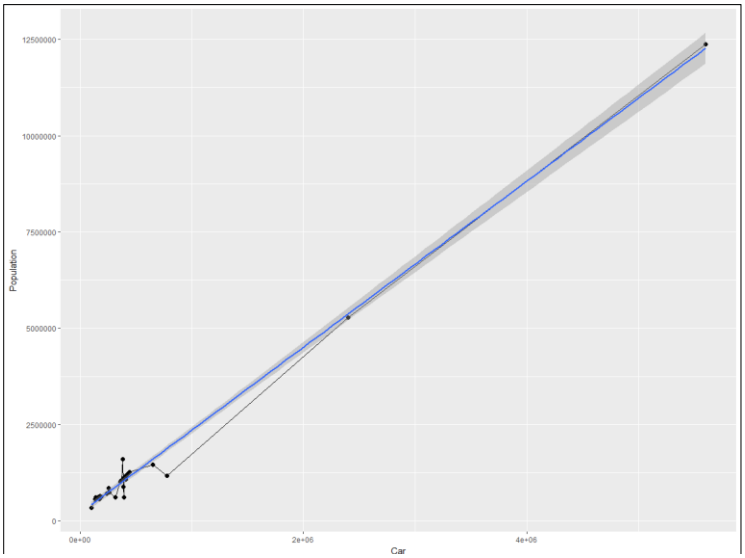

Рис.10. Команды для получения графиков зависимости параметров Car и Population

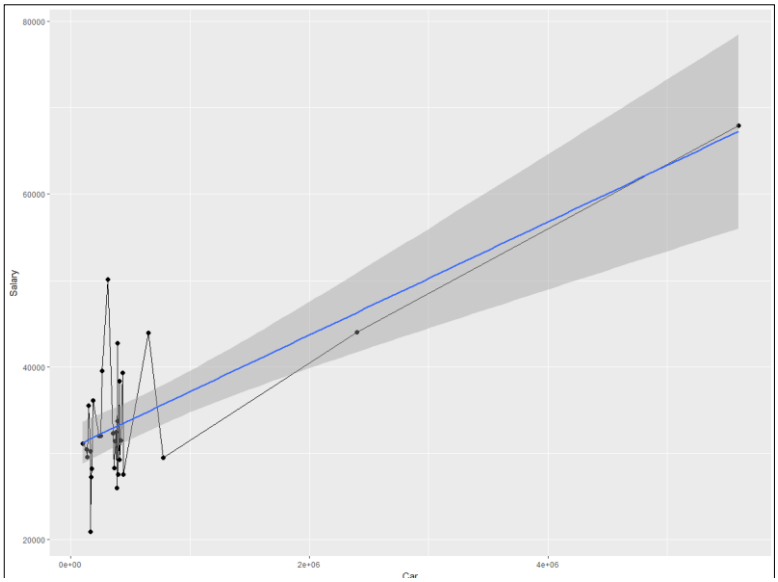

Рис.11. График зависимости параметров Car и Growth

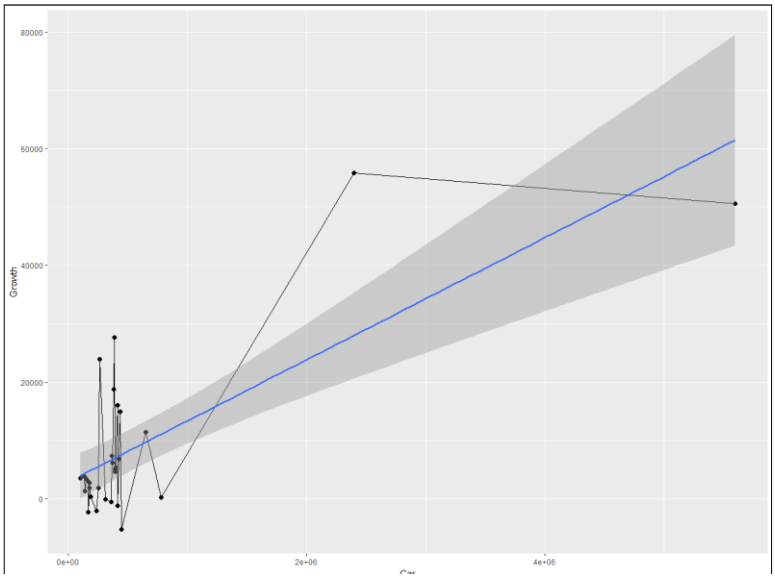

Рис.12. График зависимости параметров Car и Salary

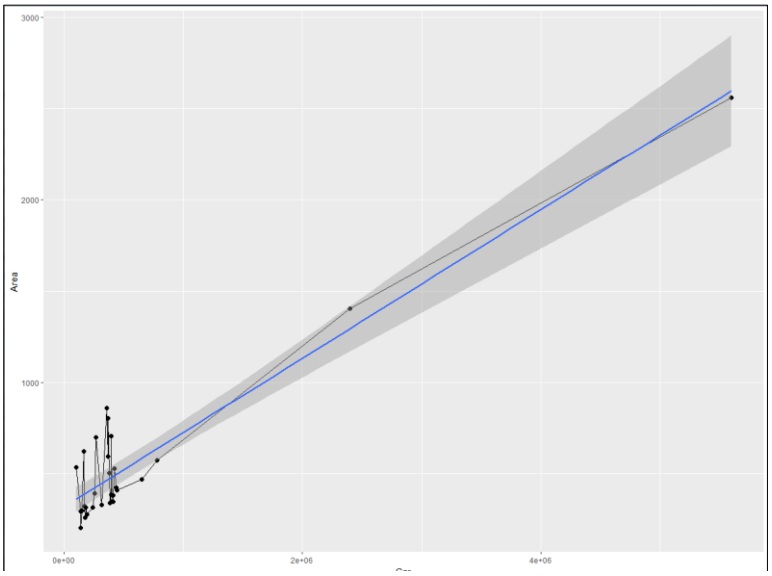

Рис.13. График зависимости параметров Car и Area

Из результатов следует, что «population» (население в регионе) – это параметр от которого больше всего зависит итоговое значение количества автомобилей, следующие по зависимости, находятся «salary» (средняя зарплата в регионе), потом «growth» (прирост населения за год в регионе), меньше всего влияет параметр «area» (площадь региона).

В результате исследования были получены при помощи корреляции Пирсона результаты влияния различных факторов на количество автомобилей в регионах. Эти результаты могут понадобится для вычисления определённых данных

Библиографический список

1. Калугина М.И., Бегичева С.В. Современные возможности визуализации результатов исследований в среде R. // В сборнике: BI-технологии и

корпоративные информационные системы в оптимизации бизнес-процессов Материалы IV Международной научно-практической очно-заочной конференции. Ответственные за выпуск: Д.М. Назаров, С.В. Бегичева, Е.В. Зубкова. 2016. С. 51-55 URL: https://elibrary.ru/item.asp?id=30092779

2. Прокофьев О.В., Семочкина И.Ю. Применение языка R и среды RStudio для математической обработки данных. // Современные информационные технологии. 2017. № 25 (25). С. 47-51. URL: https://elibrary.ru/item.asp?id=30092779
3. Синчук О.В., Буга С.В. Картирование распространения инвазивных видов животных фауны Беларуси средствами RStudio. // В сборнике: Международный конгресс по информатике: информационные системы и технологии материалы международного научного конгресса. С. В. Абламейко (гл. редактор). 2016. С. 185-188. URL: https://elibrary.ru/item.asp?id=28354860
4. Шарый А.А. Моделирование дефолта на рынке автокредитования. // Достижения науки и образования. 2016. № 6 (7). С. 47-49. URL: https://elibrary.ru/item.asp?id=26240710
5. Курилов Ф. М. Использование языка R для эконометрического моделирования и обеспечения расчетов // Проблемы и перспективы экономики и управления: материалы III Междунар. науч. конф., г. Санкт-Петербург, 2014 г. URL: https://moluch.ru/conf/econ/archive/131/6801/
6. RStudio // Wikipedia URL: https://ru.wikipedia.org/wiki/RStudio (дата обращения 11.04.18)
7. Корреляция // Wikipedia URL: https://ru.wikipedia.org/wiki/Корреляция (дата обращения 11.04.18)
8. 100 Крупнейших городов России по населению 2017 список РФ // statdata URL: http://www.statdata.ru/largest_cities_russia (дата обращения 11.04.18)
9. Сколько автомобилей в России и в Москве в 2018 году // avtojurcon URL: http://avtojurcon.ru/sovety/ckolko-avtomobilej-v-rossii-i-v-moskve-v-2017-godu.html (дата обращения 11.04.18)