УДК 004

Выявление зависимостей в данных о рождаемости с помощью линейной регрессии

Азаров Андрей Евгеньевич
Приамурский государственный университет им. Шолом-Алейхема
Студент

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

Аннотация
Данное исследование проводилось с целью изучить данные о рождаемости в Австралии и найти какие-либо зависимости.
Ключевые слова: R Studio, анализ, линейная регрессия.

Identification dependent data in fertility using linear regression

Azarov Andrey Evgenevich
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
This study was conducted to study the data on fertility in Australia and find any dependencies.
Key words: R Studio, analysis, linear regression.

Во всех развитых странах, сбор больших объёмов данных стал уже неотъемлемой частью любого процесса, будь то экономического или социального. В основном сбором данных занимаются частные компании для собственных исследований, например, узнать насколько хорошо продаётся тот или иной товар, но так же, сбором данных занимаются и государственные органы, для выявления различных статистик: демографических, социальных, экономических, психологический и так далее.

Целью данного исследования является изучение зависимостей между количеством рождённых детей от года их рождения, возраста матерей и штата в котором рождались дети.

Исследование проводится на основе данных о рождаемости в Австралии, собранных австралийскими учеными с 1999 года по 2009 год.

Изучение демографических процессов является одним из приоритетных исследований проводимых государственными органами, ведь из результатов данных исследований можно сделать множество выводов, таких как: социальное и материальное благополучие населения, прогноз роста или спада количества населения в ближайшие несколько лет, отдельные показатели рождаемости в каждом штате, в каком возрасте женщины решают стать матерями и меняется ли это возраст со временем.

Анализом данных с помощью методов линейной регрессии занимаются в России очень давно, например В. В. Дергунов в 1999 году написал статью изучил анализ динамики ВНП методом линейной регрессии [1]. Исследования В.Т. Тарушкин и др. за 2011 год о линейной интервальной регрессии для ВВП России содержат сравнения прогнозов правительством РФ и их личными исследованиями, которые показали, что прогнозируемые данные совпали с реальными [2]. Исследование - линейная регрессия и «коэффициент корреляции» 2013 года В.А.Падве о вопросе не информативности «коэффициента корреляции», вычисляемого по массивам данных, имеющих значимую функциональную зависимость [3]. Исследование Е.А.Дмитриева представило линейную регрессию как один из методов машинного обучения [4]. Линейная регрессия параметров артериального давления для определения риска развития вторичной гипотензии 2013 М.В. Войтикова представляло метод классификации медицинских сигналов артериального давления по степени риска развития гипотензии, основанный на построении линейной регрессионной модели для параметров артериального давления с последующим применением классификатора по методу опорных векторов [5].

После детального изучения собранных данных учёными, они были отсортированы и оформлены в форму, удобную для чтения в программе R Studio и для изучения данных интеллектуальным методом анализа линейной регрессии.

	years	yy	old15	old20	old25	old30	old35	old40	old45
1	1999	0.3	19.0	64.5	112.7	110.0	49.3	9.3	0.3
2	2000	13.3	16.9	61.9	111.2	112.9	50.9	9.6	0.4
3	2001	26.3	17.1	59.8	106.4	108.7	50.9	9.8	0.4
4	2002	39.3	16.5	58.4	106.9	112.8	54.3	10.7	0.5
5	2003	52.3	15.0	55.9	105.6	113.4	57.2	10.7	0.4
6	2004	65.3	15.0	54.0	103.6	113.9	58.9	11.2	0.5
7	2005	78.3	13.4	50.6	101.5	118.5	62.7	11.4	0.6
8	2006	91.3	13.2	49.7	100.3	120.2	64.8	11.9	0.6
9	2007	104.3	12.3	49.7	100.2	122.5	67.7	13.0	0.8
10	2008	117.3	13.9	52.2	101.1	125.8	72.2	14.8	0.8
11	2009	125.8	13.0	47.7	96.7	122.3	70.1	14.8	0.8

Рис. 1.

На рисунке 1 изображена часть исходных данных после форматирования, которые была скачаны с официального сайта Австралийского университета. В первой колонке указаны года исследований, во второй первое значение является самым минимальным значением среди всех значений в таблице, а в 11-ой максимальное, в остальных столбцах значения

Далее можно увидеть и изучить результаты проведенной работы. На рисунке 2 изображен график зависимости между 15-ти летними девушками которые родили детей по отношению к годам. На графике видно, что с 1999 года по 2009 год, 15-ти летние стали намного меньше рожать. После проведения регрессионного анализа была выявлена хорошая регрессия, так как R квадрат равен 0.83. Значения по оси у представляют собой количество рождённых детей, а по оси х год рождения этих детей. Синяя линия это линия тренда, с помощью которой можно спрогнозировать дальнейшие значения.

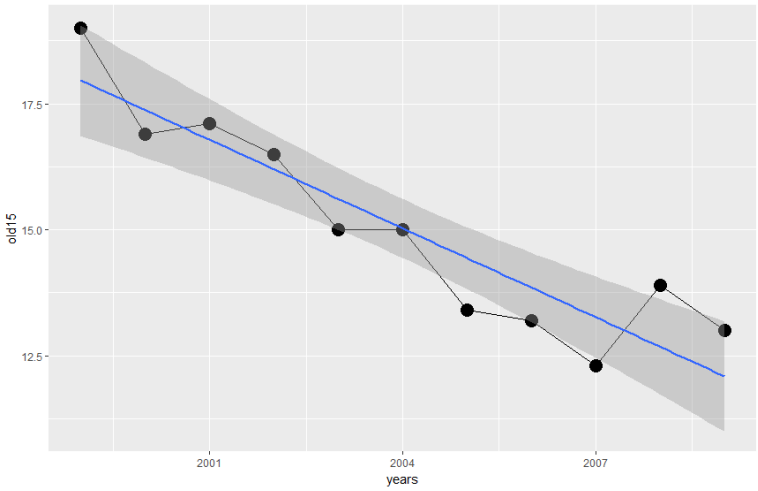

Рис. 2.

Далее на рисунке 3 проведен аналогичный анализ, где R квадрат равен 0.84, следовательно, регрессия хорошая, но уже видно большое отличие, во всех предыдущих графиках шёл спад рождаемости среди женщин до 30 лет, а после 30 рождаемость начала расти.

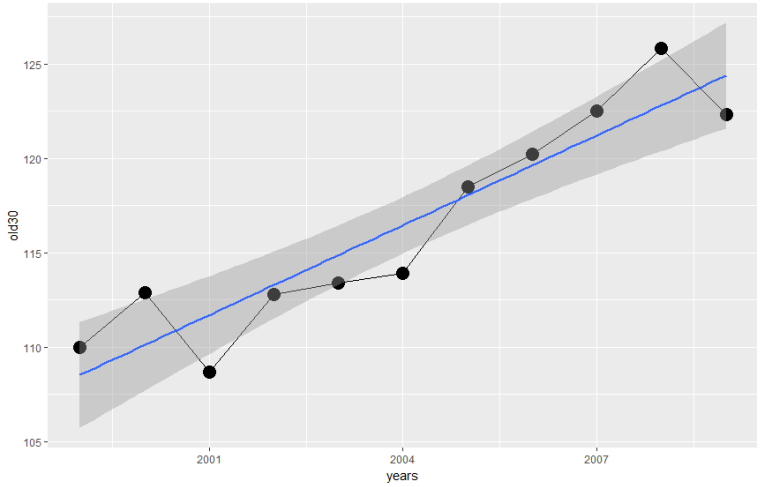

Рис. 3.

На рисунке 4 изображен график, содержащий в себе данные обо всех возрастных группах одновременно. Средний R квадрат по всем возрастам равен 88,9.

По рисунку видны общие тенденции, даже если и рождаемость падает в определенной возрастной группе, она всё равно может иметь довольно высокие показатели относительно рождаемости.

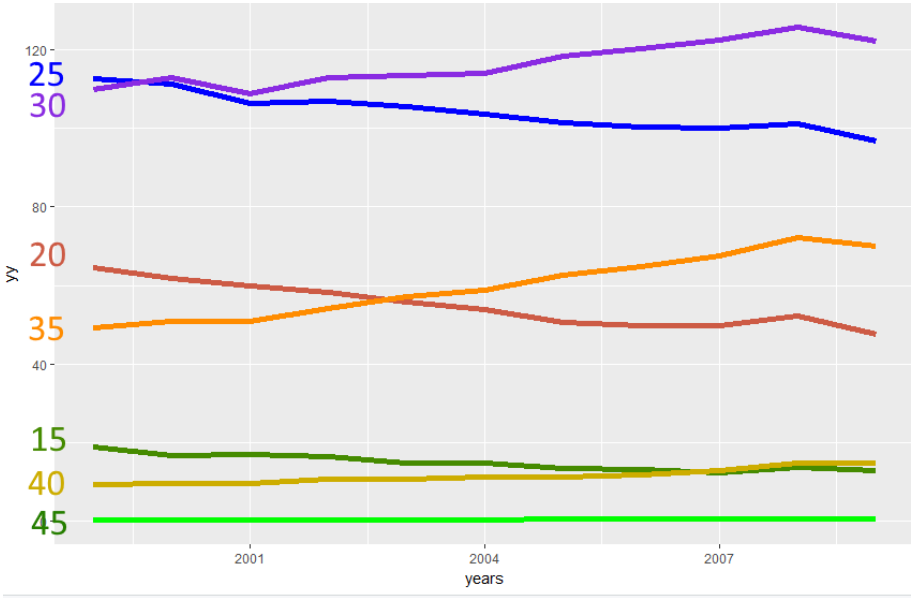

Рис 4.

Для реализации последнего графика использовались возможности языка R, программного обеспечения R Studio и библиотеки ggplot2 для данной среди разработки.

```
15  g=ggplot(data, aes(years, yy))
16
17  g=g+geom_line(aes(y=old15), colour="chartreuse4",size=2)
18  g=g+geom_line(aes(y=old20), colour="coral3",size=2)
19  g=g+geom_line(aes(y=old25), colour="blue",size=2)
20  g=g+geom_line(aes(y=old30), colour="blueviolet",size=2)
21  g=g+geom_line(aes(y=old35), colour="darkorange",size=2)
22  g=g+geom_line(aes(y=old40), colour="gold3",size=2)
23  g=g+geom_line(aes(y=old45), colour="green",size=2)
24
25  plot(g)
```

Рис. 5 Отрывок исходного кода

На рисунке 5 изображён отрывок исходного кода, в пятнадцатой строке объявляется график, где по оси x выводятся года исследований, а по оси y выводится всё множество чисел, которые имеют все следующие столбцы данных.

В заключении можно сказать, что после проведения данного исследования было получено множество данных и математически доказана взаимосвязь определенных величин друг с другом. Доказано, что существует взаимосвязь между возрастом матери и количеством детей которые были

рождены в этот год, среди матерей этого же возраста. Также наглядно продемонстрированы различные взаимосвязи и выявлены линии тренда со всеми отдельными возрастными группами на графиках, что может помочь спрогнозировать рождаемость в следующие года, также узнали среди каких возрастных групп рождаемость будет идти на спад, а среди каких на подъём.

Библиографический список

1. Дергунов В.В. Анализ динамики внп методом линейной регрессии // вестник финансовой академии. 1999. №4. С. 98-108.
2. Тарушкин В.Т., Тарушкин П.В., Тарушкина Л.Т. Линейная интервальная регрессия для ввп россии за 2010 год // Международный журнал прикладных и фундаментальных исследований. 2011. №6. С. 136.
3. Падве В.А. Линейная регрессия и «коэффициент корреляции» // Интерэкспо гео-сибирь. 2013. №3. С. 78-81.
4. Войтикова М.В., Хурса Р. В. линейная регрессия параметров артериального давления для определения риска развития вторичной гипотензии // Весці нацыянальнай акадэміі навук беларусі. Серыя фізіка-матэматычных навук. 2013. №1. С. 117-122.
5. R Studio URL: https://www.rstudio.com (дата обращения: 21.4.2018).
6. Colors in R // Department of Statistics URL: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf (дата обращения: 21.4.2018).
7. R Documentation URL: https://www.rdocumentation.org (дата обращения: 21.4.2018).