УДК 519.23+519.25(045)

Анализ данных и регрессионное моделирование с применением языков программирования Python и R

Чибирова Марина Эльбрусовна,
студентка
финансово-экономического факультета,
Финансовый университет, Владикавказский филиал,
Владикавказ, Россия
m.chibirova@rambler.ru

Аннотация. *В статье рассматривается методика регрессионного анализа данных и разработки математических моделей, которые используются для прогнозирования процессов. Автор подробно показывает каждый этап разработки регрессионных моделей, используя метод наименьших квадратов. Особое внимание уделено оценке адекватности разработанных математических моделей (оценка адекватности моделей проводилась по F-критерию Фишера — Снедекора и по коэффициенту корреляции). Проведены расчеты прогнозных оценок. Анализ данных, разработка регрессионных моделей и прогнозирование (в том числе с помощью метода скользящей матрицы) реализованы на двух современных языках программирования: Python и R, которые в настоящее время являются наиболее востребованными для решения подобных задач. В заключение кратко сформулированы основные преимущества проведения математических расчетов и реализации алгоритма на двух языках программирования для более глубокого понимания процесса анализа данных.*

Ключевые слова: *регрессионный анализ данных; математическая статистика; малая выборка данных; регрессионное моделирование; прогнозирование; метод скользящей матрицы; язык программирования R; язык программирования Python*

Data Analysis and Regression Modelling Using the Python and R Programming Languages

Chibirova Marina Elbrusovna,
student,
Faculty of Finance and Economics,
Financial University (Vladikavkaz Branch),
Vladikavkaz, Russia
m.chibirova@rambler.ru

Abstract. *This article discusses the method of regression data analysis and the development of mathematical models that are used to predict processes. The author describes in detail each step of the development of regression models using the least squares method. Particular attention the author paid to assess the adequacy of the developed mathematical models (assessment of the adequacy of the models was carried out by the Fisher-Snedecor F-test and by the correlation coefficient). The author has calculated the forecast estimates. Data analysis, regression model development and forecasting (including using the*

Научный руководитель: ***Дзгоев А.Э.,*** *кандидат технических наук, доцент, доцент кафедры «Математика и информатика», Финансовый университет, Владикавказский филиал, Владикавказ, Россия.*

sliding matrix method) are implemented in two modern programming languages: Python and R, which are currently the most popular for solving such tasks. In conclusion, the article briefly describes the main advantages of performing mathematical calculations and implementing the algorithm in two programming languages for a deeper understanding of the data analysis process.

Keywords: *regression data analysis; mathematical statistics; small data sample; regression modelling; prediction; sliding matrix method; R programming; Python programming language*

Ученому по анализу данных необходимо знать научные методы обработки данных, уметь разрабатывать адекватные математические (регрессионные) модели для прогнозирования поведения систем или процессов и владеть навыками программирования на нескольких языках.

Специалисты по Data Science используют для анализа данных методы математической статистики, теории вероятностей и регрессионного анализа данных.

С помощью теории вероятностей специалисты находят вероятности «сложных» событий через вероятности «простых», связанных с ними событий, а с помощью математической статистики оценивают вероятности этих событий по выборке данных. Цель использования указанных методов состоит в изучении закономерностей массовых случайных явлений и прогнозировании их характеристик, минуя сложное (а зачастую и невозможное) исследование отдельного случайного явления [1, с. 13].

В практике исследований имеющиеся данные не всегда можно считать выборкой из многомерной нормальной совокупности, например когда линия регрессии не является прямой. В этом случае пытаются определить кривую, которая дает наилучшее (в смысле наименьших квадратов) приближение к исходным данным. Соответствующие методы приближения получили название регрессионного анализа [1, с. 457].

Регрессионный анализ является основой для прогнозирования поведения случайного явления за пределами данных, а математические модели, разработанные с помощью данного анализа, называются регрессионными моделями.

Корреляционно-регрессионный анализ является широко используемым методом прогнозирования. Функция регрессии выражает отношение зависимой переменной к одной или нескольким независимым переменным. С другой стороны, корреляция предназначена для измерения направления и интенсивности этих отношений. Обычно только те переменные, которые показывают значительный уровень корреляции, подвергаются регрессионному анализу [2, с. 160].

Математическая формализация расчетов

В статье описан процесс построения регрессионных моделей для прогнозирования на примере анализа общего количества доменных имен в зонах.ru и.рф. Домен — это набор символов, который составляет имя сайта. Это имя нельзя купить раз и навсегда, его регистрируют на один год с возможностью пролонгации.

Актуальность разработки связана с индустриальной потребностью в регулярном получении точных данных об объемах и направлениях развития интернет-рынков в России. Также прогнозирование общего количества доменов позволит регистраторам доменных имен рассчитывать свою прибыль.

В ходе работы были определены зависимая переменная (Y) и факторы, влияющие на Y,— независимые переменные (X):

1) X_0 — фиктивная переменная (для формирования свободного коэффициента B_0 регрессионной модели). Фиктивные переменные позволяют строить и оценивать так называемые кусочно-линейные модели, которые можно применить для исследования структурных изменений [3, с. 113–115];

2) X_1 — независимая переменная, период с 2003 по 2015 г.;

3) X_2 — независимая переменная, численность интернет-аудитории в России за месяц, млн чел.;

4) X — объединенная матрица независимых переменных X_0, X_1, X_2;

5) Y — зависимая переменная, общее число доменных имен в зонах .ru и .рф, млн (*рис. 1*).

Для расчетов было отобрано 13 наблюдений, доступных в открытых источниках (минимальное рекомендуемое число наблюдений для анализа данных на автокорреляцию — 12).

X0 =

	0
0	1
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1

X1 =

	0
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10
10	11
11	12
12	13

X2 =

	0
0	11.6
1	14.2
2	20.1
3	23.9
4	27.5
5	33.3
6	41.1
7	50.3
8	57.8
9	64.4
10	68.7
11	73.8
12	76

X =

	X0	X1	X2
0	1	1	11.6
1	1	2	14.2
2	1	3	20.1
3	1	4	23.9
4	1	5	27.5
5	1	6	33.3
6	1	7	41.1
7	1	8	50.3
8	1	9	57.8
9	1	10	64.4
10	1	11	68.7
11	1	12	73.8
12	1	13	76

Y =

	0
0	0.2
1	0.3
2	0.4
3	0.7
4	1.2
5	1.9
6	2.6
7	3.8
8	4.6
9	5
10	5.7
11	5.7
12	5.9

Рис. 1. **Матрицы *X* и *Y***

Источник: составлено автором на основе данных Российской ассоциации электронных коммуникаций. URL: http://old.raec.ru/upload/files/EconomicaRunetaItogy2016.pdf (дата обращения: 02.11.2018) и данных Координационного центра национального домена сети Интернет. URL: https://cctld.ru/files/stats/report_ru-2017_rus2.pdf (дата обращения: 02.11.2018).

N = 13; k = 3, где N — число наблюдений; k — число коэффициентов уравнения регрессии.

Проведен расчет коэффициентов регрессионной модели по формуле

$$B = \left(X^T X\right)^{-1} X^T Y, \quad (1)$$

где B — коэффициенты регрессионной модели.

В результате (1) получены 3 коэффициента регрессии:

$$B = \begin{pmatrix} -1{,}361 \\ -0{,}35 \\ 0{,}156 \end{pmatrix}.$$

Разработана регрессионная модель, которая имеет вид

$$Y = -1{,}361 - 0{,}35 \times X_1 + 0{,}156 \times X_2. \quad (2)$$

Оценка адекватности разработанной регрессионной модели

Оценка адекватности разработанной регрессионной модели проводилась по *F*-критерию Фишера — Снедекора (*F-test*), где необходимо проанализировать дисперсию и сравнить расчетное значение критерия *FR* с табличным (критическим) значением $F_{табл.}$

Если расчетное значение *F*-критерия Фишера — Снедекора больше табличного, то делается вывод о том, что разработанная регрессионная модель адекватна. В каждом случае критический уровень зависит от числа независимых переменных и от числа степеней свободы ($N - k$) [4, с. 117].

Расчетные значения зависимой переменной были рассчитаны по формуле

$$YR = X \times B. \quad (3)$$

Рассчитано среднее арифметическое зависимой переменной:

$$YSR = \frac{\sum Y}{N} = 2{,}923. \quad (4)$$

Из уравнения (4) найдена дисперсия зависимой переменной:

$$DY = \frac{\sum \left(Y - YSR\right)^2}{N - 1} = 5{,}159. \quad (5)$$

Далее рассчитана дисперсия адекватности по формуле

$$Dad = \frac{\sum \left(Y - YR\right)^2}{N - k} = 0{,}039. \quad (6)$$

Затем найдено расчетное значение *F*-критерия Фишера:

$$FR = \frac{DY}{Dad} = 132{,}188. \quad (7)$$

Для проверки регрессионного уравнения на адекватность сравним найденное расчетное зна-

чение с табличным значением *F*-критерия Фишера [4, с. 435]:

$$F = qF\left(0{,}95, N-1, N-k\right) = 2{,}913. \quad (8)$$

Вывод: в связи с тем, что расчетное значение (132,188) больше табличного значения *F*-критерия Фишера (2,913), разработанная регрессионная модель признана адекватной на уровне значимости 0,05 или с доверительной вероятностью p = (1–0,05) × 100 = 95%.

Полученная адекватная регрессионная модель позволяет провести детальное исследование изучаемого объекта и прогнозирование его поведения с учетом различных факторов [5, с. 36.].

Далее проведено ранжирование независимых переменных по силе их влияния на зависимую переменную для уравнения регрессии (2) (*табл. 1*).

Для выявления рангов:

1. Переписываем диагональные элементы матрицы, обратной матрице нормальных уравнений *G*, и вычисляем квадратный корень каждого диагонального значения ($\sqrt{G_{j,j}}$, j = 3).

2. Рассчитываем среднеквадратическую ошибку *S* ($S = \sqrt{DY} = 2{,}271$).

3. Рассчитываем значения рангов независимых переменных:

$$X_j = \frac{\left|B_j\right|}{S \times \sqrt{G_{j,j}}}, \quad (9)$$

Таким образом, значения рангов независимых переменных равны:

1) для X_0 = 0,966;
2) для X_1 = 0,636;
3) для X_2 = 0,24.

В результате было выявлено, что на первом месте по силе влияния расположился фактор X_0, на втором месте — X_2 (численность интернет-аудитории в России за месяц, млн чел.), на третьем месте — X_1 (период с 2003 по 2015 г.).

Далее проведен анализ данных на автокорреляцию.

Автокорреляция — статистическая взаимосвязь между последовательностями величин одного ряда, взятыми со сдвигом, например для случайного процесса — со сдвигом по времени.

Метод наименьших квадратов в случае автокорреляции возмущений дает несмещенные и состоятельные оценки параметров, однако их интервальные оценки могут содержать грубые ошибки. В случае выявления автокорреляции возмущений необходимо вернуться к проблеме выбора функции тренда, пересмотреть набор включенных в него переменных и провести все расчеты заново. Если автокорреляция в статистических данных присутствует, то такие данные не пригодны для прогнозирования.

Таблица 1

Ранги независимых переменных по силе их влияния на зависимую переменную

Ранг	Сила влияния	X_j
I	0,966	X_0
II	0,636	X_2
III	0,24	X_1

Источник: составлено автором.

Наиболее простым и достаточно надежным критерием определения автокорреляции возмущений является критерий Дарбина — Уотсона (*d*-статистика). С помощью этого критерия проверяется гипотеза об отсутствии автокорреляции между соседними остаточными членами ряда лагом, равным 1.

Найдены отклонения расчетных значений от экспериментальных:

$$e = Y - YR. \quad (10)$$

Статистика Дарбина — Уотсона имеет вид

$$d = \frac{\sum\left(e_0 - e_1\right)^2}{\sum e^2} = 2{,}136. \quad (11)$$

Для *d*-статистики существуют верхняя $d_в$ = 1,82 и нижняя $d_н$ = 0,72 критические границы на уровне значимости α = 0,05. Если $d_в < d < (4 - d_в)$, автокорреляция отсутствует, следовательно, данные зависимой переменной *Y* пригодны для прогнозирования [1, с. 511].

Вывод: так как статистика Дарбина — Уотсона (2,136) больше, чем верхняя критическая граница (1,82), и меньше, чем (4–1,82), то гипотеза об отсутствии автокорреляции принимается.

Таблица 2

Результаты прогноза и прогнозные ошибки

Год	X_2	Y	YP	ΔY	Отн. ошибка |%|
2016	80,5	6,3	6,297	0,003	0,198

Источник: составлено автором.

Расчет прогнозных оценок

Одна из важнейших целей моделирования заключается в прогнозировании поведения исследуемого объекта или процесса. Для регрессионных моделей термин «прогнозирование» имеет более широкое значение. Данные могут не иметь временной структуры, но и в этих случаях вполне может возникнуть задача оценки значения зависимой переменной для некоторого набора независимых переменных, которых нет в исходных наблюдениях. Именно в этом смысле — как построение оценки зависимой переменной — и следует понимать прогнозирование в эконометрике [3, с. 204].

Проведен расчет прогнозных значений общего числа доменных имен в зонах.рф и.ru (*YP*) на 2016 г. на основе разработанной регрессионной модели (2) и рассчитан доверительный интервал коридора ошибок:

$$Y_{min} = 5{,}844,\ YP = 6{,}263,\ Y_{max} = 6{,}681.$$

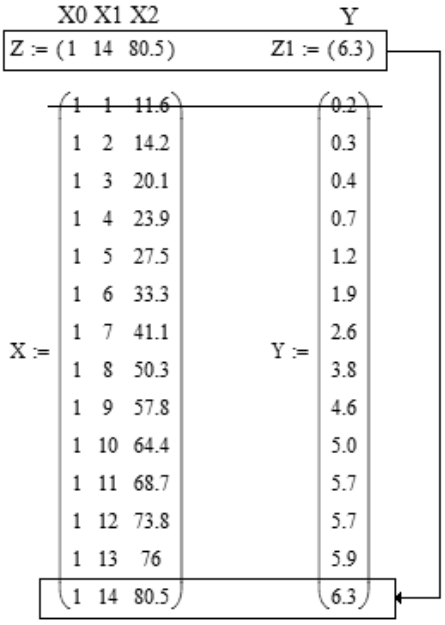

Рис. 2. **Иллюстрация метода скользящей матрицы**

Источник: составлено автором.

Среднемесячная численность интернет-аудитории в России (X_2) за 2016 г. равна 80,5 млн фактическое число доменных имен в зонах .рф и .ru на 2016 г.— 6,3 млн[1], прогнозное значение — 6,263 млн.

При обработке временных рядов, как правило, наиболее ценной бывает информация последнего периода, так как необходимо знать, как будет развиваться тенденция, существующая в данный момент, а не тенденция, сложившаяся в среднем на всем рассматриваемом периоде [6, с. 244].

Для расчетов прогнозной оценки на следующий период будет применен метод скользящей матрицы (*рис. 2*).

Метод скользящей матрицы состоит в последовательном исключении первых строк матриц независимых и зависимой переменных и добавлении новых строк с фактическими значениями переменных. Так как матрицы независимых переменных (X) и зависимой переменной (Y) изменят свои значения, необходимо проверить на адекватность новую полученную регрессионную модель [7, с. 30–37]. Таким образом, метод скользящей матрицы позволяет учесть степень «устаревания» данных, что делает прогноз более корректным.

Рассчитаны абсолютная (ΔY) и относительная (отн. ошибка |%|) ошибки прогноза (*табл. 2*):

$$\Delta Y = Y - YP = 0{,}003, \qquad (12)$$

$$\text{отн. ошибка}|\%| = \frac{|\Delta| \times 100}{Y} = 0{,}198. \qquad (13)$$

Вывод: в связи с тем, что рассчитанные прогнозные оценки зависимой переменной (*YP*) попадают в доверительный интервал коридора ошибок

[1] Координационный центр национального домена сети Интернет «Российское доменное пространство 2016: итоги и перспективы развития». URL: https://cctld.ru/files/stats/report_ru-2017_rus2.pdf (дата обращения: 02.11.2018).

Поиск минимума

Начальные приближения

X1 := 1 X2 := 1

Given

$1 \le X1 \le 13$

$11.6 \le X2 \le 76$

Q := Minimize(f ,X1,X2)

$Q = \begin{pmatrix} 1 \\ 11.6 \end{pmatrix}$

f(1,11.6) = 3.517

Поиск максимума

Начальные приближения

X1 := 1 X2 := 11.6

Given

$1 \le X1 \le 13$

$11.6 \le X2 \le 76$

Q := Maximize(f ,X1,X2)

$Q = \begin{pmatrix} 13 \\ 76 \end{pmatrix}$

f(13,76) = 17.744

Рис. 3. **Поиск экстремумов функции**

Источник: составлено автором.

Таблица 3

Данные для расчетов

X_0	X_1	X_2	Y
1	1	11,60	0,2
1	2	14,20	0,3
1	3	20,10	0,4
1	4	23,90	0,7
1	5	27,50	1,2
1	6	33,30	1,9
1	7	41,10	2,6
1	8	50,30	3,8
1	9	57,80	4,6
1	10	64,40	5
1	11	68,70	5,7
1	12	73,80	5,7
1	13	76,00	5,9
1	14	80,50	6,3

Источник: составлено автором на основе данных Российской ассоциации электронных коммуникаций. URL: http://old.raec.ru/upload/files/EconomicaRunetaItogy2016.pdf (дата обращения: 02.11.2018) и данных Координационного центра национального домена сети Интернет. URL: https://cctld.ru/files/stats/report_ru-2017_rus2.pdf (дата обращения: 02.11.2018).

Таблица 4

Импорт данных, расчет коэффициентов уравнения регрессии и проверка математической модели на адекватность в R

```
library(xlsx)
maindata <- read.xlsx(«D:/data.xls», 1)
X <- matrix(c(maindata[1:13, 1:3]))
Y <- matrix(c(maindata[1:13, 4]))
B <- solve(t(X)%*%X)%*%t(X)%*%Y
YR <- X%*%B
YSR <- sum(Y) / N
DY <- (sum((Y - YSR)^2)) / (N – 1)
Dad <- (sum((Y - YR)^2)) / (N – k)
FR <- DY / Dad
F <- qf(0.95, N — 1, N — k)
```

```
> B
           [,1]
[1,] -1.3612608
[2,] -0.3503387
[3,]  0.1556375
> YSR
[1] 2.923077
> DY
[1] 5.15859
> Dad
[1] 0.0390246
> FR
[1] 132.1881
> F
[1] 2.912977
```

Рис. 4. **Результат выполнения кода из *табл. 4***

Таблица 5

Анализ данных на автокорреляцию в R

```
e <- Y — YR
e0 <- matrix(e[1:12])
e1 <- matrix(e[2:13])
d <- (sum((e0 — e1)^2)) / sum(e^2)
```

```
> d
[1] 2.135257
```

Рис. 5. **Результат выполнения кода из *табл. 5***

и регрессионная модель адекватна, прогнозное значение можно считать корректным.

Проведена оптимизация. Решение задач оптимизации является важнейшей сферой применения численных математических методов. Задача оптимизации — найти такие значения оптимизирующих параметров, которые бы соответствовали экстремуму функции оптимизации при соблюдении ограничений на оптимизирующие параметры. Оптимальными называют параметры процесса, позволяющие получить наилучший желаемый результат в рамках выбранного критерия оптимизации и ограничений.

Функция оптимизации:

$$f(X_1, X_2) = -1{,}361 - 0{,}35 \times X_1 + 0{,}156 \times X_2 . \quad (14)$$

Установлены начальные приближения и двухсторонние ограничения для переменных, найдены минимальное и максимальное значения функции (*рис. 3*). Расчеты были проведены в математическом пакете Mathcad.

В результате проведения оптимизации рассчитано минимальное (3,517) и максимальное (17,744) значения функции $f(X_1, X_2)$.

Реализация расчетов на языке программирования R

Расчет выполнен с помощью интегрированной среды разработки (IDE) R-Studio для языка программирования R и пакетов “plyr”, “Rcpp”, “ggplot2”, “reshape2”, “xlsx”, “xlsxjars”.

Пакеты — это собрания функций R, данных и скомпилированного программного кода. Пакеты необходимо скачивать и устанавливать. После установки они загружаются во время сессии по мере необходимости [8, с. 45]. Только официальный репозиторий R насчитывает более 4300 пакетов.

Язык программирования R позволяет обрабатывать большие объемы данных. Для удобства эти данные были импортированы из таблицы Microsoft Excel (*табл. 3*) с помощью пакета “xlsx” командой “read.xlsx(“xlsxFile”, sheet index)”:

1. “xlsxFile” — название файла (включая путь к нему на диске).

2. sheet index — номер страницы в файле Microsoft Excel.

Далее рассчитаны коэффициенты уравнения регрессии, и математическая модель проверена на адекватность (*табл. 4, рис. 4*).

Проведен анализ данных на автокорреляцию (*табл. 5, рис. 5*).

Проведен расчет прогнозных оценок, и с помощью метода скользящей матрицы были получены новые матрицы *X* и *Y* для составления прогноза на 2017 г. (*табл. 6, рис. 6*).

Реализация расчетов на языке программирования Python

В области анализа данных и интерактивных научно-исследовательских расчетов Python неизбежно приходится сравнивать с другими языками программирования, например с R. Наличие библиотек для анализа данных в Python делает его серьезным конкурентом в решении задач манипулирования данными. Python является отличным выбором для создания приложений обработки данных с учетом его достоинств как универсального языка программирования [9, с. 14].

Расчет выполнен с помощью IDE PyCharm (Community Edition) и библиотек “pandas”, “numpy”, “scipy.stats”, “math” (*табл. 7, рис. 7*).

Данные также были импортированы из таблицы Microsoft Excel (см. *табл. 3*) с помощью библиотеки “pandas” командой “pd.read_excel(“xlsxFile”, sheet_name)”:

1. “ xlsxFile” — название файла (включая путь к нему на диске).

2. sheet_name — название страницы в файле Microsoft Excel.

Заключение

В статье проведен анализ данных, разработана адекватная регрессионная модель, на основе которой были рассчитаны прогнозные оценки. Процесс анализа данных реализован на языках программирования Python (библиотеки “pandas”, “numpy”, “scipy.stats”, “math”) и R (пакеты “plyr”, “Rcpp”, “ggplot2”, “reshape2”, “xlsx”, “xlsxjars”).

Языки программирования Python и R имеют свои сильные и слабые стороны. Часто задачи анализа данных бывает легче и быстрее решить на одном языке, чем на другом. Если Data Scientist знает и применяет несколько языков, то сможет быстрее и эффективнее справиться с разного рода задачами, встающими перед специалистами по научному анализу данных. Также необходимо отметить, что специалист, знающий несколько языков программирования, несомненно, будет наиболее конкурентоспособен на рынке труда в сфере IT.

Таблица 6

Прогнозирование и применение метода скользящей матрицы в R

```
# новые данные независимых
переменных для прогнозирования
XP <- t(matrix(c(1, 14, 80.5)))
YP <- B[1] + B[2] * XP1 + B[3] * XP2
# скользящая матрица
X <- matrix(c(maindata[2:14, 1:3]))
Y <- matrix(c(maindata[2:14, 4]))
```

```
> XP
     [,1] [,2] [,3]
[1,]    1   14 80.5
> YP
[1] 6.262815
> Ymax
[1] 6.681457
> Ymin
[1] 5.844172
> X                          > Y
      [,1] [,2] [,3]              [,1]
 [1,]    1    2 14.2          [1,]  0.3
           ...                        ...
[13,]    1   14 80.5          [13,]  6.3
```

Рис. 6. **Результат выполнения кода из *табл. 6***

Таблица 7

Расчеты на языке программирования Python

```
mainData = pandas.read_excel(«D:\
data.xlsx», sheet_name = «Лист1»)
X = numpy.asmatrix(mainData[0:13, 0:2])
Y = numpy.asmatrix(mainData[0:13, 3])
B = (numpy.transpose(X) * X)**(–
1) * numpy.transpose(X) * Y
YR = X * B
YSR = sum(Y) / N
DY = (sum((Y - YSR)**2)) / (N - 1)
Dad = (sum((Y - YR)**2)) / (N - k)
FR = DY / Dad
F = scipy.stats.f.ppf(0.95, N — 1, N — k)
# анализ данных на автокорреляцию
e = Y — YR
e0 = numpy.asmatrix(e[0:12]);
e1 = numpy.asmatrix(e[1:13])
d = (sum((e0 — e1)**2)) / sum(e**2)
# прогнозирование
XP = numpy.matrix([1, 14, 80.5])
YP = B[0] + B[1] * XP1 + B[2] * XP2
# скользящая матрица
X = numpy.asmatrix(mainData[1:14, 0:3])
Y = numpy.asmatrix(mainData[1:14, 3])
```

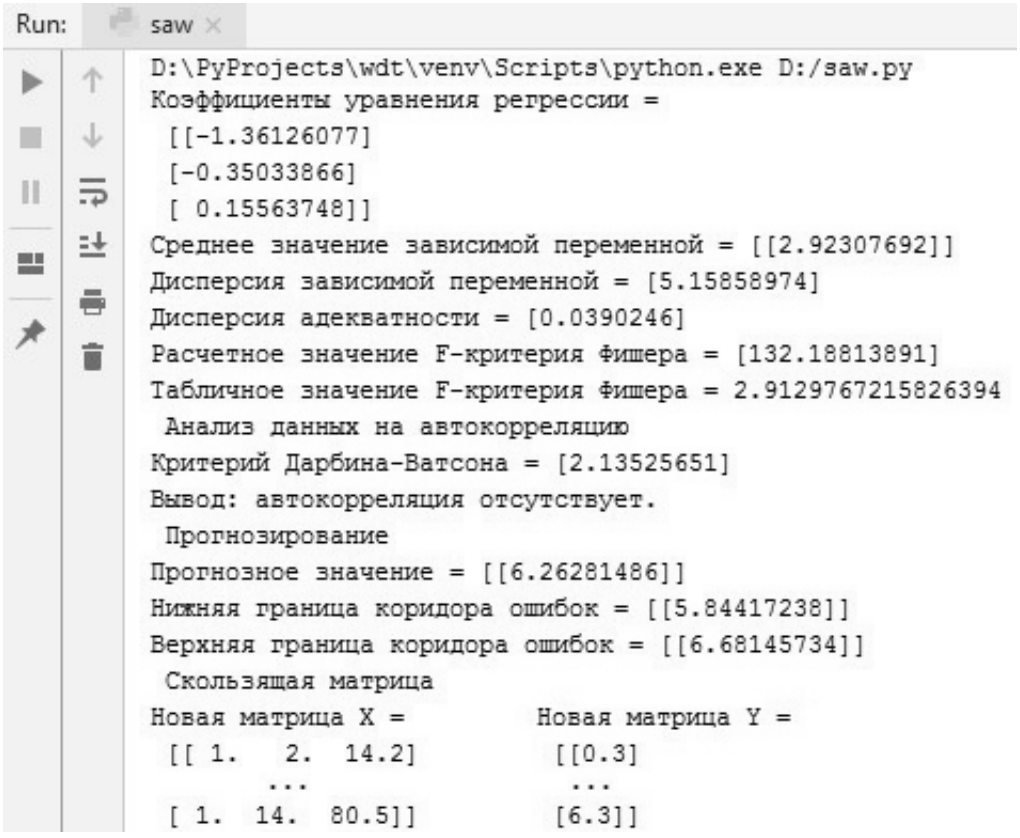

Рис. 7. **Результат расчета на Python**

Список источников

1. Кремер Н.Ш. Теория вероятностей и математическая статистика. М.: ЮНИТИ-ДАНА; 2004. 573 с.
2. Bolt G.J. Marketing and sales forecasting: a total approach. London; 1988. 347 p.
3. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело; 2004. 576 с.
4. Доугерти К. Введение в эконометрику. М.: ИНФРА-М; 2010. XIV, 465 с.
5. Зарубин В.С. Математическое моделирование в технике. М.: Изд-во МГТУ им. Н.Э. Баумана; 2003. 496 с.
6. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование. М.: Финансы и статистика; ИНФРА-М; 2010. 320 с.
7. Алкацев М.И., Дзгоев А.Э., Бетрозов М.С. Исследование и разработка метода прогнозирования потребления электроэнергии в системе управления электроснабжения региона. *Известия вузов. Проблемы энергетики*. 2012;(5–6):30–37.
8. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R. М.: ДМК Пресс; 2014. 588 с.
9. Маккинли У. Python и анализ данных. М.: ДМК Пресс; 2015. 482 с.

References

1. Kremer N. Sh. Probability theory and mathematical statistics. Moscow: UNITY-DANA Publ.; 2004. 573 p. (In Russ.).
2. Bolt G.J. Marketing and sales forecasting: a total approach. London; 1988. 347 p.
3. Magnus Ya.R., Katyshev P.K., Peresetsky A.A. Econometrics. Moscow: Delo; 2004. 576 p. (In Russ.).
4. Dougherty Ch. Introduction to Econometrics. Moscow: INFRA-M; 2010. 576 p. (In Russ.).
5. Zarubin V.S. Mathematical Modeling in Engineering. Moscow: Bauman MSTU Publ. House; 2003. 496 p. (in Russ.).
6. Afanasiev V.N., Yuzbashev M.M. Analysis of temporary ranks and forecasting. Moscow: Finansy i statistika, INFRA-M; 2010. 320 p. (In Russ.).
7. Alkatsev M.I., Dzgoev A.E., Betrozov M.S. Research and development of electric energy consumption prediction method in the electric supply management system in the region. *Proceedings of the higher educational institutions. Energy sector problems*. 2012;(5–6):30–37. (In Russ.).
8. Kabacoff R.I. R in Action. Data analysis and graphics with R. Moscow: DMK press; 2014. 588 p.
9. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. Moscow: DMK press; 2015. 482 p.