

Министерство науки и высшего образования РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Пензенский государственный университет»

Факультет вычислительной техники
Кафедра математики и суперкомпьютерного моделирования

ОТЧЕТ

по производственной (преддипломной) практике
«Проверка статистической гипотезы о значимости уравнения модели парной
линейной регрессии»

Выполнил: студент гр. 19ВМ1

Яковлева А. Д.

Руководитель: Мартынов И. В.

Пенза 2023

Министерство науки и высшего образования РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Пензенский государственный университет»

«УТВЕРЖДАЮ»

Зав. кафедрой МСМ, д.ф.-м.н.

_____ Ю.Г. Смирнов

ЗАДАНИЕ

на производственную (преддипломную) практику

(8 семестр)

Вариант 6

Тема «Проверка статистической гипотезы о значимости уравнения модели
парной линейной регрессии»

1. Изучение теоретических сведений;
2. Поиск исходных данных для проведения исследования;
3. Статистический анализ данных;
4. Получение и анализ численных результатов.

Руководитель работы

Мартынов И. В.

Содержание

Введение	4
Глава I. Теоретическая часть	6
§1. Парная регрессия и корреляция.....	6
Глава II. Практическая часть	12
§1. Результаты исследования «Gas Turbine CO and NOx Emission Data Set Data Set / Набор данных о выбросах газовых турбин CO и NOx»	12
§2. Анализ исходных данных.....	14
§3. Построение парной регрессионной модели зависимости выбросов оксидов азота от температуры окружающей среды	18
Заключение	24
Список используемых источников.....	25

Введение

В процессе проведения различных научных исследований возникает вопрос определения причинно-следственных связей между отдельными элементами и возможность возникновения закономерностей таких связей.

В процессе своей деятельности человек оказывает на окружающую среду преимущественно негативное воздействие, которое оборачивается серьезными экологическими проблемами. В ходе изучения влияния деятельности человека на окружающую среду возникло понятие «мониторинг».

Экологический мониторинг или мониторинг окружающей среды – комплексная система наблюдений за состоянием окружающей среды, оценки и прогноза изменений состояния окружающей среды под воздействием природных и антропогенных факторов.

Существуют различные методы мониторинга, они проводятся по тому или иному явлению, процессу, а также в зависимости от среды, в которой проходят исследования.

При обработке экспериментальных данных экологические исследования зависимостей случайных величин приводят к моделям регрессии и регрессионному анализу. Выбор нужной модели основывается на статистических доводах. Статистические исследования направлены на оценивание параметров регрессии; проверку гипотез о статистической значимости этих параметров; проверку адекватности построенной модели. Для оценки параметров регрессии часто используют метод наименьших квадратов. Регрессионный анализ тесно связан с корреляционным и является его логическим продолжением, углубляя представления о корреляционной связи. Регрессионный анализ используется, во-первых, для описания зависимости между переменными и определения причинной связи, во-вторых, для построения прогнозных значений зависимой переменной. Мерой зависимости является величина коэффициента корреляции. Регрессионный анализ – набор статистических методов исследования влияния одной или нескольких независимых переменных на зависимую.

Одним из наиболее распространенных методов регрессионного анализа является линейная регрессия. Линейная регрессия – выраженная в виде прямой зависимость среднего значения какой-либо величины от некоторой другой. Если установленная зависимость может быть записана в виде уравнения прямой: $y = a + b \cdot x$, то эта регрессионная зависимость называется линейной регрессией.

Несмотря на ценность применения методов статистики в экологических исследованиях, существуют некоторые нюансы, на которые стоит обратить внимание.

Первая из них – это механическое использование количественных методов анализа в исследованиях, без понимания их сути и применения к тем или иным биологическим явлениям и экологическим процессам. Важно учитывать особенности и условия использования тех или иных статистических процедур, поскольку любой метод имеет свои ограничения.

Если не учитывать эти ограничения, применение соответствующего метода становится математически неправомерно, что приводит к фальсификации результатов и выводов, к отклонению проверяемой гипотезы там, где на самом деле ее нужно было бы принять, к установлению влияния фактора, который в реальности не влияет, к подтверждению не существующих связей между элементами системы.

Вторая опасность связана с широко распространенным мнением о том, что математическая обработка данных может если не полностью учесть, то свести к минимуму те технические, организационные и методические ошибки, которые возникли при проведении исследования. На это часто надеются недобросовестные исследователи. Данное мнение глубоко ошибочно, статистические методы можно с равным успехом применять как к верным данным, так и к неправильно полученным [1, 5].

Глава I. Теоретическая часть

§1. Парная регрессия и корреляция

Регрессионный анализ исследует и оценивает связь между зависимой или объясняемой переменной и независимыми или объясняющими переменными. Зависимую переменную называют также результирующей, эндогенной переменной, результативным признаком, а независимую переменную – предсказывающей, предикторной, экзогенной переменной, факторным признаком.

В зависимости от количества факторов, включенных в уравнение регрессии, различают простую (парную) и множественную регрессии.

Парная регрессия – уравнение связи двух переменных y и x :

$$y = \hat{f}(x), \quad (1.1)$$

где y – зависимая переменная (результативный признак), а x – независимая, объясняющая переменная (признак-фактор).

Множественная регрессия, в свою очередь, представляет собой регрессию результативного признака с двумя и более факторами, т.е. модель вида:

$$y = \hat{f}(x_1, x_2, \dots, x_k). \quad (1.2)$$

Различают линейные и нелинейные регрессии.

Линейная регрессия: $y = a + b \cdot x + \varepsilon$.

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам.

Регрессии, нелинейные по объясняющим переменным:

- 1) полиномы разных степеней $y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + \varepsilon$;
- 2) равносторонняя гипербола $y = a + \frac{b}{x} + \varepsilon$.

Регрессии, нелинейные по оцениваемым параметрам:

- 1) степенная $y = a \cdot x^b \cdot \varepsilon$;
- 2) показательная $y = a \cdot b^x \cdot \varepsilon$;
- 3) экспоненциальная $y = e^{a+b \cdot x} \cdot \varepsilon$.

Построение уравнения регрессии сводится к оценке ее параметров a и b . Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических минимальна:

$$\sum_{i=1}^n (y_i - \hat{y}_{x_i})^2 = \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min. \quad (1.3)$$

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b :

$$\begin{cases} n \cdot a + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum y \cdot x. \end{cases} \quad (1.4)$$

Решая систему уравнений (1.4) либо методом последовательного исключения переменных, либо методом определителей, получим искомые оценки параметров a и b , для нахождения которых можно воспользоваться следующими формулами, которые следуют непосредственно из решения системы (1.4):

$$a = \bar{y} - b \cdot \bar{x}, \quad (1.5)$$

$$b = \frac{cov(x, y)}{\sigma_x^2}, \quad (1.6)$$

где $cov(x, y) = \overline{yx} - \bar{y} \cdot \bar{x}$ – ковариация признаков x и y , $\sigma_x^2 = \overline{x^2} - \bar{x}^2$ – дисперсия признака x .

Параметр b называется коэффициентом регрессии, величина которого показывает, на сколько единиц в среднем изменяется результат при увеличении фактора на одну единицу. Знак при коэффициенте регрессии показывает направление зависимости: при $b > 0$ – прямая зависимость, $b < 0$ – обратная зависимость.

Уравнение регрессии всегда дополняется показателем тесноты связи. Для линейной регрессии тесноту связи изучаемых явлений оценивает линейный коэффициент корреляции Пирсона r_{xy} :

$$r_{xy} = b \frac{\sigma_x}{\sigma_y} = \frac{cov(x, y)}{\sigma_x \sigma_y} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sigma_x \sigma_y}. \quad (1.7)$$

Как известно, линейный коэффициент корреляции находится в границах: $-1 \leq r_{xy} \leq 1$.

Если коэффициент регрессии $b > 0$, то $0 \leq r_{xy} \leq 1$; при $b < 0$, $-1 \leq r_{xy} \leq 0$.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических:

$$\bar{A} = \frac{1}{n} \cdot \sum \left| \frac{y - \hat{y}_x}{y} \right| \cdot 100\%. \quad (1.8)$$

Построенное уравнение регрессии считается удовлетворительным, если значение $\bar{A}$ не превышает 8-10%.

В задачах статистики, связанных с выявлением связей между объектами вычисляется коэффициент эластичности:

$$\varepsilon = b \cdot \frac{\bar{x}}{\bar{y}}. \quad (1.9)$$

Коэффициент эластичности показывает на сколько процентов изменится зависимая переменная y при изменении фактора x на 1%.

Долю дисперсии результативного признака y , объясняемую регрессией, в общей дисперсии результативного признака характеризует коэффициент детерминации:

$$r_{xy}^2 = \frac{\sigma_{y_{\text{объясн}}}^2}{\sigma_{y_{\text{общ}}}^2}. \quad (1.10)$$

После того как уравнение линейной регрессии найдено, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли аналитическая модель, выражающая зависимость между

переменными, экспериментальным данным, и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой.

Оценка значимости уравнения регрессии в целом осуществляется с помощью F -критерия Фишера.

В начале исследования выдвигается нулевая гипотеза H_0 : коэффициент регрессии $b = 0$, и, следовательно, фактор x не влияет на результат y .

F -критерий Фишера заключается в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи. Для этого выполняется сравнение фактического $F_{\text{факт}}$ и критического (табличного) $F_{\text{табл}}$ значений F -критерия Фишера.

$F_{\text{факт}}$ определяется из соотношения значений факторной и остаточной дисперсий, рассчитанных на одну степень свободы:

$$F_{\text{факт}} = \frac{D_{\text{факт}}}{D_{\text{ост}}} = \frac{\frac{\sum(\hat{y}_x - \bar{y})^2}{1}}{\frac{\sum(y - \hat{y}_x)^2}{n - 2}} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2). \quad (1.11)$$

$F_{\text{табл}}$ – максимально возможное значение критерия под влиянием случайных факторов при степенях свободы $k_1 = 1$, $k_2 = n - 2$ и уровне значимости α .

Уровень значимости α – вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно величина α принимается равной 0,05.

Если нулевая гипотеза справедлива, то факторная и остаточная дисперсии равны ($D_{\text{факт}} = D_{\text{ост}}$). Сопоставляя фактическое и критическое (табличное) значения F -статистики $F_{\text{факт}}$ и $F_{\text{табл}}$ принимают или отвергают H_0 . Вычисленное значение F -отношения считается достоверным (отличным от единицы), если оно больше табличного, $|F_{\text{факт}}| > F_{\text{табл}}$, и, следовательно, нулевая гипотеза об отсутствии связи признаков отклоняется, и можно сделать вывод о существенности этой связи. В ином случае, если величина будет меньше табличной, т.е. $|F_{\text{факт}}| < F_{\text{табл}}$, то вероятность H_0 выше заданного уровня значимости

α , и нулевая гипотеза не отклоняется, и признается статистическая незначимость (ненадежность) уравнения регрессии.

В линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров.

Для проверки существенности коэффициента регрессии применяется t -распределение Стьюдента.

В начале исследования выдвигается нулевая гипотеза H_0 о случайной природе показателей, т.е. о их незначимом отличии от нуля.

Для того, чтобы оценить значимость коэффициента регрессии, необходимо определить его стандартную ошибку m_b :

$$m_b = \sqrt{\frac{\sum(y - \hat{y}_x)^2}{(n - 2) \cdot \sum(x - \bar{x})^2}} = \sqrt{\frac{S_{\text{ост}}^2}{\sum(x - \bar{x})^2}} = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}}, \quad (1.12)$$

где $S_{\text{ост}}^2 = \frac{\sum(y - \hat{y}_x)^2}{n - 2}$ – остаточная дисперсия на одну степень свободы.

Для проверки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, т.е. определяется фактическое значение t -критерия Стьюдента: $t_b = \frac{b}{m_b}$. Значение t_b необходимо сравнить с табличным значением при $df = n - 2$ степенях свободы.

Для расчета доверительного интервала определяется предельная ошибка для коэффициента регрессии: $\Delta_b = t_{\text{табл}} \cdot m_b$

Доверительный интервал определяется по формулам: $\gamma_b = b \pm \Delta_b$, $\gamma_{b_{\min}} = b - \Delta_b$, $\gamma_{b_{\max}} = b + \Delta_b$.

Знак коэффициента регрессии указывает:

- 1) на рост результативного признака y при увеличении фактора x , если $b > 0$;
- 2) на падение результативного признака y при увеличении независимой переменной x , если $b < 0$;
- 3) на независимость результативного признака от фактора, если $b = 0$.

Если нижняя граница отрицательна ($\gamma_{b_{min}} < 0$), а верхняя положительна ($\gamma_{b_{max}} > 0$), т.е. в границы доверительного интервала попадает ноль (например, $-1,3 \leq b \leq 0,7$), то оцениваемый принимается нулевым, так как он не может одновременно принимать положительное и отрицательное значения.

Стандартная ошибка параметра a определяется по формуле:

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n - 2} \cdot \frac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}} = \sqrt{S_{\text{ост}}^2 \cdot \frac{\sum x^2}{n^2 \cdot \sigma_x^2}} = S_{\text{ост}} \frac{\sqrt{\sum x^2}}{n \cdot \sigma_x}. \quad (1.13)$$

Процедура оценивания значимости данного параметра не отличается от рассмотренной ранее для коэффициента регрессии; t -критерий вычисляется по формуле: $t_a = \frac{a}{m_a}$, его величина сравнивается с табличным значением при $df = n - 2$ степенях свободы.

Существенность линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции $m_{r_{xy}}$:

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n - 2}}. \quad (1.14)$$

Фактическое значение t -критерия Стьюдента определяется как:

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n - 2}. \quad (1.15)$$

Значение $t_{r_{xy}}$ впоследствии сравнивается с табличным значением при необходимом уровне значимости α и числе степеней свободы $(n - 2)$.

Сравнивая фактическое и табличное значения t -статистики $t_{\text{факт}}$ и $t_{\text{табл}}$, принимают или отвергают гипотезу H_0 . Если $t_{\text{факт}} > t_{\text{табл}}$, то H_0 отклоняется, т.е. a, b и r_{xy} значимо отличаются от нуля и сформировались под влиянием фактора x . Если $t_{\text{факт}} < t_{\text{табл}}$, то H_0 не отклоняется и признается случайная природа показателей: a, b и r_{xy} [2, 3].

Глава II. Практическая часть

§1. Результаты исследования «Gas Turbine CO and NO_x Emission Data Set Data Set / Набор данных о выбросах газовых турбин CO и NO_x»

Рассмотрим открытый набор данных, который содержит 36733 экземпляра измерений 11 датчиков, агрегированных за один час с газовой турбины, расположенной в северо-западном регионе Турции, с целью изучения выбросов дымовых газов, а именно монооксида углерода CO и оксидов азота NO_x (монооксид азота NO + диоксид азота NO₂). Собранные данные содержат среднечасовые измерения девяти входных параметров, из которых можно выделить переменные, характеризующие состояние окружающей среды: температура, давление, влажность; переменные, характеризующие технологический процесс: перепад давления в воздушном фильтре, давление выхлопа газовой турбины, температура на входе в турбину, температура после турбины, давление нагнетания компрессора, выход энергии турбины; и два целевых показателя – объемы выбросов CO и NO_x [4, 7].

Переменная	Сокраще- ние	Единица измере- ния	Мин. значе- ние	Макс. значе- ние	Сред- нее зна- чение
Температура окружающей среды	AT	°C	-6,23	37,1	17,71
Давление окружающей среды	AP	мбар	985,85	1036,6	1013,07
Влажность окружающей среды	AH	%	24,09	100,2	77,87
Перепад давления в воздушном фильтре	AFDP	мбар	2,09	7,61	3,93
Давление выхлопа газовой турбины	GTEP	мбар	17,7	40,72	25,56
Температура на входе в турбину	TIT	°C	1000,8	1100,9	1081,43
Температура после турбины	TAT	°C	511,04	550,61	546,16
Давление нагнетания компрессора	CDP	мбар	9,85	15,16	12,06
Выход энергии турбины	TEY	МВтч	100,02	179,5	133,51
Монооксид углерода	CO	мг/м ³	0,00039	44,1	2,37
Оксиды азота	NO _x	мг/м ³	25,91	119,91	65,29

Таблица 1. Основная статистическая информация о данных, использованных в исследовании.

Подробно контролируемые переменные приведены в таблице 1.

Часть данных экологического исследования представлена в таблице 2.

АТ	АР	АН	АФДР	ГТЕР	ТІТ	ТАТ	СДР	ТЕУ	СО	NOx
4,59	1018,70	83,68	3,58	23,98	1086,20	549,83	11,90	134,67	0,33	81,95
4,29	1018,30	84,24	3,57	23,95	1086,10	550,05	11,89	134,67	0,45	82,38
3,90	1018,40	84,86	3,58	23,99	1086,50	550,19	12,04	135,10	0,45	83,78
3,74	1018,30	85,43	3,58	23,91	1086,50	550,17	11,99	135,03	0,23	82,51
3,75	1017,80	85,18	3,58	23,92	1085,90	550	11,91	134,67	0,27	82,03
3,89	1017,70	83,95	3,58	23,90	1086	549,98	11,87	134,67	0,23	81,75
3,67	1018	84,11	3,58	23,89	1085,90	550,04	11,88	134,68	0,44	84,59
3,59	1018,20	83,87	3,58	23,88	1086	549,88	11,89	134,66	0,80	84,19
3,71	1018,50	84,95	3,60	23,96	1086,30	549,98	11,87	134,65	0,69	83,98
4,83	1018,50	85,35	3,52	23,42	1083,10	549,80	11,69	132,67	1,03	82,65
6,03	1018,40	86,05	3,57	23,99	1085,80	550,04	11,80	135,24	1,02	81,95
7,99	1017,90	87,54	4,39	30,13	1094,50	534,81	13,33	152,60	0,94	73,29
9,87	1017,20	84,25	4,70	32,34	1100,20	530,58	13,93	159,99	1,04	70,56
10,72	1016,50	80,26	4,75	32,74	1100,20	528,93	14,05	160,60	1,35	71,12
11,02	1016,30	78,13	4,74	32,67	1100,10	529,06	14,02	160,40	1,07	70,11
10,95	1016,20	77,96	4,69	32,42	1100,30	530,67	13,95	160,09	0,22	67,42
10,76	1015,90	79,27	4,71	32,63	1100,20	529,35	14,03	160,66	0,24	67,49
10,18	1015,80	81,42	4,64	32,29	1099,80	530,67	13,83	160,64	0,63	66,86
9,51	1015,70	82,52	4,62	32,22	1099,80	530,75	13,91	160,84	0,57	67,36
5,39	1014,30	90,46	3,63	24,06	1084,80	549,11	11,81	134,26	0,52	82,14
5,49	1013,90	88,63	3,28	21,25	1069,30	549,65	11,13	121,95	1,17	84,91
5,45	1013,60	87,77	3,63	23,99	1085,10	549,71	11,79	133,51	0,47	82,95
5,48	1013,40	87,77	3,51	23,01	1079,70	550,13	11,61	129,19	0,68	82,62
5,38	1013,10	86,56	3,56	23,42	1082,20	550,10	11,67	131,08	0,68	84,05
5,22	1012,80	87,09	3,51	22,94	1079,40	550,09	11,69	129,06	0,90	85,06
5,25	1012,70	87,37	3,62	23,80	1084,40	549,93	11,69	132,86	0,66	82,95
5,27	1012,50	88,53	3,42	22,21	1074,90	550,06	11,45	125,76	1,05	83,45
5,98	1012,50	90,68	3,60	23,45	1082,30	550,02	11,62	130,63	0,77	82,30
6,77	1012,80	90,88	3,70	24,21	1086,30	550,04	11,78	133,52	0,68	82,36
7,57	1013	90,30	3,71	24,20	1085,80	549,87	11,80	132,98	0,71	80,61
9,27	1013,20	91,48	3,80	24,63	1088,50	549,87	11,92	133,67	0,84	79,25
11,35	1013,10	88,32	3,70	24,38	1086,60	549,85	11,88	133,68	0,95	74,72
12,52	1012,60	79,37	3,68	24,36	1085,90	549,71	11,90	133,67	1,33	78,37
13,10	1011,90	74,68	3,71	24,39	1086,20	549,63	11,91	133,67	1,74	78,34
13,46	1011,70	72,25	3,73	24,38	1086,60	549,83	11,86	133,64	1,85	77,35
13,43	1011,50	71,49	3,40	22,02	1072,50	550,12	11,22	123,68	1,15	77,84
12,86	1011,70	76,20	3,41	22,43	1074,50	549,70	11,32	125,72	1,34	77,93
12,26	1011,80	79,06	4,66	32,29	1099,60	531,41	13,73	158,68	0,39	66,90
...	...	...	...	...	...	...	...	...	...	...

Таблица 2. Таблица исходных данных.

§2. Анализ исходных данных

Проведем первоначальный статистический анализ данных. В таблицах 3 и 4 представлена линейная зависимость между двумя случайными величинами x и y , изученная с помощью статистического пакета SPSS.

	AT	AP	АН	AFDP	GTEP	TIT	TAT	CDP	TEY
AT	1	-0,407	-0,476	0,252	0,046	0,184	0,282	0,015	-0,091
AP	-0,407	1	-0,015	-0,04	0,058	-0,005	-0,226	0,103	0,118
АН	-0,476	-0,015	1	-0,148	-0,235	-0,222	0,023	-0,196	-0,137
AFDP	0,252	-0,04	-0,148	1	0,678	0,691	-0,467	0,703	0,665
GTEP	0,046	0,058	-0,235	0,678	1	0,874	-0,7	0,978	0,964
TIT	0,184	-0,005	-0,222	0,691	0,874	1	-0,381	0,908	0,91
TAT	0,282	-0,226	0,023	-0,467	-0,7	-0,381	1	-0,706	-0,682
CDP	0,015	0,103	-0,196	0,703	0,978	0,908	-0,706	1	0,989
TEY	-0,091	0,118	-0,137	0,665	0,964	0,91	-0,682	0,989	1

Таблица 3. Корреляционная матрица входных переменных.

Опираясь на данные таблицы 3, можно утверждать, что существование очень сильной линейной зависимости между входными данными возможно. В частности, между давлением нагнетания компрессора (CDP) и выходом энергии турбины (TEY) коэффициент корреляции Пирсона $r_{xy} = 0,989$, между CDP и давлением выхлопа газовой турбины (GTEP) $r_{xy} = 0,978$. GTEP также имеет очень сильную корреляцию с TEY, $r_{xy} = 0,964$. Основываясь на данные таблицы 3, можно сделать вывод: вариация параметра TEY практически не зависит от изменений переменных окружающей среды.

	CO	NOx
AT	-0,174	-0,558
AP	0,067	0,192
АН	0,107	0,165
AFDP	-0,448	-0,188
GTEP	-0,519	-0,202
TIT	-0,706	-0,214
TAT	0,058	-0,093
CDP	-0,551	-0,171
TEY	-0,57	-0,116

Таблица 4. Попарная корреляция между признаками и двумя целевыми переменными CO и NOx.

Анализирую таблицу 4, наблюдаем более сильную корреляцию между СО и параметрами турбины, которые характеризуют технологический процесс. Также видим, самое большое влияние на СО оказывает температура на входе в турбину (TIT), корреляция Пирсона $r_{xy} = -0,706$. И это неудивительно, ведь известно, что при неполном сгорании с более низкой температурой на входе в турбину происходит больший объем выбросов СО. Что касается NOx, самую сильную корреляцию наблюдаем с температурой окружающей среды (AT), $r_{xy} = -0,558$, что свидетельствует о том, что для снижения выбросов выхлопных газов целесообразно работать при более высоких температурах.

Особенностью данного набора данных является достаточно длительный срок сбора данных – 5 лет (с 2011 по 2015 гг.). На протяжении этого времени неоднократно проводилось профилактическое обслуживание или ремонт оборудования, могли существенно измениться технологические режимы работы турбины. Поэтому был поставлен вопрос о проведении анализа данных с тем, чтобы установить наличие существенных изменений статистических характеристик каждого показателя.

На рис. 4 приведены три графика boxplot для выхода энергии турбины и выбросов СО и NOx по годам. Средние горизонтальные линии соответствуют медианным значениям, граничные показывают нижний (25%) и верхний (75%) квартили. На графиках рис. 4 черными маркерами отмечены резко выделяющиеся, как в большую, так и в меньшую сторону, значения наблюдений. Выбросы монооксида углерода в ряде случаев могут увеличиваться в десятки раз по сравнению с медианным значением. Объем выбросов оксидов азота больше по абсолютной величине, максимальное увеличение NOx превышает медианное значение в 1.5–2 раза.

Был проведен анализ соответствующих технологических параметров, чтобы исключить возможные ошибки в измерении или фиксации значений переменных, так как в исследуемой задаче наибольшую важность имеет правильное предсказание именно резкого роста выбросов СО и NOx. Анализ

показал, что резкий рост выбросов связан с изменением параметров технологического процесса. На рис. 4 также видно, что в 2014-2015 гг. произошло снижение выбросов оксидов азота, причем это снижение не может быть объяснено уменьшением выработки энергии.

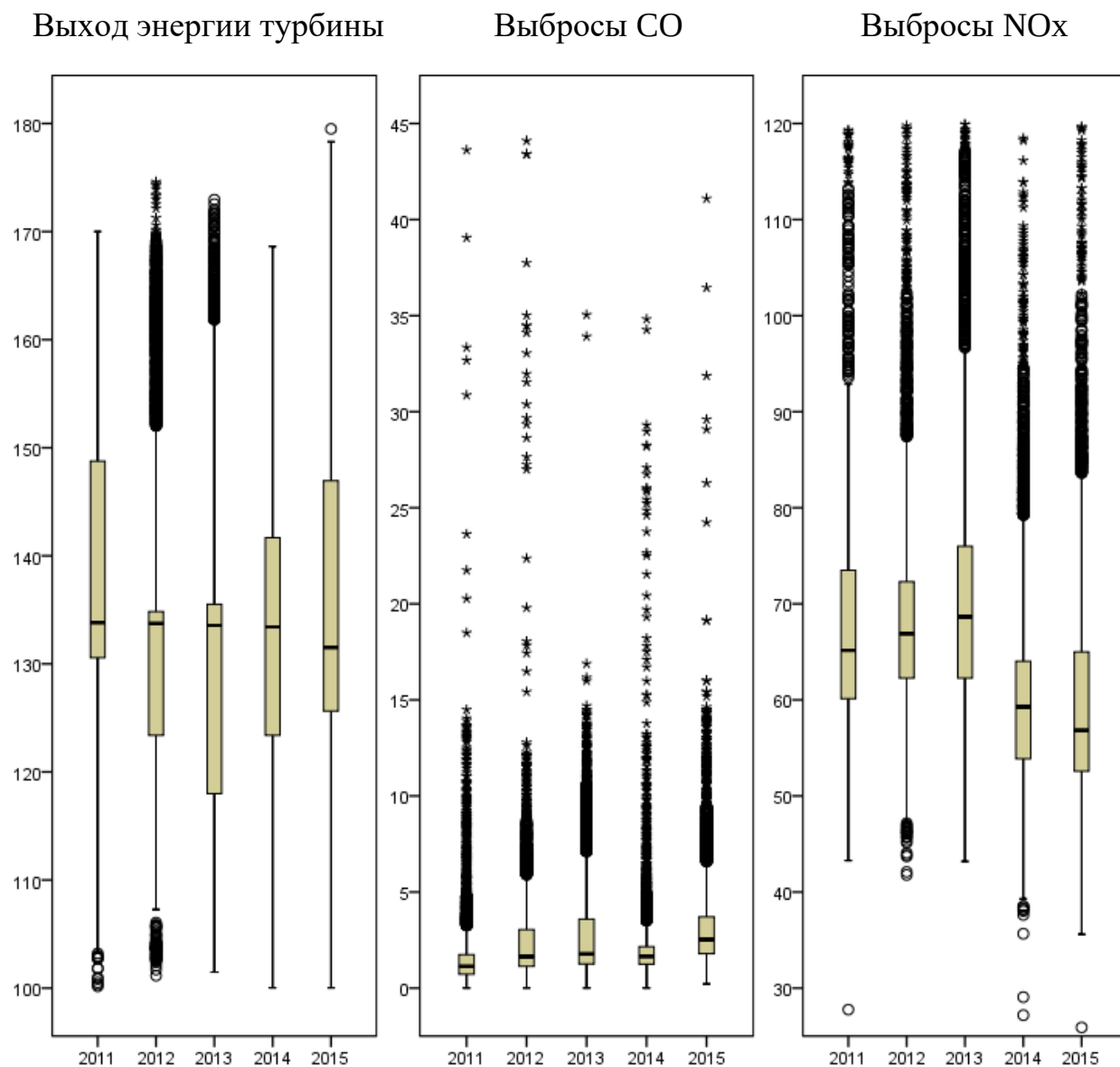


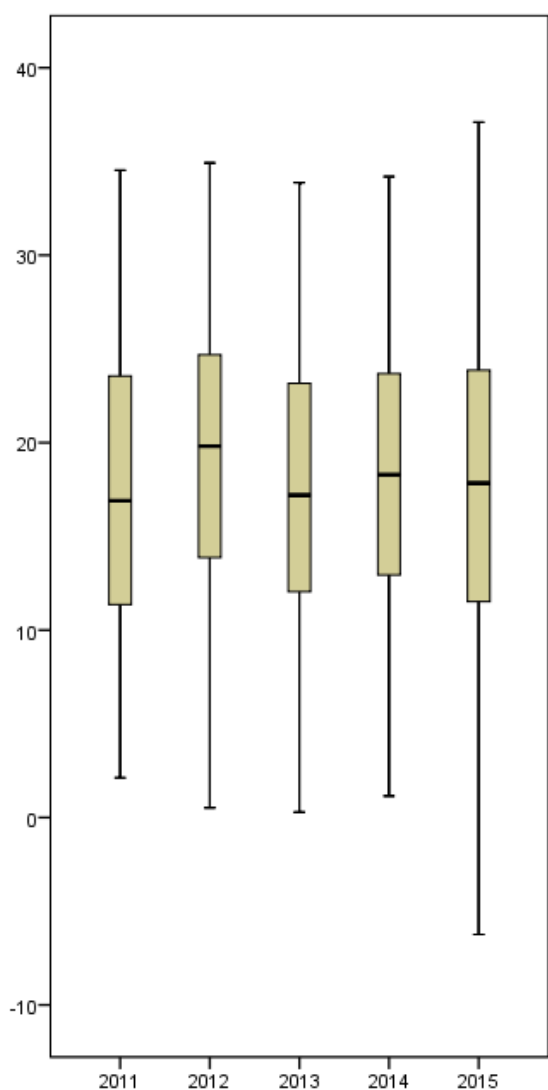
Рис. 4. Графики boxplot для выхода энергии турбины, выбросов монооксида углерода и оксидов азота.

На рис. 5 приведены 2 графика boxplot для АТ и выбросов NOx по годам. Проанализировав рис. 5, можно сделать вывод: модель прогнозирования выбросов оксидов азота претерпела качественное изменение в 2014-2015 гг.

Так как качественная зависимость выбросов оксидов азота от температуры окружающей среды изменилась, для анализа моделей прогноза выбросов

NO_x следует исследовать результаты, тестируя модели отдельно на данных 2011–2013 гг. и на данных 2014–2015 гг. [4, 6].

Температура окружающей среды



Выбросы NO_x

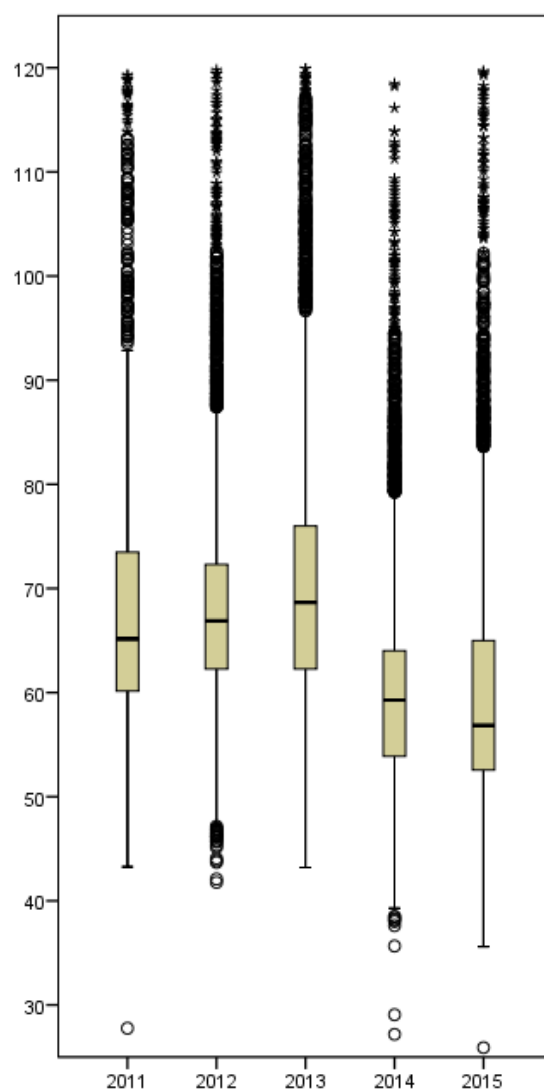


Рис. 5. Графики boxplot для температуры окружающей среды и выбросов оксидов азота.

§3. Построение парной регрессионной модели зависимости выбросов оксидов азота от температуры окружающей среды

Рассмотрим подробнее более актуальную (за 2015 г.) регрессионную модель прогноза выбросов оксидов азота.

На основе качественного анализа определили независимую и зависимую переменные (фактор (x) и результат (y)): АТ (температура окружающей среды) – независимая переменная, а NO_x (выбросы оксидов азота) – зависимая переменная

Для расчета параметров a и b линейной регрессии $y = a + b \cdot x$ решили систему нормальных уравнений относительно a и b :

$$\begin{cases} n \cdot a + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum y \cdot x. \end{cases}$$

По исходным данным рассчитали $\sum y$, $\sum x$, $\sum y \cdot x$, $\sum x^2$, $\sum y^2$ и составили следующую таблицу 5.

N	x	y	$y \cdot x$	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	$\bar{A}_i$
1	21,65	55,92	1210,78	468,81	3127,05	56,35	-0,43	0,77
2	18,51	55,84	1033,62	342,66	3117,88	59,13	-3,29	5,89
3	15,76	70,61	1112,90	248,41	4985,91	61,56	9,05	12,81
4	18,22	56,84	1035,63	332,00	3230,44	59,39	-2,55	4,48
5	21,87	60,30	1318,73	478,21	3636,57	56,16	4,14	6,87
6	16,13	70,21	1132,24	260,08	4929,16	61,24	8,97	12,78
7	25,47	53,01	1350,33	648,92	2809,85	52,97	0,04	0,07
8	24,35	50,41	1227,46	592,87	2541,27	53,96	-3,55	7,05
9	23,82	54,86	1306,62	567,34	3009,18	54,43	0,42	0,77
10	18,94	62,63	1186,31	358,76	3922,77	58,75	3,88	10,34
11	17,53	66,92	1172,97	307,27	4477,75	60,00	6,92	0,50
12	21,55	56,16	1210,20	464,40	3153,72	56,44	-0,28	3,73
13	19,87	60,17	1195,82	394,98	3620,43	57,92	2,25	0,63
...	...	...	...	...	...	...	...	...
5967	9,40	57,38	539,37	88,38	3291,89	67,19	-9,81	17,11
Итого	107337	355584	6086526	2281109	21618401	355584	0	431
Ср. знач.	17,99	59,59	1020,03	382,29	3622,99			7,22
σ	7,66	8,47						
σ^2	58,70	71,82						

Таблица 5. Работа с исходными данными.

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sigma_x^2} = \frac{1020,03 - 59,59 \cdot 17,99}{58,70} = -0,88,$$

$$a = \bar{y} - b \cdot \bar{x} = 59,59 + 0,88 \cdot 17,99 = 75,51.$$

Уравнение регрессии: $y = 75,51 - 0,88x$. С увеличением температуры окружающей среды на 1 °С доля выбросов оксидов азота снижается в среднем на 0,88 мг/м³.

Рассчитали линейный коэффициент парной корреляции:

$$r_{xy} = b \frac{\sigma_x}{\sigma_y} = -0,88 \cdot \frac{7,66}{8,47} = -0,80.$$

Связь очень высокая, обратная.

Также можно рассчитать коэффициент корреляции в Excel с помощью функции *KORREL*.

Определили коэффициент детерминации:

$$r_{xy}^2 = (-0,80)^2 = 0,64.$$

Вариация результата y на 64% объясняется вариацией фактора x .

На основе исходных данных для каждого значения фактора x нашли теоретические (расчетные) значения $\hat{y}_x$ (см. таблицу 5) по уравнению парной линейной регрессии с помощью функции *ТЕНДЕНЦИЯ*.

Определили величину средней ошибки аппроксимации $\bar{A}$:

$$\bar{A} = \frac{1}{n} \cdot \sum \left| \frac{y - \hat{y}_x}{y} \right| \cdot 100\% = \frac{431}{5967} = 7,22\%.$$

В среднем расчетные значения отклоняются от фактических на 7,22%. Так как $\bar{A} < 8 - 10\%$, следовательно, уравнение достаточно хорошо аппроксимирует модель.

Затем рассчитали коэффициент эластичности:

$$\varepsilon = b \cdot \frac{\bar{x}}{\bar{y}} = -0,88 \cdot \frac{17,99}{59,59} = -0,27.$$

Так как $\varepsilon = -0,27$, следовательно, увеличение показателя АТ на 1% приводит к уменьшению показателя NO_x на 0,27%.

Оценили качество уравнения регрессии в целом с помощью F -критерия Фишера, нашли фактическое значение:

$$F_{\text{факт}} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2) = \frac{0,64}{1 - 0,64} \cdot (5967 - 2) = 10604,44.$$

Значение F -критерия Фишера при уровне значимости $\alpha = 0,05$ рассчитали с помощью функции $F.ОБР.ПХ$, т.е. $F_{\text{табл}} = 3,84$.

Так как $|F_{\text{факт}}| > F_{\text{табл}}$, следовательно, нулевую гипотезу H_0 об отсутствии связи признаков отклонили, и на уровне значимости $\alpha = 0,05$ признали статистическая значимость уравнения в целом.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитали t -критерий Стьюдента и доверительные интервалы каждого из показателей.

Случайные ошибки параметров линейной регрессии и коэффициента корреляции ($S_{\text{ост}}^2 = \frac{\sum(y - \hat{y}_x)^2}{n - 2} = \frac{154375}{5967} = 25,87$):

$$m_b = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}} = \frac{\sqrt{25,87}}{7,66 \cdot \sqrt{5967}} = 0,0086;$$

$$m_a = S_{\text{ост}} \frac{\sqrt{\sum x^2}}{n \cdot \sigma_x} = \frac{\sqrt{25,87 \cdot 2281109}}{5967 \cdot 7,66} = 0,1681;$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} = \sqrt{\frac{1 - 0,64}{5967 - 2}} = 0,0777.$$

Фактические значения t -статистик:

$$t_b = \frac{b}{m_b} = \frac{-0,88}{0,0086} = -102,326;$$

$$t_a = \frac{a}{m_a} = \frac{75,51}{0,1681} = 449,197;$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{-0,80}{0,0777} = -10,296.$$

Табличное значение критерия для уровня значимости 0,05 при $n = 5967$ приняли равным $t_{\text{табл}} = 1,96$. Так как $t_b > t_{\text{табл}}$, $t_a > t_{\text{табл}}$ и $t_{r_{xy}} > t_{\text{табл}}$, то

признали статистическую значимость параметров регрессии и показателя тесноты связи. Рассчитали доверительные интервалы для параметров регрессии a и b : $\gamma_a = a \pm \Delta_a$ и $\gamma_b = b \pm \Delta_b$, где $\Delta_a = t_{\text{табл}} \cdot m_a$, $\Delta_b = t_{\text{табл}} \cdot m_b$. Получили, что $a \in [75,181; 75,869]$ и $b \in [-0,872; -0,888]$.

Также характеристики уравнения модели регрессии можно получить с помощью встроенной функции *ЛИНЕЙН*:

- 1) Оценки параметров двумерной модели регрессии – коэффициенты уравнения и их ошибки;
- 2) Значение коэффициента детерминации r_{xy}^2 ;
- 3) Расчетное значение F -критерия Фишера.

Результат функции *ЛИНЕЙН* представлен в таблице 6.

Значение коэффициента b	-0,8834	75,5058	Значение коэффициента a
Стандартная ошибка для b	0,0086	0,16806	Стандартная ошибка для a
Коэффициент детерминации r_{xy}^2	0,63975	5,08724	Стандартная ошибка для y
F -статистика	10593	5965	Число степеней свободы
Регрессионная сумма квадратов	274148	154375	Остаточная сумма квадратов

Таблица 6. Результат выполнения функции *ЛИНЕЙН*.

Определили значения доверительного интервала для $\hat{y}_x$:

$\hat{y}_x$	Δ	Нижняя граница	Верхняя граница
56,35	0,32811	56,02259565	56,6788083
59,13	0,31481	58,8146702	59,44429297
61,56	0,42088	61,14146686	61,98323584
59,39	0,32473	59,06131236	59,71076516
56,16	0,32901	55,83060081	56,48862142
61,24	0,31231	60,92624475	61,55087225
52,97	0,31391	52,65554044	53,28337024
53,96	0,31136	53,65335748	54,27608301
54,43	0,30001	54,13358888	54,73361232
58,75	0,32865	58,42041852	59,07771992
60,00	0,31313	59,68510395	60,31137057
56,44	0,30495	56,13599037	56,74588828
57,92	0,31139	57,61227556	58,23505016
50,05	0,42145	49,62768108	50,47057281
52,58	0,30359	52,2721803	52,87936526
...	...	...	...

Таблица 7. Значения доверительного интервала для $\hat{y}_x$.

Построили диаграмму результатов для первых ста значений (см. рис. 6). Диаграмма содержит график реальных значений зависимой переменной, график значений уравнения парной линейной регрессии с доверительным интервалом.

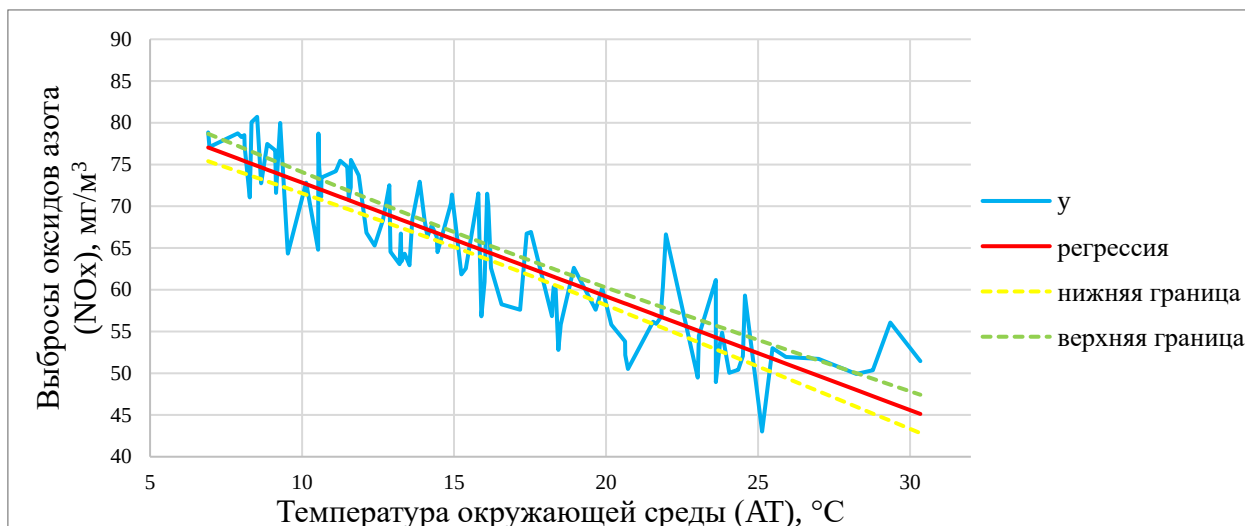


Рис. 6. Зависимость количества выбросов оксидов азота от изменения температуры окружающей среды.

Аналогичные расчеты провели в статистическом пакете SPSS.

Загрузили данные из Excel в SPSS и с помощью функции *Парные корреляции* исследовали взаимосвязь между NOx (выбросы оксидов азота) и AT (температурой окружающей среды), получили следующие данные (рис. 7)

Корреляции		AT	NOx
AT	Корреляция Пирсона	1	-,800**
	Знач. (двухсторонняя)		,000
	N	5967	5967
NOx	Корреляция Пирсона	-,800**	1
	Знач. (двухсторонняя)	,000	
	N	5967	5967

** . Корреляция значима на уровне 0,01 (двухсторонняя).

Рис. 7. Результат работы функции *Парные корреляции*.

Затем для выявления связи между NOx и AT воспользовались функцией *Диаграмма рассеяния*. Диаграмма рассеяния или поле корреляции позволяет

без математической обработки экспериментальных данных о значениях двух переменных на основе графического представления этих данных оценить характер и тесноту связи между ними.

Добавив на поле корреляции линию аппроксимации получили следующий график (см. рис. 8).

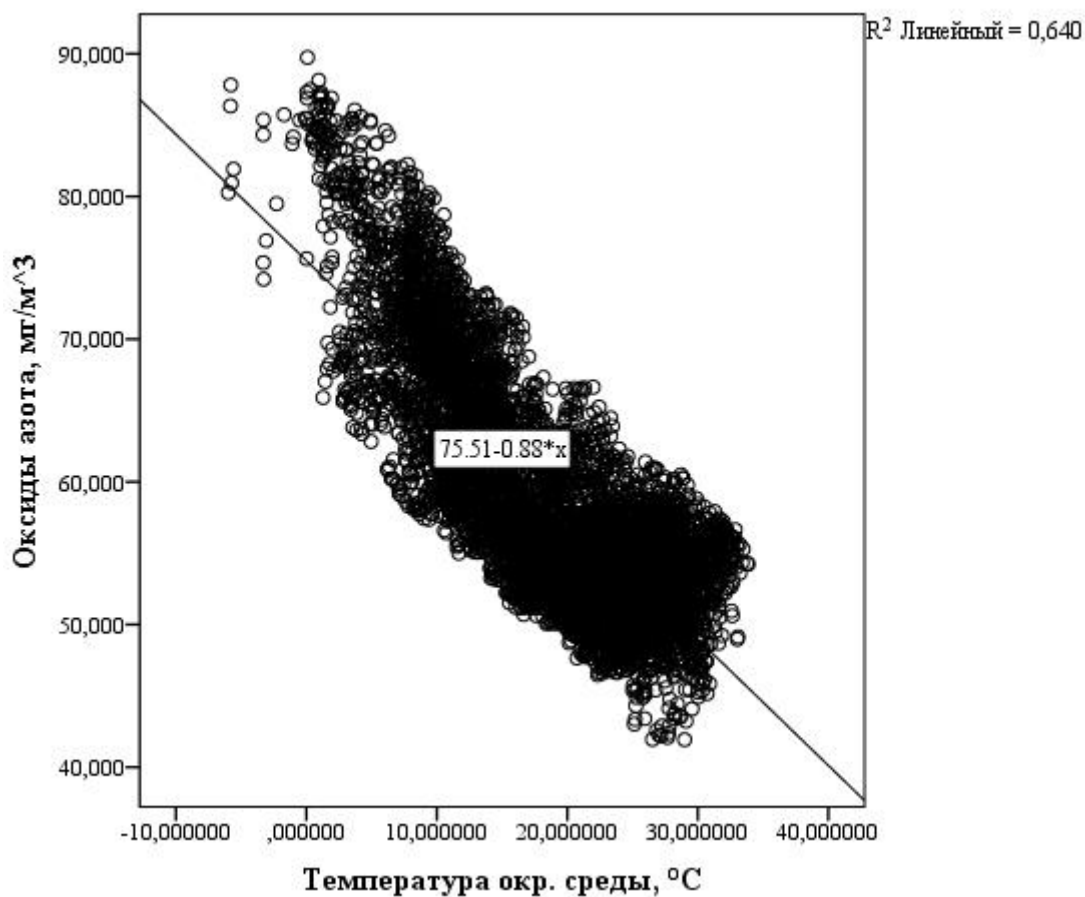


Рис. 8. Диаграмма рассеяния.

Заключение

В теоретической части были рассмотрены такие понятия, как парная и множественная регрессии, коэффициент корреляции Пирсона, средняя ошибка аппроксимации, коэффициент эластичности, коэффициент детерминации, F -критерий Фишера, t -критерий Стьюдента.

В практической части были получены характеристики уравнения регрессионной модели; была проведена оценка качества как уравнения в целом, с помощью F -критерия Фишера, так и коэффициентов регрессии и корреляции, с помощью t -критерия Стьюдента. Также была построена диаграмма, содержащая графики реальных значений зависимой переменной и значений уравнения парной линейной регрессии с доверительным интервалом.

Список литературы

1. Горшков, М. В. Экологический мониторинг. Учеб. пособие / М. В. Горшков. – Владивосток: Изд-во ТГЭУ, 2010. – 313 с.
2. Елисеева И. И. Практикум по эконометрике: учеб. пособие / И. И. Елисеева, С. В. Курышева, Н. М. Гордеенко и др.; под ред. И. И. Елисеевой. – М.: Финансы и статистика, 2005. – 192 с.
3. Елисеева, И. И. Эконометрика: учебник / И. И. Елисеева, С. В. Курышева, Т. В. Костеева и др.; под ред. И. И. Елисеевой. – 2-е изд., перераб. и доп. – М.: Финансы и статистика, 2007. – 576 с.
4. Кочуева, О. Н. Разработка моделей прогнозирования выбросов оксидов углерода и азота газовых турбин / О. Н. Кочуева // Neftegaz.RU. – 2022. – № 5-6. – С. 15-20.
5. Никифорова, Ю. Ю. Статистические методы в экологии и природопользовании: учеб. пособие / Ю. Ю. Никифорова; под общ. ред. И. С. Белюченко. – Краснодар: КубГАУ, 2019. – 88 с.
6. Predicting CO and NO_x emissions from gas turbines: novel data and a benchmark PEMS / H. Kaya, P. Tüfekci, E. Uzun // Turkish Journal of Electrical Engineering and Computer Sciences. – 2019. – Vol. 27. – No. 6. – P. 4783-4796.
7. Gas turbine CO and NO_x emission data set data set. [Электронный ресурс]. Режим доступа:
<https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set>