МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

На правах рукописи

УДК 004.415.2

Рычка Ольга Валентиновна

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ ВЫЯВЛЕНИЯ И КОРРЕКТИРОВКИ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки)

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель
кандидат технических наук, доцент
Григорьев А.В.

Идентичность всех экземпляров
ПОДТВЕРЖДАЮ
Ученый секретарь
диссертационного совета Д 01.024.04
канд. тех. наук, доцент Т.В. Завадская

Донецк 2021

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. В условиях использования большого объёма данных, а также высокой значимости результатов их анализа важной задачей является обнаружение аномальных данных (выбросов, несогласованных наблюдений, исключений). Появление аномальных измерений может быть обусловлено «человеческим фактором» или влиянием на систему внешних факторов. Игнорирование их наличия может привести к существенному ухудшению качества модели, поэтому важное значение имеет предварительная обработка исходных статистических данных, которая включает этапы выявления аномальных измерений, их корректировку или удаление.

Качественной можно считать модель, которая обладает следующими свойствами: простота, максимальное соответствие реальным данным (необходимо стремиться к максимально возможному значению коэффициента детерминации R^2), высокие прогнозные качества.

Одним из главных инструментов анализа экспериментальных данных и обнаружения закономерностей в них является регрессионный анализ. С его помощью находится математическая модель влияния одной или нескольких независимых переменных X_i на зависимую переменную Y.

Частным, но важным случаем регрессионного анализа являются парные линейные модели. В пользу использования линейных моделей говорит и широкая область их практического применения. Так, на основе линейных моделей построено множество сложных методов машинного обучения, в том числе и нейронные сети. Найденное линейное уравнение может быть начальной точкой для построения более сложных моделей. Ещё одним преимуществом парных линейных регрессионных моделей является возможность приведения большинства нелинейных моделей к линейному виду. Поэтому парные линейные регрессионные модели применяются в различных областях науки.

Следует отметить, что несмотря на многообразие методов обнаружения выбросов в исходных статистических данных, существующие подходы, в

большинстве своём, применимы только для одномерных выборок (чаще всего временных рядов) и на каждой итерации метода осуществляется анализ лишь одного подозрительного значения. Когда речь идёт о зависимости между несколькими переменными, методы, основанные на поиске отклонения от среднего или предыдущего (последующего) значения не дают положительных результатов. Это связано с тем, что не учитывается наклон линии регрессии. Помимо этого, описанные в литературе методы являются чувствительными к объёму исходной выборки.

Поэтому, разработка и реализация алгоритма поиска аномальных измерений в исходных данных и основанных на нём методов последующей обработки данных с целью повышения качества парных регрессионных моделей для дальнейшего их использования в прогнозировании, проектировании, здравоохранении и других областях является актуальной научно-прикладной задачей.

Степень разработанности темы исследования. Исследования, связанные с обнаружением аномальных данных, проводились уже в XIX веке. По данной теме написано большое количество статей и книг различными учёными, например, такими как Россиув и Лерой, Барнет и Левис, Бекман и Кук, Хоакинс, Ходж и Остин, Маркоу и Синх и др. Также необходимо отметить работы Дж. Тьюки, Д. Хоглина, Ф. Мостеллера, В.М. Бухштабера, С.А. Айвазяна, П. Веллемана, И.С. Енюкова, Л.Д. Мешалкина. В современной литературе описаны десятки различных методов нахождения и устранения выбросов из исходных статистических данных. Основными из них являются: метод Граббса, метод Титьена-Мура-Бекмана, методы Эктона и Прескотта-Лунда. Их главные преимущества – простота понимания и применения. Однако, существующие методы имеют ряд общих недостатков:

– методы плохо формализованы и заключаются в поиске лишь одного аномального значения на каждом шаге;

– большинство из них применяются только для одномерных выборок, т.е. поиск аномалий осуществляется только по одной из переменных;

− нет конкретных рекомендаций о дальнейших действиях исследователя после нахождения выбросов в исходных данных;

− существующие методы обнаружения ненадежных и аномальных измерений опираются на конкретные законы распределения вероятностей, однако исследователю они априорно не известны;

− чувствительность большинства методов к объёму выборки;

– большинство существующих методов реализованы в таких специализированных программных пакетах, как MathCad, MatLab, Statistica, Mathematica и др.; однако данные программные средства, в основном, ориентированы на математиков и инженеров, поэтому являются достаточно сложными в использовании для специалистов других областей; помимо этого, в них не уделяется достаточное внимание последующему анализу и обработке аномальных значений.

Таким образом, существующие методы обнаружения аномальных данных имеют ряд существенных недостатков, что делает актуальным новые исследования и разработки в этой области для повышения качества регрессионных моделей.

Целью диссертационной работы является совершенствование методов предварительной обработки исходных статистических данных для повышения точности линейных регрессионных моделей при построении эффективных прогнозов.

Для достижения цели поставлены и решены следующие задачи:

1. Выполнить сравнительный анализ существующих методов обнаружения аномальных и ненадёжных измерений.

2. Предложить и обосновать усовершенствованные методы обработки статистических данных, эффективность которых не должна зависеть от объёма исходной выборки, а их применение – не приводить к ухудшению качественных характеристик модели.

3. Дать рекомендации по выбору конкретного метода обработки исходных статистических данных в зависимости от вида модели.

4. На основе усовершенствованных методов поиска и корректировки исходных данных выполнить построение алгоритмов и осуществить разработку программного комплекса.

5. Провести программное моделирование с использованием разработанного комплекса программ для оценки эффективности предложенных методов.

Объект исследования. Объектом исследования является процесс анализа данных, основанный на использовании парных линейных регрессионных моделей**.**

Предмет исследования. Предметом исследования являются математические модели, методы и алгоритмы обработки статистических данных для повышения качества регрессионных моделей.

Научная новизна полученных результатов заключается в следующем.

1. Разработан новый метод поиска аномалий, основанный на построении области надёжности, которая зависит от наклона уравнения регрессии, доверительной вероятности и соответствующего коэффициента, что позволяет одновременно обнаруживать аномальные измерения как по независимой переменной (X), так и по зависимой переменной (Y). Это приводит к повышению качества прогнозов (от 10%), полученных по линейным регрессионным моделям, возрастанию коэффициента детерминации (от 10% до 30%) и уменьшение трудоёмкости (количество элементарных операций) – до $2*10^n$ раз, по сравнению с существующими методами.

2. Получил дальнейшее развитие метод корректировки аномалий, отличием которого является изменение значений аномальных статистических данных на значения, соответствующие рассчитанной области надёжности, а также отсутствие сокращения объема исходных статических данных из-за отбрасывания, что особенно важно при моделировании на выборках малого объёма.

3. Предложены два упрощения алгоритма поиска аномальных данных, позволяющие сократить трудоёмкость анализа, выбор которых зависит от надёжности исходных значений X и Y, что приводит к обнаружению аномальных данных по одной из соответствующих переменных. Спецификой первого

упрощения является то, что оно может быть использовано для поиска аномальных данных при многомерной линейной зависимости.

4. Обоснована возможность применения предлагаемого в работе метода обнаружения выбросов не только для линейных регрессионных прогнозных уравнений, но и для нелинейных моделей с внутренней линейностью.

Теоретическая и практическая значимость работы. Теоретическая значимость результатов работы заключается в том, что предлагаемые методы повышения качества регрессионных моделей, основанные на обнаружении и последующей обработке аномальных измерений в исходных статистических данных, являются эффективным инструментом для последующей разработки точных прогнозов, использующихся в различных отраслях науки и техники. В частности:

1. Показано, что предложенный в работе подход позволяет обнаружить выбросы и скорректировать вид модели без дополнительного графического отображения (на примере трёх наборов данных из «квартета Энскомба»).

2. Предложенные методы поиска и корректировки аномалий не имеют ограничений на объём выборки, в отличие от существующих.

3. Предложенные методы поиска аномальных данных и их последующей корректировки в дальнейшем могут быть дополнены и расширены для применения при построении многомерных регрессионных моделей.

Практическая значимость работы состоит в том, что результаты работы могут применяться в различных предметных областях, таких как здравоохранение, экономика и других, при решении задач прогнозирования, проектирования, оптимизации и т.д. В работе определены оптимальные параметры использования предложенных методов корректировки исходных данных, на основе которых даны практические рекомендации по выбору конкретного метода. Разработан оригинальный комплекс программ, реализующий новый алгоритм поиска аномальных данных и методы их последующей обработки, отличающийся наличием различных модулей для автоматизированной обработки исходных

статистических данных, их графического отображения и построения наилучшей модели.

Методология и методы исследования. Для решения поставленных задач в работе использовались методы теории вероятности, математической статистики, математического моделирования, численные методы, регрессионный анализ.

Связь работы с научными программами, планами, темами. Работа выполнена в соответствии с тематическими планами Донецкого национального технического университета и является частью исследований, в которых автор принимала участие как исполнитель: гостемы Н-22-10 «Программное обеспечение высокопроизводительных вычислительных, интеллектуальных и моделирующих систем»; гостемы Н-1-16 «Анализ современных методов инженерии программного обеспечения для информационно-вычислительных и интеллектуальных систем»; гостемы Н-16-18 «Исследование методов, технологий и средств инженерии программного обеспечения на различных классах приложений»; гостемы Н-2020-14 «Усовершенствование средств инженерии программного обеспечения для актуальных классов IT-приложений».

Научные положения, выносимые на защиту.

1. Показано, что построение области надёжности («коридора»), представляющей собой прямоугольную область, размер которой зависит от заданного значения вероятности и величин среднеквадратических отклонений, позволяет эффективно обнаруживать ненадежные данные, как данные, которые не попали в эту область, в результате чего достигается повышение качества исходной модели (значение коэффициента детерминации R^2 может увеличиваться до 30%). При этом, при отбрасывании данных не происходит ухудшения качественных характеристик модели, поскольку число отброшенных наблюдений не является критическим и, как правило, составляет от 5% до 20% исходных данных.

2. Показано, что предлагаемый алгоритм применим также к нелинейным регрессионным моделям с внутренней линейностью (экспоненциальная, логарифмическая, степенная и др.).

3. Доказано, что применение алгоритма построения области надежности, а также соответствующей стратегии исключения/изменения данных, исходя из объема имеющейся выборки, позволяет сократить временные затраты за счет уменьшения количества вычислительных операций (при этом сокращение может быть от 4*n до $2*10^s$, где n – количество исходных данных, s – число аномальных измерений) и получить регрессионные модели более высокой точности.

Степень достоверности и апробация результатов. Достоверность результатов исследования обеспечивается достаточным количеством проведенных экспериментальных вычислений с использованием реальных и модельных данных. Подготовка, анализ исходных данных и интерпретация итоговых результатов базируются на современных методах обработки информации и статистического анализа.

Практическая ценность исследований подтверждается внедрением в ООО НПО «Интермет» (справка о внедрении от 23 июня 2021 г.), в научно-исследовательские работы ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ» (справка о внедрении № 29-13/15 от 05 июля 2021 г.), в учебный процесс ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ» при чтении лекций и проведении лабораторных работ на кафедре «Программная инженерия» им. Л.П. Фельдмана по дисциплинам: «Эмпирические методы программной инженерии», «Численные методы в информатике» (справка о внедрении № 29-12/15 от 05 июля 2021 г.).

Апробация результатов диссертации.

Основные положения диссертационной работы докладывались и обсуждались на: VIII Международной научно-практической конференции «Математическое и программное обеспечение интеллектуальных систем (MPZIS-2010)», г. Днепропетровск: ДНУ им. Олеся Гончара, 2010 г.; «Донбас-2020 перспективи розвитку очами молодих вчених», г. Донецк, ДонНТУ, 2010 г; XIX Международной научно-практической конференции MicroCAD-2011, г. Харьков, 2011 г.; Одиннадцатой международной научно-технической конференции «Проблемы информатики и моделирования» г. Харьков: НТУ «ХПИ», 2011 г.; IV

Международной научно-технической конференции «Моделирование и компьютерная графика - 2011», г. Донецк, 2011 г., VII Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2016)», г. Донецк, 2016 г.; II Международной научно-практическая конференция «Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2018)», г. Донецк, 2018 г.; VIII Международной научно-практической конференции «Современные тенденции развития и перспективы внедрения инновационных технологий в машиностроении, образовании и экономике», г. Азов, 2021 г.

Личный вклад соискателя. Все результаты и положения, составляющие основное содержание диссертации, вынесенные на защиту, получены соискателем самостоятельно в процессе научных исследований. Личный вклад автора заключается в обосновании идеи и цели работы, её реализации, а также в проведении теоретических и экспериментальных исследований, разработке вычислительных алгоритмов и комплекса программ для их компьютерной реализации, разработке рекомендаций по практическому применению результатов.

Публикации. Основные научные результаты диссертации опубликованы в 17 научных работах, из них 2 статьи в специализированных изданиях, рекомендованных ВАК ДНР, 4 – в изданиях, входящих в перечень научных изданий, утверждённых ВАК Украины, 2 – в других научных изданиях (в том числе 1 монография), 9 – в материалах международных научных конференций.

Соответствие темы и содержания диссертации паспорту специальности.

По направлению исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки), в частности: п.3 «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий»; п.5 «Комплексные исследования научных и технических проблем с применением современных

технологий математического моделирования и вычислительного эксперимента»; п.6 «Разработка новых математических методов и алгоритмов проверки адекватности математических моделей объектов на основе данных натурного эксперимента».

Структура и объём работы. Диссертация состоит из введения, четырех разделов, заключения, списка литературы и 4 приложений. Изложена на 162 страницах машинописного текста, включая 44 рисунка, 36 таблиц, список литературы из 98 наименований.

РАЗДЕЛ 1
ПОСТАНОВКА ЗАДАЧИ И АНАЛИЗ ИЗВЕСТНЫХ МЕТОДОВ ОБНАРУЖЕНИЯ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ

1.1 Обзор общих методов обнаружения аномалий

Поиск и обработка аномальных данных является важной задачей, которой посвящено множество различных исследований. Аномалиями (выбросами, исключениями) называются значения, которые существенно отличаются от остальных данных или не согласуются с ними. Обнаружение выбросов относится к проблеме поиска образцов в данных, которые не соответствуют ожидаемому поведению. Исследования, связанные с обнаружением аномальных данных, проводились уже в XIX веке. Так в 1887 г. английским экономистом и математиком Фрэнсисом Эджуортом была опубликована статья «О противоречивых наблюдениях» [1]. Статья посвящена поиску необычных или аномальных образцов в массиве данных. По данной теме написано большое количество статей и книг различными учёными, например, такими как Россиув и Лерой, Барнет и Левис, Бекман и Кук, Хоакинс, Ходж и Окстин, Маркоу и Синх и др. [2]. Также необходимо отметить работы Дж. Тьюки, Д.Хоглина, Ф.Мостеллера, В.М. Бухштабера, С.А. Айвазяна, П. Веллемана, И.С. Енюкова, Л.Д. Мешалкина. О важности поиска аномальных измерений, говорит и наличие государственных стандартов в этой области, в частности ГОСТ 8.736-2011 [3], ГОСТ Р ИСО 16269-4-2017 [4]. Несмотря на большое количество работ, поиск и дальнейшая обработка аномалий является актуальной проблемой.

Аномальные значения делятся на 2 класса:

- искусственные – вызванные некорректной работой программы, оборудования, ошибками измерений или ввода данных;

- естественные – это реальные факты и события, которые происходят редко.

На сегодняшний день существуют следующие методы распознавания аномалий:

- статистический анализ;

- кластеризация;

- алгоритм ближайшего соседа;

- классификация;

- спектральные методы;

- гибридные методы.

При использовании статистического анализа определяется разница между построенной моделью и реальными данными. Если эта разница превышает определённый порог, то в данных существуют аномалии. Выделяют следующие группы методов статистического анализа:

- параметрические методы (на основе Гауссовой модели, на основе регрессионной модели, их комбинация);

- непараметрические методы (методы на основе гистограмм или функций ядра).

Кластеризация заключается в том, что все похожие экземпляры группируются в кластеры, если какой-либо экземпляр удален от центров кластеров более чем на определенную величину, то он считается аномальным, при этом дополнительные знания о свойствах возможных исключений не требуются.

В алгоритмах ближайшего соседа осуществляется определение расстояния или меры сходства между двумя экземплярами данных.

Метод классификации заключается в том, что наблюдения делятся на один или несколько классов, а те наблюдения, которые не принадлежат ни к одному из классов, признаются выбросами. Самыми распространенными подходами в этом методе являются:

- нейронные сети;

- Байесовы сети;

- метод на основе правил;

- метод опорных векторов. Суть метода опорных векторов в том, что создается линия или гиперплоскость, которая разделяет данные на классы.

При использовании спектрального метода на основе частотных характеристик данных строится модель, которая должна учесть большую часть изменчивости в данных.

В таблице 1.1 приведены примеры основных областей и решаемых задач, в которых применяется поиск выбросов, а также наиболее часто используемые методы применительно к каждой области.

Таблица 1.1 – Примеры применения методов поиска выбросов в различных областях

Область	Пример задачи	Метод
Медицина	вспышки заболеваний, отклонения в состоянии пациентов, ошибки записи	параметрические статистические методы, нейронные сети, байесовские нейронные сети, методы на основе правил, алгоритм ближайших соседей
Астрономия	отделение квазаров (активное ядро галактики) от звёзд	алгоритм ближайшего соседа
Компьютерные сети	обнаружение сетевых вторжений, взломов	все виды статистических методов, все виды классификации, кластеризация, ближайшего соседа
Обнаружение мошенничества	мошенничество с кредитными картами, мобильными телефонами, страховые агентства	статистические методы с использованием гистограмм, параметрические статистические методы, нейронные сети, методы на основе правил, кластеризация
Промышленность	поломки оборудования	параметрические и непараметрические статистические методы, нейронные сети, спектральный анализ
Торговля	выявление аномального спроса	параметрические статистические методы
Обработка изображений	спутниковые изображения, распознавание цифр, медицинские снимки	параметрические статистические методы, нейронные и байесовские сети, кластеризация, алгоритм ближайшего соседа

Как видно из таблицы, параметрические статистические методы используются для поиска аномальных значений в выборке практически во всех предметных областях.

Такое положение делает актуальной задачу совершенствования параметрических статистических методов поиска и обработки аномалий.

1.2 Обзор общих методов обработки аномальных данных

В анализе данных выделяются два направления, занимающиеся поиском аномалий – обнаружение выбросов и обнаружение новизны.

Новизна отличается от выброса тем, что указывает на некоторое изменение в системе и не является следствием ошибок в данных, вызванных неточными измерениями, некорректным вводом данных и т.д. В этом случае задача заключается именно в своевременном выявлении аномалии при её появлении в выборке. Целью такого поиска может служить обнаружение сетевых вторжений, мошенничества с кредитными картами, неисправностей функционирования оборудования. Т.е. в данных ситуациях пользователя интересуют сами выбросы, как индикаторы отклонения от нормы и именно они являются целью анализа. Поэтому, после обнаружения такие данные не подвергаются дальнейшей обработке (отбрасыванию или изменению).

В случае, если данные исследуются с целью построения корректной модели, наиболее точно описывающей определённую зависимость и использования этой модели для дальнейшего анализа, то после обнаружения аномалий они либо исключаются из выборки, либо корректируются.

Удаление аномальных значений из выборки можно производить тогда, когда она содержит достаточное количество данных для анализа, чтобы не снизить существенно репрезентативность выборки.

В противном случае, используются различные методы корректировки. Основные из них:

- ручная замена – используется при небольшом числе аномальных измерений;

- замена на наиболее вероятное значение;

- интерполяция данных – замена выбросов, значениями, полученными на основе ближайших соседей;

- сглаживание данных.

1.3 Анализ методов решения задачи поиска аномалий и их обработки методами регрессионного анализа

Проведем детальный анализ методов решения задачи поиска аномалий и их обработки методами регрессионного анализа с целью:

- определения специфики решения задачи поиска аномалий таким путем;
- выявления перспективных направлений решения данной задачи.

1.3.1 Основные этапы решения задач моделирования методами регрессионного анализа

Для решения задач с помощью методов регрессии как инструмента математического моделирования необходимо пройти ряд этапов [5, 6]:

1. Математическая постановка задачи. На этом этапе строится математическая модель. Математическая модель – приближенное описание объектов с помощью математических соотношений. Для того, чтобы правильно поставить задачу, для корректного построения модели, предварительно необходимо осуществить детальный анализ предметной области и изучить существующую проблему [7, 8, 9, 10].
2. Применение численных методов. Численные методы – раздел вычислительной математики, изучающий приближенные методы решения различных задач на уровне математических моделей. На данном этапе необходимо выбрать наиболее подходящий метод решения поставленной задачи.
3. Разработка алгоритма. Алгоритмы используются для наглядного представления последовательности необходимых действий.
4. Программирование. На данном этапе осуществляется программная реализация разработанного алгоритма.
5. Тестирование и отладка программы. Проводится проверка правильности работы программы на ряде тестовых задач.

6. Вычислительный эксперимент. На данном этапе производятся расчеты по исходным данным [11, 12, 13, 14].

7. Анализ результатов. Проводится анализ, полученных на предыдущем этапе результатов. Если результаты не соответствуют ожидаемым, то возникает необходимость возврата на один из предыдущих этапов.

При оценке эффективности выбранных для решения поставленной задачи численных методов, необходимо учитывать такие его свойства, как устойчивость, сходимость и простота реализации. Также следует помнить о том, что численное решение задачи, как правило, содержит некоторую погрешность, величина которой зависит от ряда факторов [15].

Одной из важных задач при использовании численных методов является нахождение баланса между уменьшением трудоёмкости и увеличением точности. В данном случае трудоёмкость измеряется объёмом памяти необходимым для поиска решения и временем, т.е. количеством элементарных операций, требующимся для выполнения всех вычислений.

Примерами типовых задач, решаемых в данной работе, с использованием численных методов, являются численное решение уравнений, численное интегрирование, аппроксимация и т.д.

1.3.2 Специфика и проблемы развития современных методов регрессионного анализа как инструмента решения задачи поиска аномалий и их обработки

В современном мире практически в каждой сфере человеческой деятельности возникает необходимость в получении информации о дальнейшем развитии какого-либо объекта или процесса на основании имеющихся на текущий момент данных, т.е. в прогнозировании [16, 17, 18]. Регрессионные модели широко используются для решения различных прикладных задач (в экономике, медицине, социологии, при инвестиционном анализе и т.д.), в том числе и для осуществления прогнозов. В настоящее время в любой сфере деятельности можно получить большое количество эмпирических данных. Сбор этой информации необходим для

дальнейшего анализа. Регрессионный анализ позволяет выявить взаимосвязь между наблюдаемыми величинами, а также предсказать значения переменной [19, 20, 21, 22, 23, 24, 25, 26]. Если же зависимость достаточно сложная, необходимо попробовать подобрать простую приближенную функцию для определённой ограниченной области. Это позволит выявить поведение модели на одном из участков.

Частным случаем регрессионного анализа является метод линейной регрессии. Простая (парная) линейная регрессия – статистический метод, позволяющий предсказывать значения зависимой переменной Y по значениям независимой переменной X [27, 28, 29].

Линейное уравнение в этом случае имеет следующий вид:

$$Y = \alpha X + \beta + \varepsilon, \quad (1.1)$$

где α и β – параметры модели, определяемые в результате регрессионного анализа;

ε – случайные ошибки (невязки) переменной Y.

Параметры генеральной совокупности α и β нельзя определить точно, но можно найти их оценки a и b соответственно.

Таким образом, уравнение простой линейной регрессии примет вид:

$$\hat{Y}_i = aX_i + b, \quad (1.2)$$

где $\hat{Y}_i$ – предсказанное значение переменной Y для i-го наблюдения (i = 1..n);

X_i – значение переменной X для i-го наблюдения (i = 1..n);

a – коэффициент регрессии, определяющий наклон к оси OX;

b – коэффициент регрессии, который определяет точку пересечения прямой регрессии с осью ординат [30, 31].

Для того, чтобы найти оценки a и b широко используется метод наименьших квадратов (МНК, англ. Ordinary Least Squares, OLS). Основная идея метода заключается в том, что находятся такие параметры a и b, при которых квадрат разности между фактическими значениями Y_i и предсказанными $\hat{Y}_i$ будет минимальным [32, 33, 34]

$$S = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \rightarrow min\ . \quad (1.3)$$

Согласно условиям Гаусса-Маркова, для получения наилучших результатов в ходе регрессионного анализа, случайный член должен удовлетворять следующим условиям [35]:

1) математическое ожидание случайного члена в любом наблюдении должно быть равно 0;

2) дисперсия случайного члена должна быть постоянна для всех наблюдений;

3) значения случайного члена должны быть независимы между собой;

4) случайная ошибка должна быть распределена независимо от объясняющих переменных, т.е. $\sigma_{xi,ei}=0$ для всех наблюдений.

В диссертационной работе будут рассматриваться методы повышения качества прогнозов, построенных на основании парных линейных регрессионных моделей. Выбор линейных моделей был обусловлен их преимуществами [36, 37, 38, 39, 40, 41, 42]:

1) эти зависимости являются хорошо изученными;

2) данный вид модели является широко используемым, благодаря своей простоте и наглядности;

3) делают возможным применение математического аппарата для четкой статистической обработки;

4) с помощью полученной модели можно сделать однозначные выводы о влиянии определённого фактора на результат.

Для того, чтобы построить адекватную регрессионную модель, использование которой позволит получить наиболее точный прогноз, необходимо иметь достаточное количество надёжных исходных данных. Поэтому задача обнаружения и последующей обработки (устранения или корректировки) аномальных значений является важной и актуальной.

1.3.3 Детальный анализ основных методов решения задачи поиска аномалий и их обработки

Если какие-либо результаты измерений (экспериментов) значительно отклоняются от средних значений, то следует проверить, являются ли эти результаты случайными или нет. На текущий момент существует множество различных критериев обнаружения аномальных данных. Их можно разбить на две основные группы – оценка выбросов во временных рядах или одномерной выборке и оценка выбросов в регрессии (Рисунок 1.1)

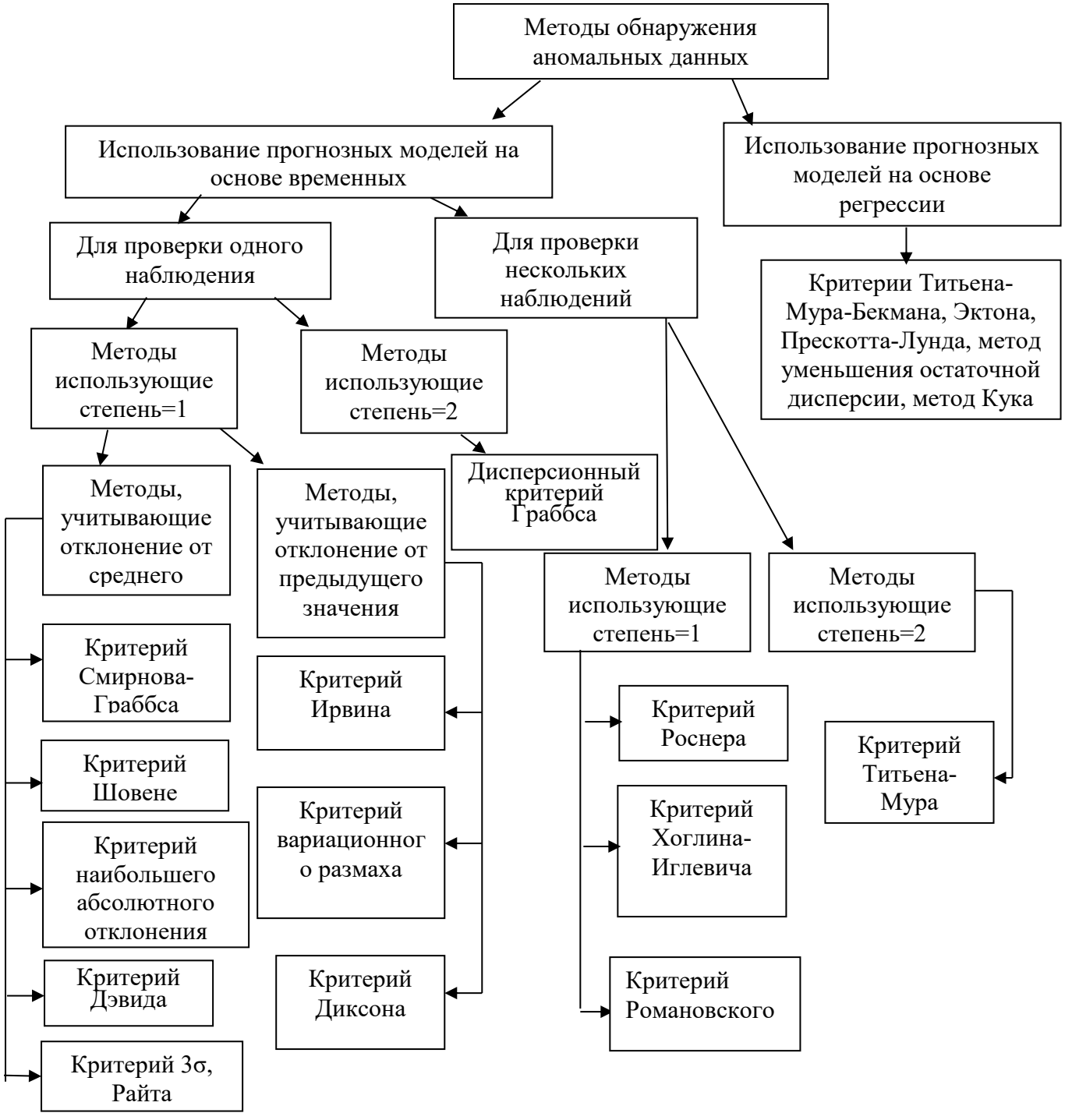


Рисунок 1.1 – Методы обнаружения аномальных данных

К первой группе относятся критерии: Смирнова-Граббса [43], Шовене [44], Дэвида, Романовского [45, 46], 3σ, Райта [47], наибольшего абсолютного отклонения, вариационного размаха, Диксона, Роснера, Хоглина-Иглевича, Титьена-Мура [48]. Ко 2-й группе – критерии Титьена-Мура-Бекмана, Эктона, Прескотта-Лунда, метод уменьшения остаточной дисперсии, расстояние Кука [49].

1.3.3.1 Методы обнаружения и устранения аномальных данных временных рядов

1.3.3.1.1 Методы для проверки одного подозрительного наблюдения

В данном подразделе будут проанализированы методы для проверки одного подозрительного значения. Они, в свою очередь, включают в себя методы, учитывающие при расчете, отклонение от среднего значения и методы, учитывающие отклонение от предыдущего значения.

Рассмотрим вначале методы, учитывающие отклонение от среднего. К ним относятся: критерий Смирнова-Граббса, критерий Шовене, критерий наибольшего абсолютного отклонения, критерий Дэвида, критерий 3σ и критерий Райта.

Пусть $x_1, x_2, \ldots, x_n$ – наблюдаемая выборка, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ – построенный по ней вариационный ряд (т. е. упорядоченный по возрастанию) [50].

Статистической характеристикой критерия Смирнова-Граббса является стандартизованное предельное отклонение аномального значения x_n (x_1) от среднего $\bar{x}$. Если речь идет о максимальном значении x_n, то по формуле 1.4, находится значение T_1:

$$T_1 = \frac{x_n - \bar{x}}{\sigma}, \tag{1.4}$$

где $\bar{x}$ и σ определяются для совокупности в целом:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \tag{1.5}$$

Если проверяется минимальное значение x_1, то

$$T_2 = \frac{\bar{x} - x_1}{\sigma} . \tag{1.6}$$

Фактическое значение T_1 или T_2 сравнивается с критическим значением $T_1(\alpha)$ или $T_2(\alpha)$. Если фактическое значение T_n меньше критического, то отклонения с вероятностью α считаются случайными, а если превышает критическое, то отклонения считаются существенными, а значение – аномальным. В таком случае это значение исключается и критерий применяется к x_{n-1} и т. д., пока не будет признано, что выбросов нет, а, следовательно, совокупность однородная [51, 52, 53, 54].

Критические значения $T_1(\alpha)$ и $T_2(\alpha)$ при $n \leq 25$ приведены в таблице А.1 приложения А. Если же n больше этого числа, то критические значения $T_1(\alpha)$ можно определить, воспользовавшись следующим приближением:

$$T_1(\alpha) = u_{1+\frac{\alpha-1}{n}}\sqrt{\frac{n-1}{n}}, \tag{1.7}$$

где u_γ – γ-квантиль стандартного нормального распределения.

Для $T_2(\alpha)$ используется приближение при α=0,95

$$T_2(0{,}95) = \begin{cases} 1{,}31 + 0{,}435\ln(n-2{,}7) \;\; при 5 \leq n < 35; \\ 1{,}962 + 0{,}281\ln(n-15) \;\; при 35 \leq n < 500. \end{cases} \tag{1.8}$$

Критерий Смирнова-Граббса может применяться в случае, когда дисперсия известна заранее, а также когда дисперсия оценивается по выборке с помощью соотношения 1.5.

Схематически данный метод можно представить в виде рисунка 1.2:

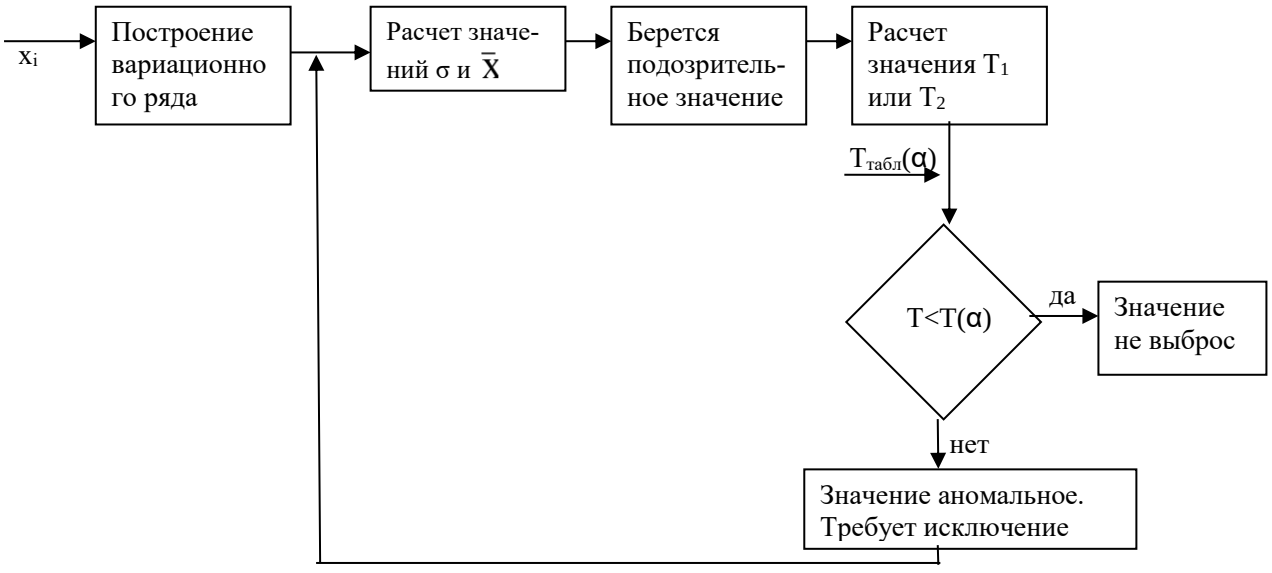


Рисунок 1.2 – Схематическое представление критерия Смирнова-Граббса

Следующий рассматриваемый критерий – критерий Шовене [55, 56].

Из полученного ряда, содержащего n значений, выбирается элемент x_k, проверяемый на наличие погрешности и вычисляется отношение модуля разности его величины и среднего значения к среднему квадратическому отклонению:

$$Z = \frac{|x_k - \bar{x}|}{\sigma}. \quad (1.9)$$

Затем, из таблицы нормированного нормального распределения по величине z вычисляется вероятность этого отклонения (1.10), а также ожидаемое число n измерений, которые дадут отсчеты, имеющие отклонение Z не меньшее, чем испытуемый (1.11).

$$P(z\sigma<|x_k - \bar{x}|), \quad (1.10)$$

$$n_{ож} = nP. \quad (1.11)$$

Если получено $n_{ож}<0.5$, то элемент x_k считается промахом [51, 53].

Ещё один критерий – критерий наибольшего абсолютного отклонения основан на нахождении отношения максимального отклонения значения от среднего и СКО (1.12). Здесь в отличие от двух предыдущих методов σ находится по формуле 1.13.

$$\tau = \frac{\max|x_i - \bar{x}|}{\sigma}, \quad (1.12)$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \quad (1.13)$$

Если фактическое значение $\tau \geq \tau(\alpha)$, то значение x_i признается выбросом. Критические значения $\tau(\alpha)$, приведены в таблице А.2 приложения А.

Для большого объёма выборки в [57] предложено весьма точное приближение (с точностью менее 0,1) для $\tau(\alpha)$ при α=0,95 (1.14):

$$\tau(0,95) = \begin{cases} 1,39 + 0,462\ln(n-3) \text{ при } 5 \leq n < 35; \\ 2,136 - 0,281\ln(n-15) \text{ при } 35 \leq n \leq 500. \end{cases} \quad (1.14)$$

Критерий Дэвида представляет собой модификацию критерия Смирнова-Граббса.

Здесь также в расчетах используются максимальные и минимальные значения соответственно:

$$T = \frac{x_n - \bar{x}}{\sigma}, \tag{1.15}$$

$$T^* = \frac{\bar{x} - x_1}{\sigma}, \tag{1.16}$$

где $\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$ – выборочная дисперсия, оцениваемая по отдельной независимой выборке (объема m) значений y_i.

Критические значения T(α) статистики Дэвида можно определить по таблице А.3. Если n превышает 25, то можно воспользоваться аппроксимацией:

$$T(\alpha) \approx u_{1+\frac{\alpha-1}{n}}\left(1 + \frac{3}{m-1}\sqrt{\frac{n-1}{n}}\right), \tag{1.17}$$

где u_γ – γ-квантиль стандартного нормального распределения.

При $T(T^*) \geq T(\alpha)$ значение с вероятностью α признается выбросом.

Достаточно простым критерием является критерий «трёх сигм». Определяется разность между проверяемым значением x_i и средним, рассчитанным по исходной выборке. Если найденное значение превышает величину равную 3σ, то измерение x_i признаётся выбросом и исключается.

Аналогичным критерием является критерий Райта. Разница лишь в том, что для сравнения используется величина 4σ.

Во всех вышеописанных критериях при определении коэффициентов используется первая степень. В отличии от них в дисперсионном критерии Граббса применяется вторая степень. Он основан на сравнении сумм квадратов отклонений от среднего исходной и сокращенной (без крайнего элемента) выборок.

По формуле 1.18, находится статистика дисперсионного критерия Граббса для проверки максимального значения выборки.

$$G_{оп} = \frac{\sum_{i=1}^{n-1}(x_i - \tilde{x}_1)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1.18}$$

где $\overline{x}$ – среднее по всей выборке;

$\tilde{x}_1$ – среднее по выборке без сомнительных результатов

$$\tilde{x}_1 = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i . \qquad (1.19)$$

Если значения статистики G_{on} окажется меньше критического значения, то наблюдение относят к грубым ошибкам.

После отсева наблюдений, признанных нетипичными, проверку на грубые ошибки повторяют для сокращенной выборки.

Таким образом, исследовав перечисленные выше методы, можно сделать следующие выводы:

1) основным преимуществом данных методов является относительная простота их применения и возможность использования для небольшого объема выборки;

2) недостатками данных методов является то, что для выборок большого объема трудоемкость будет увеличиваться, т.к. расчеты необходимо проводить для каждого подозрительного значения в отдельности. Также существует вероятность того, что в выборке большее число аномальных измерений, чем исследуется на выбросы, что может отрицательно сказаться на результатах анализа.

3) рассмотренные критерии применимы для проверки на аномальность единичных наблюдений (минимального или максимального), однако в ситуации, когда выборка содержит группу близких по значениям аномальных наблюдений, они могут не дать нужного результата.

4) помимо перечисленных недостатков критерий Граббса имеет еще один существенный недостаток – при определении статистики критерия используется вторая степень, что уменьшает точность расчетов;

5) критерии «трёх сигм» и Райта не рекомендуется использовать дважды.

Теперь проанализируем методы, в которых сравниваются определённые значения ряда. К ним относятся критерии Ирвина, вариационного размаха и Диксона.

Метод Ирвина принадлежит к этапу предварительного анализа временных рядов и заключается в проверке однородности ряда, т.е. в выявлении аномальных значений. Данный метод основан на определении отношения отклонения подозрительного значения x_i от предыдущего значения ряда x_{i-1} (где i=2..n) к СКО. Выявление аномальных наблюдений производится по следующей формуле:

$$\lambda_i = \frac{|x_i - x_{i-1}|}{\sigma}. \quad (1.20)$$

Затем сравнивается коэффициент λ_i с табличным значением критерия Ирвина λ_α. Если расчетное значение превысит уровень критического, то оно признается аномальным. Табличное значение λ_α определяется при уровне значимости α и числе степеней свободы k = n-2 (Таблица А.4).

Второй критерий – критерий вариационного размаха. Он представляет собой достаточно простой метод исключения грубой погрешности измерения. Критерий заключается в определении разности (размаха) между наибольшим и наименьшим значениями признака:

$$R_n = x_n - x_1. \quad (1.21)$$

Далее проверяется подозрительное значение x_i. Для этого находится интервал надёжности по формуле:

$$\overline{X} - z \cdot R_n < x_i < \overline{X} + z \cdot R_n, \quad (1.22)$$

где $\overline{X}$ – выборочное среднее арифметическое значение, вычисленное после исключения предполагаемого промаха;

z – критериальное значение, зависящее от числа членов вариационного ряда. Определяется согласно таблице 1.2.

Если значение x_i попадает в интервал, то делается вывод о его надёжности. Если условие (1.22) не выполняется, то такое измерение следует исключить из ряда.

Таблица 1.2 – Критерий вариационного размаха

n	5	6	7	8-9	10-11	12-15	16-22	23-25	26-63	64-150
z	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8

Третий критерий – критерий Диксона. Он определяется в зависимости от количества измерений n по формулам:

– при $3 \leq n \leq 7$ для проверки одного сомнительного наблюдения происходит сравнение отклонения подозрительного (n-го) значения от предыдущего (для проверки последнего значения ряда) и размаха (формула 1.23), или отклонение первого значения от второго (для проверки первого, т. е. минимального) значения ряда (формула 1.24):

$$r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}, \tag{1.23}$$

$$r_{10} = \frac{x_2 - x_1}{x_n - x_1}; \tag{1.24}$$

– при $8 \leq n \leq 10$ сравнивается отклонение подозрительного значения и размах, но размах считается независимо от противоположного крайнего наблюдения.

$$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}, \tag{1.25}$$

$$r_{11} = \frac{x_n - x_{n-1}}{x_n - x_2}; \tag{1.26}$$

– при $11 \leq n \leq 13$ рассчитывается отношение подозрительного значения без учета, следующего по величине и размаха, который считается независимо от противоположного крайнего наблюдения:

$$r_{21} = \frac{x_3 - x_1}{x_{n-1} - x_1}, \tag{1.27}$$

$$r_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}; \tag{1.28}$$

– при $17 \leq n \leq 25$ расчет производится аналогично предыдущему методу, но размах считается независимо от двух противоположных крайних:

$$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1}, \tag{1.29}$$

$$r_{22} = \frac{x_n - x_{n-2}}{x_n - x_3}. \quad (1.30)$$

После расчета значений критериев Диксона они сравниваются с критическими значения статистик r(α) (Таблица А.5). Если рассчитанное значение r превышает критическое r(α), то значение признаётся выбросом.

Проанализировав перечисленные выше методы можно сделать вывод о том, что они являются сравнительно простыми.

Недостатками методов является то, что рассчитывается разница между значением ряда, например, с номером n и n-1 (для максимальных значений), однако если эти значения оба являются аномальными, то вывод об отклонении исследуемого значения может быть не корректным. Т.е. этот критерий может не дать реального результата при наличии нескольких близких выбросов на том или ином конце ряда. Для наибольшей точности результатов, рекомендуется использовать при проверке одновременно несколько критериев, а это усложняет анализ и увеличивает его длительность.

1.3.3.1.2 Методы для обнаружения нескольких выбросов в выборке

К методам обнаружения сразу нескольких выбросов в выборке относятся критерии Хоглина-Иглевича, Роснера и Романовского. Также к такого рода методам относятся метод уменьшения остаточной дисперсии и критерий Титьена-Мура, однако при вычислениях используется вторая степень.

Согласно критерию Хоглина-Иглевича [58] наблюдение считается выбросом, если его значение выходит за пределы интервала $((1+k)x_{[l]} - kx_{[n+1-l]}; (1+k)x_{[n+1-l]} - kx_{[l]})$.

Для выбора значения *l* используются следующие формулы:

$$l_1 = \frac{1}{2}\left[\frac{n+3}{2}\right];\ l_2 = \left[\frac{n}{4} + \frac{5}{12}\right];\ l_3 = \left[\frac{n}{4} + \frac{1}{4}\right], \quad (1.31)$$

где [...] − целая часть числа.

В таблице А.6 представлены значения коэффициентов k.

В критерии Титьена–Мура (L-критерий) сравнивается сумма квадратов отклонений от среднего исходной и сокращенной выборок без k сомнительных результатов.

Для выявления k максимальных выбросов используется формула:

$$L_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x}_k)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (1.32)$$

где $\bar{x}_k = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i$ – среднее по выборке без сомнительных результатов.

Для нахождения k минимальных наблюдений используется формула:

$$L_k^* = \frac{\sum_{i=k+1}^{n}(x_i - \bar{x}_k^*)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (1.33)$$

где $\bar{x}_k^* = \frac{1}{n-k}\sum_{i=k+1}^{n} x_i$ – усеченное выборочное среднее.

Когда расчетная статистика не превышает критического значения, то измерение признается выбросом.

Если предполагается, что выбросами могут быть, как наименьшие, так и наибольшие наблюдения, то используется обобщенный E-критерий Титьена-Мура. Вначале необходимо найти модули отклонений $d_i = |x_i - \bar{x}|$. После этого ряд упорядочивается по возрастанию от d_1 до d_n. Элементы полученного нового ряда обозначаем через z_i. Расчет статистики осуществляется по формуле 1.35:

$$E_k = \frac{\sum_{i=1}^{n-k}(z_i - \bar{z}_k)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \qquad (1.34)$$

где $\bar{z} = \bar{x}$;

$\bar{z}_k = \frac{1}{n-k}\sum_{i=1}^{n-k} z_i$ – среднее усеченное по k значениям.

Наличие выбросов признается значимым с достоверностью α, если $E_k \le E_k(\alpha)$.

Значения $E_k(\alpha)$ приведены в таблице А.7.

Поскольку количество выбросов заранее может быть неизвестно, Роснером был предложен критерий, который предназначен для определения наличия выбросов и их количества.

Суть критерия Роснера заключается в последовательном применении критерия Граббса для определения одного выброса, основанного на статистике.

$$T_1^* = \max\left(\frac{\bar{x} - x_1}{\sigma}, \frac{x_n - \bar{x}}{\sigma}\right). \quad (1.35)$$

Для реализации критерия Роснера необходимо выполнить следующие шаги:

1) по исходной выборке объема n вычисляются среднее значение $\bar{x}$, среднеквадратическое отклонение σ и статистика T_1^*;

2) необходимо определить какое из значений (наибольшее или наименьшее) находится дальше от среднего;

3) из выборки удаляется, выбранное на предыдущем шаге значение;

4) шаги 1-3 повторяются k раза;

5) сравниваем найденные значения статистик T_{1i}^* (i=1,…, k) с критическими значениями (таблица А.8). Если статистика больше, чем критическое значение, то определяются выбросы и их количество. Количество при этом равно номеру вычисленной статистики.

6) расчеты статистик ведутся до тех пор, пока $T_{1(i+1)}^* > T_{1i}^*$.

Последним рассматриваемым критерием в данном подразделе является критерий Романовского, который заключается в следующем. Гипотеза о наличии грубых отклонений в подозрительных результатах подтверждается, если абсолютное отклонение подозрительного значения от среднего без учета сомнительных результатов наблюдений превышает либо равно произведению СКО

S, которое находится также без учета подозрительных значений, и коэффициента Стьюдента:

$$| x_i - \bar{x} | \geq t_p S, \tag{1.36}$$

где t_p – квантиль распределения Стьюдента при заданной доверительной вероятности с числом степеней свободы $k = n - k_n$ (k_n – число подозрительных результатов наблюдений) [59, 60, 61].

Основным достоинством данных методов является то, что есть возможность проводить оценку сразу нескольких выбросов в выборке. Однако при исследовании на аномальность нескольких значений, под подозрение может попасть значение, которое не является выбросом, и для проверки одного конкретного значения необходимо будет использовать другие методы. Также критерий Титьена-Мура имеет такой же недостаток, как и метод Граббса – при расчете статистики критерия используется вторая степень, что снижает точность вычислений. Критерий Титьена-Мура предполагает, что количество выбросов k заранее известно, однако это не всегда так. Проблема выявления количества выбросов рассматривается критерием Роснера. Он позволяет автоматически оценивать количество выбросов в выборке.

Для корректного решения о наличии резко выделяющихся наблюдений, также, как и критерии из предыдущего подраздела, данные критерии следует применят совместно.

1.3.3.2 Методы обнаружения и устранения аномальных данных в регрессии

1.3.3.2.1 Простые методы выявления ненадёжных измерений

Наибольший интерес в данной работе представляют методы обнаружения выбросов в регрессии, такие как критерий Эктона, Титьна-Мура-Бекмана, Прескотта-Лунда, уменьшения остаточной дисперсии.

Расчетное значение критерия Эктона (формула 1.37) представляет собой отношение абсолютной разницы между остатком предполагаемого выброса e_k и

средним по всем другим остаткам $\bar{e}$ (найденному по формуле 1.38) к среднеквадратическому отклонению S_k экспериментальных точек линии регрессии с учетом отбрасывания подозрительного наблюдения [62]:

$$V = \frac{|e_k - \bar{e}|}{S_k}, \tag{1.37}$$

$$\bar{e} = \frac{1}{n-1}\sum_{\substack{i=1 \\ i \neq k}}^{n} e_i, \tag{1.38}$$

где $e_i = Y_i - \hat{Y}_i$.

Остаток e_i с вероятностью α считается выбросом, если расчетное значение V больше критического V_α (Таблица А.9).

Критерий Эктона применяется для нахождения только одного выброса в парной линейной модели, в случае, если n≥30.

Ещё одним критерием для выявления одного промаха в линейной модели вида (1.2) является критерий Титьена-Мура-Бекмана [63, 64, 65]. Он заключается в том, что рассчитывается значение Rm, как максимальное значения из всех отношений остатков к среднеквадратическим отклонениям остатков. После этого, полученное значение Rm сравнивается с критическим значением R_α (Таблица А.10). Если полученное значение оказывается больше, то значение Y_i является выбросом.

$$Rm = \max\left|\frac{e_i}{S_i}\right|, \tag{1.39}$$

$$\text{где } S_i^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{1.40}$$

Статистика критерия Прескотта-Лунда является упрощенной модификацией предыдущего критерия и определяется по формуле (1.41).

$$R^* = \sqrt{n}\max\frac{|e_i|}{\sqrt{\sum_{i=1}^{n} e_i^2}}. \tag{1.41}$$

Критическое значение можно определить по таблице А.11 или с помощью отношения:

$$R_{\alpha}^{*} = \sqrt{\frac{(n-k)F}{n-k+1+F}}\ , \qquad (1.42)$$

$F_{(1-\frac{1-\alpha}{n})}$ – квантиль F-распределения Фишера с $f_1=1$ b $f_2=n-k-1$ степенями свободы.

Если вычисленное значение оказывается больше критического, то исследуемое значение признаётся выбросом.

Недостоверные наблюдения, участвующие в регрессионной модели, вносят самый большой вклад в величину остаточной дисперсии. Вклад каждого наблюдения в величину остаточной дисперсии можно определить, исключая по отдельности из набора значений, участвующих в расчете регрессионной модели, включенные в него наблюдения. На каждом шаге из набора значений необходимо исключить то, устранение которого приводит к наибольшему уменьшению величины остаточной дисперсии.

Основные шаги метода уменьшения остаточной дисперсии при поиске аномальных наблюдений будут следующие:

1) вычисляется частная величина $S^2_{ост}$, которая представляет собой сумму квадратов отклонений фактических значений y_i относительно значений, рассчитанных по уравнению регрессии Y_i, без учёта исследуемого наблюдения.

$$S^2_{ост} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-p-1}, \qquad (1.43)$$

где p – число параметров регрессионного уравнения;

2) наблюдение, в результате исключения которого из расчетов величина остаточной дисперсии стала наименьшей, считается выбросом и исключается из набора наблюдений, участвующих в построении уравнения регрессии;

3) шаги 1 и 2 повторяются до тех пор, пока регрессионная модель не станет адекватной по F–критерию Фишера, и пока для каждого наблюдения, участвующего в вычислениях, не станет выполняться условие

$$|Y_i - \hat{Y}_i| \leq S_i. \quad (1.44)$$

Критерием остановки может быть не только условие (1.44), но и другие статистическим критерии, например, достижение определённого уровня коэффициента детерминации R^2.

Достоинством рассмотренных выше методов является простота понимания и применения. Также в качестве достоинства можно выделить, то, что, используя значения остатков, можно легко увидеть наличие грубого отклонения. Однако недостатками является, то, что методы Эктона, Титьена-Мура-Бекмана и Прескотта-Лунда используются для нахождения только одного аномального наблюдения в линейной модели вида (1.2). Также для нахождения аномального значения необходимо производить большое количество расчетов равное объему выборки. В случае, если выборка имеет значительный объем, трудоемкость данных методов будет высока.

Основным достоинством метода уменьшения остаточной дисперсии является то, что он может применяться и для множественной линейной регрессии. Недостатками является то, что в методе уменьшения остаточной дисперсии используется при расчете вторая степень, что снижает точность. Также в данном критерии расчеты ведутся методом перебора, что увеличивает трудоемкость с увеличением объема выборки.

1.3.3.2.2 Метод Кука

Обнаруженные слабые места критериев Эктона, Титьена-Мура-Бекмана и Прескотта-Лунда привели к появлению принципиально новых методов, таких как метод Кука, методы Белсли-Ку-Уэлша (DFFITS) и Аткинсона. Однако методы Белсли-Ку-Уэлша и Аткинсона представляют собой незначительную модификацию метода Кука [66].

Расстояние Кука представляет собой меру влияния определённых наблюдений на построенную регрессию. Для нахождения данной статистики можно воспользоваться различными формулами.

Первый вариант расчета осуществляется по формуле:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{pS_e^2}, \qquad (1.45)$$

где $\hat{y}_j$ – ожидаемое значение регрессии (для j-го наблюдения), построенной по всей выборке;

$\hat{y}_{j(i)}$ – ожидаемое значение регрессии, построенной по выборке без i-го наблюдения;

p – число параметров модели (для линейной оно равно 2);

$S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$ – среднеквадратическая ошибка модели, полученная при использовании всех данных.

Следующее представление расстояния Кука имеет вид:

$$D_i = \frac{(\hat{y}_i - \hat{y}_{i(i)})}{pS_e^2 h_{ii}}, \qquad (1.46)$$

где h_{ii} – показатель влияния i-го наблюдения на коэффициенты модели. Представляет диагональные элементы матрицы проекции на пространство регрессоров $H=X(X^TX)^{-1}X^T$. Для парной линейной регрессии значение h_{ii} находится по формуле (1.47).

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (1.47)$$

Третья формула для статистики Кука:

$$D_i = \frac{e_i}{S_e\sqrt{(1-h_{ii})}} \times \frac{h_{ii}}{1-h_{ii}} \times \frac{1}{p}. \qquad (1.48)$$

Первая дробь в выражении (1.48) представляет собой стьюдентизированные остатки, которые являются внутренне стьюдентизированными.

При использовании расстояния Кука для исследования сразу нескольких значений, смысл состоит в поиске при отбрасывании уравнивающих измерений, которые стабилизируют параметры, нового регрессионного уравнения по отношению к исходному. Поэтому, даже при исключении множества невязок, эффект смещения оценок прогнозного значения, рассчитанного с использованием уравнения регрессии, будет минимальным. Благодаря этому статистика Кука считается достаточно эффективной при ручном анализе. Данный критерий даёт хорошие результаты, что позволяет применять его для повышения качества прогнозов по полученной в результате применения статистики Кука модели, поэтому данный метод будет использоваться в сравнительном анализе с предложенными методами.

Однако, он имеет ряд существенных недостатков:

1) отсутствие формального критерия для определения величины влияния;

2) при исключении пар, троек и т.к. значительно возрастает количество вычислений, которые для данного метода не являются простыми;

3) в случае, когда два или более влияющих наблюдения расположены близко друг к другу критерий Кука может быть не эффективен.

1.4 Постановка цели и задач исследования

Важной проблемой анализа данных является выявление ненадёжных измерений (выбросов) и их исследование. Для того, чтобы эффективность анализа и достоверность выводов, полученных в результате этого анализа, была максимальной, необходимо предварительная оценка исходных статистических данных на наличие в них аномальных измерений. Для этого существуют различные методы и критерии, основные из которых были проанализированы в данном разделе.

Большинство методов, основанных на поиске отклонения от среднего или предыдущего (последующего) значения не дают положительных результатов, в случае, если речь идёт о зависимости между несколькими переменными. Это связано с тем, что не учитывается наклон линии регрессии.

Основные критерии, которые используются для обнаружения выбросов в регрессии: Эктона, Титьна-Мура-Бекмана, Прескотта-Лунда, уменьшения остаточной дисперсии, метод Кука. Однако, методы Эктона, Титьена-Мура-Бекмана и Прескотта-Лунда используются для нахождения только одного аномального наблюдения в линейной модели и осуществляют поиск аномальных значений с помощью проверки конкретного значения на аномальность. Однако, при большом объёме выборки это значение выявить достаточно трудно, поэтому необходимо исследовать каждое значение, а это увеличивает трудоёмкость. Критерий Кука является эффективным при обнаружении аномалий, однако его основными недостатками является трудоёмкость и отсутствие формального критерия для определения влияющего аномального значения.

Таким образом, проанализировав основные методы нахождения выбросов, можно сделать вывод, что:

1) не существует единого оптимального метода для нахождения аномальных наблюдений, который позволял бы одновременно анализировать подозрительные значения по компоненте X и Y;

2) отсутствуют эффективные методики последующей корректировки данных, требующейся после обнаружения выбросов;

3) большинство существующих методов при большом объеме выборке являются достаточно трудоёмкими.

Поэтому актуальной и необходимой является задача совершенствования методов выявления выбросов и корректировки исходных данных.

В результате проведенного анализа сформулирована основная цель исследований – совершенствование методов предварительной обработки исходных статистических данных для повышения точности линейных регрессионных моделей при построении эффективных прогнозов.

Для достижения цели поставлены и решены следующие задачи:

1. Выполнить сравнительный анализ существующих методов обнаружения аномальных и ненадёжных измерений.

2. Предложить и обосновать усовершенствованные методы обработки статистических данных, эффективность которых не должна зависеть от объёма исходной выборки, а их применение – не приводить к ухудшению качественных характеристик модели.

3. Предложить критерии оценки эффективности рассматриваемых методов.

4. Дать рекомендации по выбору конкретного метода корректировки исходных статистических данных в зависимости от вида модели.

5. На основе усовершенствованных методов поиска и корректировки исходных данных выполнить построение алгоритмов и осуществить разработку программного комплекса.

6. Провести программное моделирование с использованием разработанного комплекса программ для оценки эффективности предложенных методов.

1.5 Выводы по разделу 1

1. Для эффективного построения качественной модели необходима предварительная обработка исходных данных, для определения в них аномальных измерений.

2. Проанализированы основные методы обнаружения аномальных наблюдений в статистических данных и выявлены существующие недостатки.

3. Показано, что важным инструментом анализа экспериментальных данных являестся регрессионный анализ, частным, но эффективным инструментом которого является линейная регрессия.

4. На основании выявлениых недостатков существующих методов и нерешенных задач были сформулированы цель и задачи диссертационного иследования.

РАЗДЕЛ 2

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ ВЫЯВЛЕНИЯ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ И ИХ ПОСЛЕДУЮЩЕЙ ОБРАБОТКИ НА ОСНОВЕ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

2.1 Метод повышения качества парной линейной регрессионной модели, основанный на отбрасывании данных

2.1.1 Используемая модель статистических данных

При исследовании влияния «засорений» на предполагаемую регрессионную модель используется функция типа Тьюки-Хьюбера [67]. Функция плотности при этом имеет вид:

$$f(y) = (1-\varepsilon)\cdot\varphi(y;m;\sigma_0^2) + \varepsilon\cdot\varphi(y;m;\sigma_1^2), \qquad (2.1)$$

где $\varphi(y;m;\sigma_0^2) = \frac{1}{\sigma_0\cdot\sqrt{2\pi}} e^{-\frac{(y-m)^2}{2\sigma_0^2}}$ – плотность вероятности нормального распределения со средним значением m и дисперсией σ_0^2;

$\varphi(y;m;\sigma_1^2)$ – плотность вероятности засорений;

ε – доля «засоряющих» аномальных наблюдений.

В данной модели, предполагается, что в выборочных значениях кроме «хороших», есть и «плохие» наблюдения. Т.е. с вероятностью 1- ε, близкой к 1 в выборке встречаются «хорошие» наблюдения, а с малой вероятностью ε появляются «плохие». Дисперсия σ_1^2 в несколько раз больше, чем дисперсия σ_0^2. Исходя из этого у «плохих» значений дисперсия значительно больше, чем у «хороших», поэтому они являются выбросами.

В данной работе модель (2.1) можно расширить, благодаря отступлению от нормального закона $\varphi(y;m;\sigma^2)$. Если плотность распределения вероятностей случайных величин y_i является убывающей при увеличении значения $|y_i - m|$, а

также отвечает свойствам симметрии относительно m, то её можно использовать в качестве $\varphi(y)$ (например, двустороннее экспоненциальное распределение (распределение Лапласа), Симпсона и др.).

2.1.2 Метод наименьших квадратов для оценки параметров линейного регрессионного уравнения

Предположим необходимо найти аппроксимирующую функцию $y = f(x, a, b)$ такую, чтобы в точках $x = x_i$ она принимала значения максимально приближенные к табличным. Таким образом, график функции должен проходить как можно ближе к экспериментальным данным. Для определения неизвестных параметров a и b используется метод наименьших квадратов.

Метод наименьших квадратов (МНК) – один из методов для получения оценок параметров уравнения регрессии. Он является самым популярным классическим методом. Данный метод предложили ещё в XIX в. трое ученых независимо друг от друга – А.М. Лежандр, К.Ф. Гаусс и Р.А. Эндрейн. Суть метода состоит в определении параметров регрессионной модели (для линейной – это a и b) таким образом, чтобы функция S (формула 1.3) стремилась к минимальному значению. Чтобы минимизировать квадрат разности нужно найти частные производные по коэффициентам a, b. Для линейной модели требуется, чтобы найденная линия регрессии была наилучшей среди других прямых.

Основным достоинством МНК является простота математических вычислений и практической реализации. Данный метод имеет аналитическое решение для нахождения неизвестных параметров уравнения регрессии. Для применения МНК необходимо иметь достаточное количество экспериментальных данных (рекомендуется не менее 10 значений). При соблюдении определённых предпосылок, найденные оценки коэффициентов регрессии будут обладать рядом оптимальных свойств.

2.1.3 Сущность предлагаемого метода, основанного на отбрасывании данных

Как было представлено в предыдущей главе, проблема обнаружения выбросов при построении регрессионной модели является важной. В связи с выявленными недостатками существующих методов нахождения аномальных данных, было принято решение разработать новые методы, которые в результате обнаружения и дальнейшей корректировки ненадёжных измерений приведут к повышению качества парных линейных регрессионных моделей.

Суть первого метода, предложенного в данной работе, заключается в том, что среди имеющихся наблюдений определяются те, которые не попадают в рассчитанную область. Эти наблюдения признаются ненадёжными или иными словами аномальными, т.е. теми, которые приводят к ухудшению качества прогноза и поэтому исключаются из дальнейшего исследования.

Для того, чтобы обнаружить ненадёжные измерения необходимо выполнить следующие действия:

1) учитывая все исходные статистические наблюдения, с использованием метода наименьших квадратов, находятся коэффициенты парного линейного регрессионного уравнения вида (1.2);

2) определяются расчётные значения $\hat{Y}_i$, путём подстановки исходных значений регрессоров X_i в полученное ранее уравнение;

3) находятся невязки e_i, как разница между исходными статистическими данными и полученными:

$$e_i = Y_i - \hat{Y}_i; \tag{2.2}$$

4) рассчитывается среднеквадратическое отклонение (СКО) по формуле:

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}; \tag{2.3}$$

5) после этого необходимо найти уравнение линии перпендикулярной линии, соответствующей исходному линейному уравнению, полученному в п.1. Для этого, вначале определяем коэффициент a':

$$a' = -\frac{1}{a}, \tag{2.4}$$

где a – коэффициент исходного регрессионного уравнения.

Затем находится коэффициент b':

$$b' = \frac{\sum_{i=1}^{n} Y_i}{n} - a' \cdot \frac{\sum_{i=1}^{n} X_i}{n}; \tag{2.5}$$

6) рассчитываются значения $\hat{Y}_i'$ по уравнению:

$$\hat{Y}_i' = a' X_i + b'; \tag{2.6}$$

7) определяются невязки e_i'. В данном случае они равны разности исходных значений Y_i и $\hat{Y}'$, рассчитанных по уравнению (2.6);

8) находится СКО невязок:

$$\sigma_e' = \sqrt{\frac{\sum_{i=1}^{n} e_i'^{\,2}}{n-2}}; \tag{2.7}$$

9) теперь строится область попадания надёжных данных, соответствующая определённой вероятности, со сторонами $2k \cdot \sigma_e$ и $2k \cdot \sigma'_e$, которая имеет вид прямоугольника. Для её построения необходимо определить коэффициент k (обычно $0{,}6 \le k < 3$), соответствующий вероятности попадания в заданную область.

Коэффициент k, находится из соотношения (2.8), которое представляет собой вероятность P_0, того, что значение попадёт в область надёжности:

$$P_0 = 2 \cdot \frac{1}{\sqrt{2\pi}} \int_0^k e^{-t^2/2} dt. \tag{2.8}$$

С использованием таблицы значений интеграла Лапласса [68] находится значение k, где $P_0 = 2\Phi(k)$. Например, если требуется определить попадания значений с вероятностью 0,9, т.е. $2\Phi(k) = 0{,}9$, то по таблице находим, что для $\Phi(k) = 0{,}45$, k будет равно 1,65. В таблице 2.1, представлены значения коэффициента k для наиболее часто используемых вероятностей.

Таблица 2.1 – Значения коэффициента k при различных вероятностях

Вероятность	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
Значения k	1,95	1.65	1.45	1.27	1.15	1.05	0.93	0.85	0.76	0.7

Таблица 2.1 соответствует вероятностям при одностороннем отбрасывании данных. Однако вероятность попадания в прямоугольную область надёжных данных, будет равна квадрату вероятностей, используемых при одностороннем поиске аномальных наблюдений, поэтому значения k будут отличаться от значений, которые бы использовались при отбрасывании данных только по переменной Y или по переменной X. Т.е., если необходимо проверить надёжность данных с вероятностью 0,9, то это будет соответствовать $2\Phi(k) = 0{,}949$ (т.е. корень из 0,9). Значения коэффициентов k для прямоугольной области попадания надёжных данных, представлены в таблице 2.2.

Таблица 2.2 – Значения коэффициента k при различных вероятностях для прямоугольной области

Вероятность	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.5
Значения k	1.9	1.75	1.6	1.5	1.4	1.3	1.2	1.05

Теперь, воспользовавшись таблицей 2.2 можно построить прямоугольную область попадания надёжных данных с определенной вероятностью (Рисунок 2.1). Для этого к каждому расчётному значению $\hat{Y}_i$ вначале прибавляется значение $k \cdot \sigma_e$, а затем вычитается. Аналогичные действия выполняются и со значением $k \cdot \sigma'_e$.

После выполнения описанных выше действий необходимо определить данные, которые не попадают в построенную область. Для этого каждое исходное значение сравнивается с граничным значением области надёжности. Те данные, которые не попадают в область надёжности, считаются ненадёжными и исключаются из выборки [69, 70].

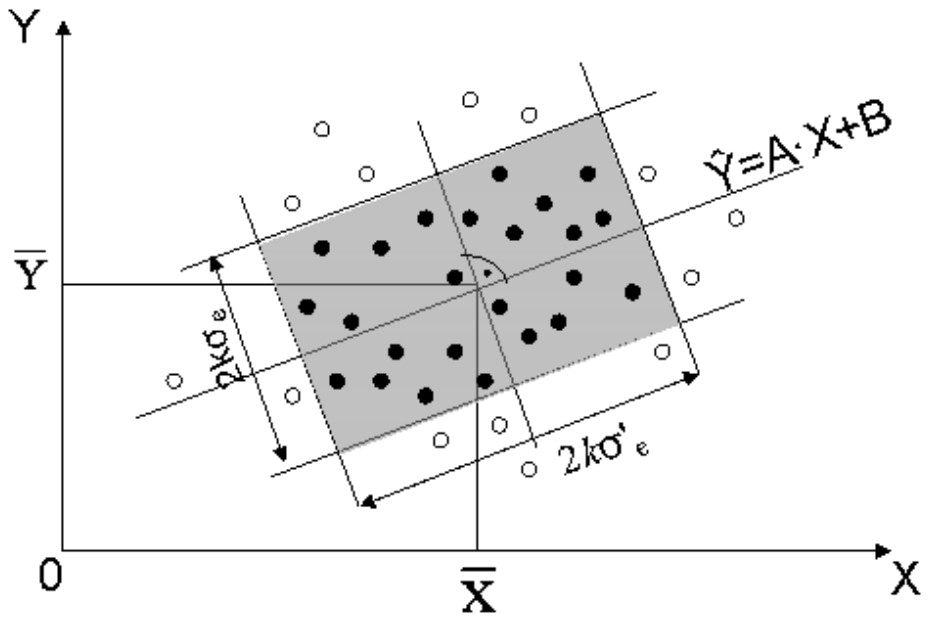


Рисунок 2.1 – Представление метода повышения точности регрессионной модели, основанного на отбрасывании «аномальных» и ненадежных данных

На рисунке 2.1 наглядно показано какие данные не попадают в заданную область. Эти данные являются аномальными наблюдениями и исключаются из выборки.

2.1.4 Критерии оценки адекватности модели и эффективности применения, предложенного метода

В качестве критериев оценки адекватности, полученной парной линейной регрессионной модели, чаще всего применяются следующие [71, 72, 73, 74]:

1) коэффициент детерминации R^2 – отражает долю дисперсии зависимой переменной, которая объясняется регрессией. Иными словами, он показывает часть значений регрессора X, которая полностью объясняет поведение случайных величин Y, построенным регрессионным уравнением. Он может принимать значения от 0 до 1. Чем сильнее зависимость между случайными величинами X и Y, тем ближе значение коэффициента детерминации к 1, т.е. построенная модель соответствует данным. Коэффициент детерминации R^2 можно определить по формуле (2.9) или (2.10):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \quad (2.9)$$

где $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ – математическое ожидание случайной величины Y;

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}. \quad (2.10)$$

2) F-критерий Фишера для оценки значимости уравнения регрессии. Он заключается в проверке гипотезы H_0 о статистической незначимости уравнения регрессии. Для этого необходимо сравнить критическое, т.е. табличное значений F-критерия Фишера $F_{табл}$ и фактическое $F_{факт}$. В случае простой линейной регрессии он рассчитывается по формуле (2.11) или (2.12):

$$F = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}(n-2), \quad (2.11)$$

$$F = \frac{R^2}{1-R^2}(n-2). \quad (2.12)$$

Чем выше значение F-критерия, тем уравнение лучше. Если фактическое значение $F_{факт}$ превышает табличное $F_{табл}$, то гипотеза H_0 о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость, т.е. модель считается адекватной.

3) величина остаточной дисперсии S^2, равна квадрату среднеквадратического отклонения, найденного по формуле (2.13). Чем её значение меньше, тем лучше подобранная модель.

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}. \quad (2.13)$$

4) средняя ошибка аппроксимации – определяется как среднее относительное отклонение расчетных значений от фактических. Если она

составляет менее 10%, то качество, подобранной модели высокое. Рассчитывается по формуле:

$$\overline{E} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \cdot 100\% . \qquad (2.14)$$

Поскольку при использовании, предложенного метода обнаружения и отбрасывания аномальных наблюдений, модели (исходная и полученная после отбрасывания данных) сравниваются между собой, то для лучшего анализа, используются следующие частные критерии эффективности [75, 76, 77, 78, 79]:

– коэффициент детерминации R^2;

– величина остаточной дисперсии S^2;

– модуль величины смещения результата прогноза, которое возникает в результате корректировки исходного уравнения и изменения вида нового уравнения после отбрасывания части статистики:

$$\Delta_{прогн}=|(a\cdot X_{прогн} + b) – (a_н\cdot X_{прогн} + b_н)|, \qquad (2.15)$$

где первое и второе слагаемое, соответственно, линейное регрессионное уравнение до отбрасывания части статистики и линейное регрессионное уравнение после отбрасывания части статистики (a, b и $a_н$, $b_н$ находятся с использованием традиционного метода наименьших квадратов);

– доверительный интервал прогнозных значений $Y_{прогн}$ – геометрическое место расположения прогнозных значений $Y_{прогн}$ при заданном значении $X_{прогн}$ и заданной доверительной вероятности $P_{дов}$;

– время, необходимое для выполнения вычислений (вычислительная сложность), т.е. количество элементарных операций ЭВМ (сложение, умножение и т.д.), необходимое для нахождения и отбрасывания аномальных и ненадежных измерений;

– точность. Это важный критерий, поскольку при использовании предложенного метода, необходимо решить до какого размера следует уменьшить прямоугольник, т.к. из-за неограниченного сужения сторон прямоугольника,

можно получить абсурдный научный результат. Точность рассчитывается по формуле:

$$T = R^2 \cdot \frac{m}{n}, \tag{2.16}$$

где m – количество данных, оставшихся после отбрасывания.

Рекомендуется выбирать ту модель, в которой коэффициент детерминации R^2 будет максимальным, но при условии, что значение T будет больше 0,5.

В данной работе применялись доверительные интервалы, которые свободны от закона распределения случайных величин невязок [80], поскольку использование квантилей Стьюдента, не будет давать корректных результатов. Это связано с тем, что используемая статистика невязок будет иметь распределение не соответствующее нормальному, а отбрасывание аномальных наблюдений ещё более это усугубит.

Суть заключается в следующем. Вначале все невязки выстраиваются по возрастанию от e_1 до e_i. Поскольку доверительный интервал (e_1, e_i) симметричен относительно медианного значения e_k, то доверительная вероятность определяется по формулам:

$$P_{дов} = I_{1/2}(1,i) - I_{1/2}(i,1), \tag{2.17}$$

$$P_{дов} = 1 - 2 \cdot I_{1/2}(b,a), \tag{2.18}$$

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1} \cdot (1-t)^{b-1} dt, \tag{2.19}$$

$$B(a,b) = \int_0^1 t^{a-1} \cdot (1-t)^{b-1} dt \text{ [81]}, \tag{2.20}$$

где $I_x(a,b)$ − неполная Бета-функция;

B(a, b) – Бета-функция.

После этого находятся пары 1, i; 2, i-1; 3, i-2 и т.д., которые будут представлять собой размах доверительного интервала и обеспечивают заданное значение $P_{дов}$, в соответствии с формулами (2.17-2.20).

Для определения вычислительной сложности необходимо учесть все производимые расчеты. Обозначим с помощью n исходное количество данных, а с помощью m − количество наблюдений, оставшихся после отбрасывания. Тогда:

1) определение значения коэффициента детерминации R^2 для исходного уравнения содержит следующие действия:

а) расчет среднего значения $\overline{Y}$ содержит n операций (n-1 для сложения всех значений Y_i и одно деление на общее количество точек);

б) определение значений $\hat{Y}_i$ требует 2n вычислений (каждое значение умножается на коэффициент a, для этого требуется n вычислений, затем к каждому произведению прибавляется коэффициент регрессии b, что также требует n операций);

в) за n вычислений, находятся разности рассчитанных значений $\hat{Y}_i$ и среднего значения $\overline{Y}$;

г) после этого каждая найденная разность возводится в квадрат (n операций);

д) n-1 элементарная операция необходима для определения суммы всех найденных в предыдущем подпункте значений;

е) далее необходимо определить знаменатель дроби. Для этого требуется то же самое количество элементарных операций, что и в подпунктах в)-г), т.е. 3n-1;

ж) числитель делится на знаменатель, что требует 1 операцию.

Итого для расчета значения коэффициента детерминации R^2 требуется: 9n – 1 элементарная операция.

2) Необходимо рассчитать значение среднеквадратического отклонения σ_e по формуле (2.3):

а) находятся все невязки e_i за n операций;

б) далее за n вычислений все значения невязок возводятся в квадрат;

в) после этого рассчитывается значение суммы квадратов невязок n-1 операцию;

г) одна операция требуется для вычитания из общего количества значений числа 2;

д) числитель делится на знаменатель. Это требует всего одной операции;

е) находится корень из полученного в предыдущем подпункте значения (одна операция).

Таким образом, определение среднеквадратического отклонения требует 3n+2 операции.

3) расчет коэффициентов уравнение перпендикулярной линии, требует:

а) одна операция для определения коэффициента a';

б) поскольку среднее значение $\overline{Y}$ уже найдено, требуется рассчитать среднее значение $\overline{X}$. Это действие состоит из n операций;

в) после этого коэффициент a' умножается на $\overline{X}$ (1 операция);

г) из $\overline{Y}$ вычитается значение, полученное в предыдущем пункте (1 операция).

Итого расчет коэффициентов требует n + 3 операции.

4) после этого находится значение σ_e' по формуле (2.7):

а) определяется значение $\hat{Y}'$ за 2n операций, по аналогии с подпунктом б пункта 1;

б) далее аналогично пункту 2, находится само значение σ_e' за 3n + 2 операции.

Таким образом, для данного пункта требуется 5n + 2 вычисления.

5) определение значений $k \cdot \sigma_e$ и $k \cdot \sigma'_e$ для построения области надёжности. Для этого необходимо выполнить всего 2 операции;

6) после этого к каждому значению $\hat{Y}_i$ прибавляется, а затем и отнимается, полученное значение $k \cdot \sigma_e$. Для этого требуется 2n число операций;

7) аналогичные действия производятся и для значения $k \cdot \sigma'_e$, что требует 2n вычислений;

8) далее, каждое исходное значение сравнивается с полученным в пунктах 6 и 7. Это требует 4n операции;

9) расчет нового значения R^2. Для его определения, по аналогии с пунктом 1, требуется 9m-1 вычисление.

Таким образом, общая вычислительная сложность метода для обнаружения и устранения аномальных наблюдений составляет:

$$K = 26 \cdot n + 9 \cdot m + 7. \qquad (2.21)$$

Если же при определении количества элементарных операций не учитывать расчет значения коэффициентов детерминации R^2, то получится:

$$K = 17 \cdot n + 9. \quad (2.22)$$

Схематически предложенный метод можно представить следующим образом (Рисунок 2.2).

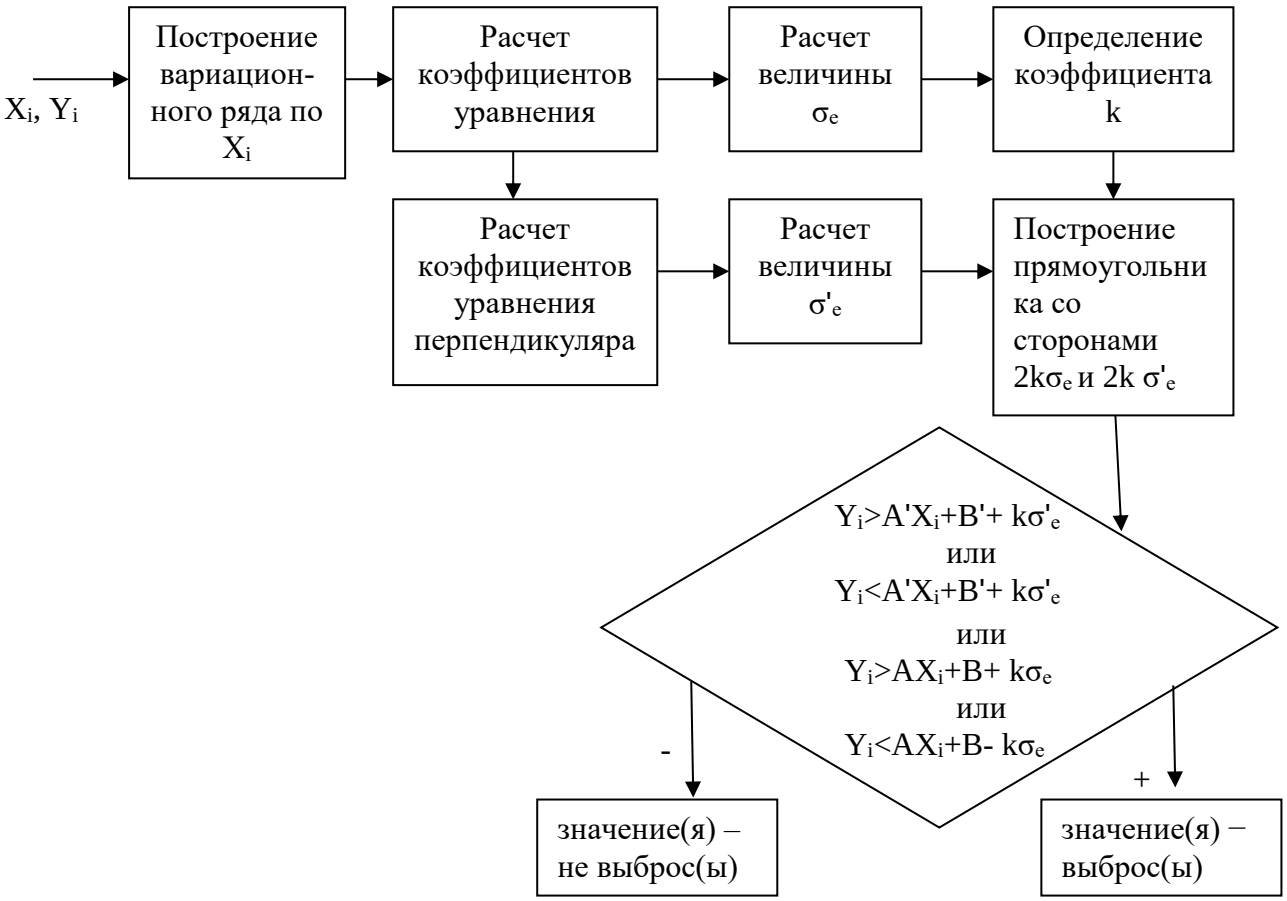


Рисунок 2.2 – Схематическое представление метода, основанного на отбрасывании данных

2.1.5 Модификации метода отбрасывания данных

В некоторых практических ситуациях достаточным будет применить одно из упрощений предложенного метода обнаружения и устранения аномальных данных.

Первое упрощение следует использовать в случае, если исследователь уверен в исходных значениях регрессора X (например, значения X_i известны заранее и в них не может быть случайных ошибок). Также данное упрощение применимо при поиске аномальных значений в данных, описываемых многомерной линейной регрессионной моделью. Суть данной модификации заключается в том, что поиск и отбрасывание ненадёжных данных, осуществляются только по зависимой переменной Y. Для этого находится область, в пределах которой, расположены надёжные наблюдения, представляющая собой не прямоугольник, а «коридор», границы которого равноудалены от линии исходного линейного уравнения на расстояние $k\sigma_e$. Таким образом, общее расстояние между границами «коридора» будет $2k\sigma_e$. Данное упрощение метода отбрасывания ненадёжных данных можно представить в виде рисунка 2.3.

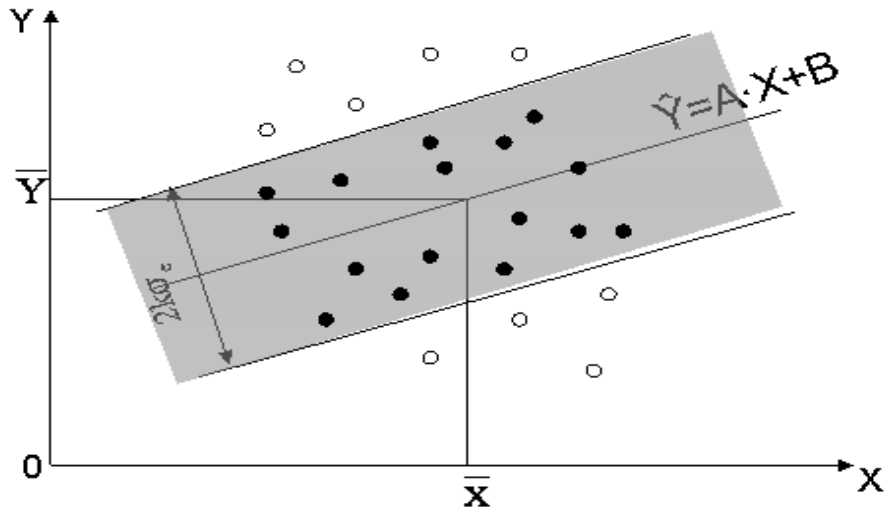


Рисунок 2.3 – Первая модификация метода обнаружения и отбрасывания аномальных данных

Вычислительная сложность в этом случае будет значительно меньше, чем у самого предложенного метода и с учетом определения коэффициента детерминации R^2 составит:

$$K = 16 \cdot n + 9 \cdot m + 1. \quad (2.23)$$

Если же при определении количества элементарных операций не учитывается R^2, имеем:

$$K = 7 \cdot n + 3. \quad (2.24)$$

В ходе исследования было выявлено, что на сегодняшний день существует метод аналогичный первой модификации предложенного метода. Это метод построения доверительного интервала для индивидуальных значений зависимой переменной y_i, который определяется по формуле (2.25):

$$\hat{y}_i - t_{1-\alpha;n-2} S_{\hat{y}_i} \le y_i \le \hat{y}_i + t_{1-\alpha;n-2} S_{\hat{y}_i}, \quad (2.25)$$

где t – распределение Стьюдента;

1-α – доверительная вероятность с которой значение, по уравнению регрессии будет находиться в доверительном интервале;

$S_{\hat{y}_i}$ – оценка дисперсии индивидуальных значений y_i рассчитывается согласно формуле (2.20)

$$S_{\hat{y}_i} = \sigma_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}. \quad (2.26)$$

Преимуществами первой модификации предложенного метода отбрасывания данных по сравнению с существующей методикой, описанной выше являются:

1) возможность использования для многомерных линейных регрессионных моделей;

2) нет привязки к конкретному количеству данных, т.е. значение коэффициента k не изменяется в зависимости от объема выборки, а значение t в существующем методе меняется при изменении количества данных;

3) простота. Предложенная методика не требует дополнительных расчетов и таблиц;

4) меньшая вычислительная сложность. Для предложенного упрощения метода, требуется 7·n +3 элементарных операций, а для существующего метода построения доверительного интервала дополнительно к 7·n +3 вычислениям, требуется:

а) определение среднего значения $\overline{X}$, которое состоит из n операций;

б) вычисление разности исходных значений X_i и среднего $\overline{X}$ за n операций;

в) возведение каждой разности в квадрат – n вычислений;

г) определение суммы всех рассчитанных ранее разностей – n-1 вычисление;

д) одна операция для деления числителя на знаменатель;

е) одна операция для деления 1 на количество данных;

ж) две операции для добавления 1 и 1/n;

и) одна заключительная операция для получения произведения σ_e на полученное в предыдущих действиях значение.

Таким образом, вычислительная сложность существующей методики, основанной на построении доверительного интервала требует 11n + 7 элементарных операций, что больше на 4n, чем в предложенной первой модификации метода обнаружения и устранения аномальных наблюдений. При этом разница будет увеличиваться при возрастании объёма экспериментальных данных.

Рассмотрим для наглядности конкретный пример. Пусть дана выборка из 22 наблюдений (Таблица 2.3):

Таблица 2.3 – Исходные данные для примера

X_i	52,46	52,95	53,71	54,08	58,99	59,74	62,98	63,69	66,8	68,44	69,33
Y_i	32,13	40,15	37,08	29,21	39,64	34,5	39,77	42,62	41,61	40,97	42,69
X_i	70,77	70,84	71,66	72,3	72,81	74,39	79,38	82,03	85,58	89,18	90,5
Y_i	48,09	44,13	38,06	44,28	42,96	47,32	45,82	52,37	53,09	53,87	50,79

Уравнение регрессии в данном случае имеет вид: $\widehat{Y}_1 = 0{,}51X_i + 7{,}53$. Подставляя соответствующие значения регрессора X в полученное уравнение, определим все значения $\widehat{Y}_1$ (Таблица 2.4).

Таблица 2.4 – Расчетные значения $\widehat{Y}_1$

$\widehat{Y}_1$	34,28	34,53	34,92	35,11	37,61	38,00	39,65	40,01	41,60	42,43	42,89
$\widehat{Y}_1$	43,62	43,66	44,08	44,40	44,66	45,47	48,01	49,37	51,18	53,01	53,69

После этого определим значения невязок и возведём каждое, полученное значение в квадрат (Таблица 2.5).

Таблица 2.5– Квадраты невязок e_i

e_i^2	4,64	31,53	4,66	34,82	4,1	12,23	0,01	6,80	0,0001	2,14	0,04
e_i^2	19,96	0,22	36,20	0,02	2,9	3,43	4,81	9,03	3,66	0,74	8,38

Далее уже можно определить значение среднеквадратического отклонения σ_e. Оно будет составлять 3,085. На следующем шаге для вероятности P=0,85, выбираем соответствующее значение k из таблицы 2.1, которое составляет 1,45. После этого произведение k и σ_e прибавляются к расчетным значениям $\hat{Y}_i$, а потом вычитаются из этих же значений.

Для определения доверительного интервала для индивидуальных значений зависимой переменной Y_i, также определяется значение σ_e. Следующим шагом будет определение квадратов разностей исходных значений регрессора X_i и среднего $\overline{X}$ (Таблица 2.6).

Таблица 2.6 – Разности исходных значений регрессора X_i и среднего $\overline{X}$

$(X_i-\overline{X})^2$	280,55	264,37	240,24	228,9	104,44	89,67	38,81	30,47	5,81	0,59	0,01
$(X_i-\overline{X})^2$	2,44	2,66	6,00	9,55	12,96	26,84	103,4	164,36	267,99	398,82	453,28

Далее можно определить значение $S_{\hat{y}_i}$ по формуле (2.20). По таблице распределения Стьюдента для вероятности равной 0,85 и числом степеней свободы 20, определяется соответствующее значение, которое составляет 1,497. К значениям $\widehat{Y}_1$ прибавляется, а затем и вычитается соответствующее значение. Таким образом, получаются значения, которые представляют собой границы доверительного интервала в пределах которого находятся надёжные измерения для заданной вероятности.

В результате работы обоих методов, получаются практически одинаковые границы надёжности. Это можно увидеть при графическом отображении, описанных выше методов (Рисунок 2.4). Данные, которые оказались за пределами,

построенных областей, признаются аномальными или ненадёжными, т.е. выбросами. Для рассматриваемого примера, при вероятности равной 0,85 выбросами являются 3 значения.

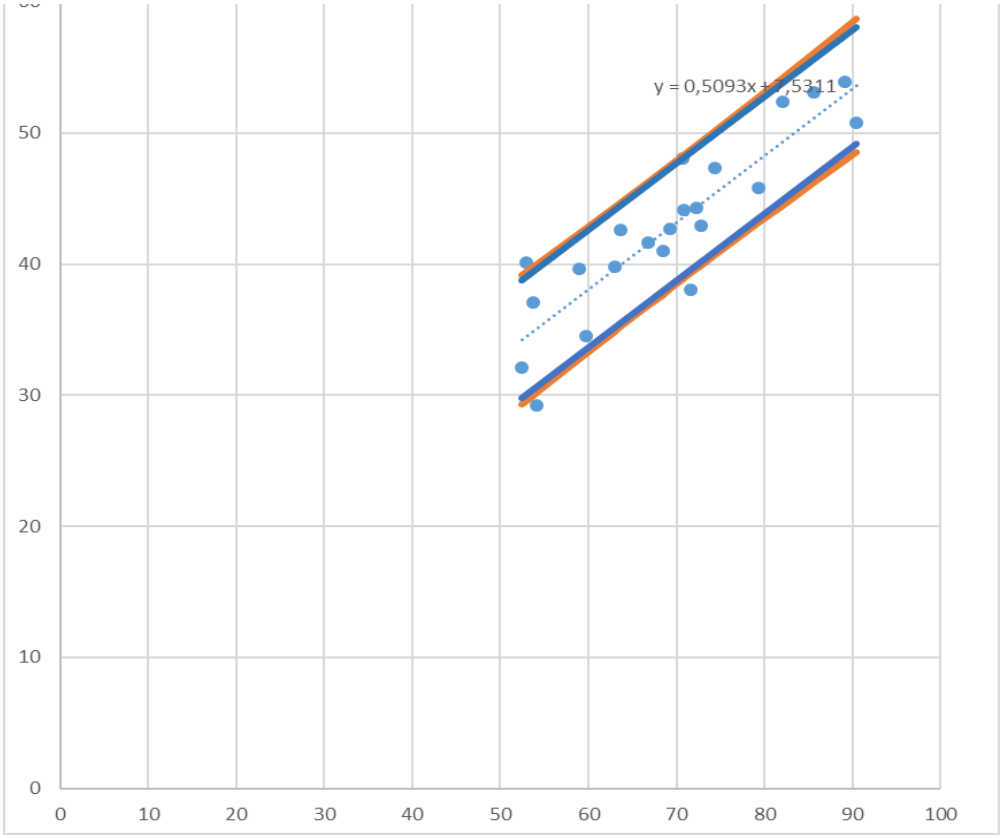


Рисунок 2.4 – Графическое представление рассматриваемых методов

На рисунке более жирными линиями отражены границы попадания надёжных данных с вероятностью 0,85 в заданную область, по упрощенной модификации, предложенного в данной диссертационной работе метода обнаружения и устранения аномальных и ненадёжных измерений. Более тонкой линией представлены пределы доверительного интервала для индивидуальных значений зависимой переменной Y_i. Как видно из рисунка, они практически совпадают. Однако, как было сказано выше и показано на практическом примере, представленный в данной работе метод имеет ряд достоинств, по сравнению с существующим, поэтому его можно рекомендовать для решения практических задач.

Схематически представить предложенную модификацию можно в виде рисунка 2.5.

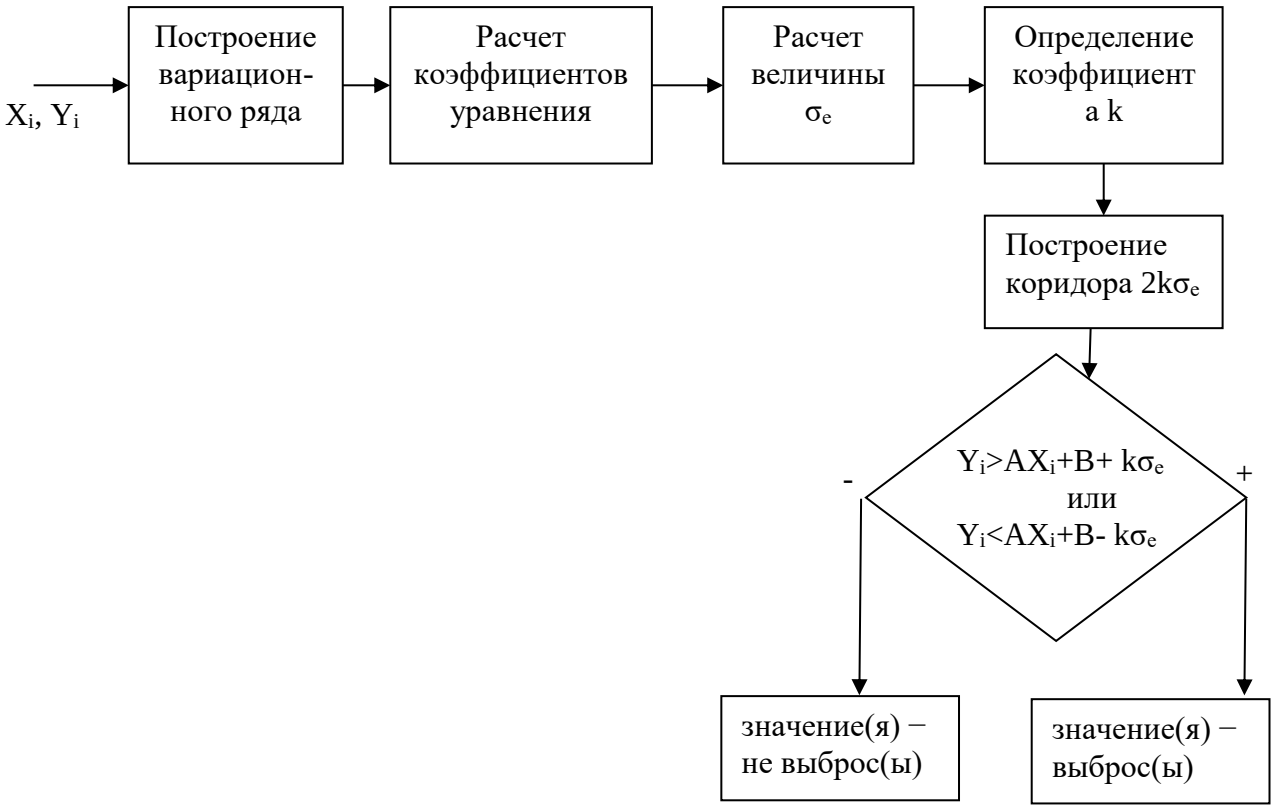


Рисунок 2.5 – Схематическое изображение первой модификации метода

В отличие от первой модификации метода повышения качества линейной регрессионной модели за счет устранения ненадёжных данных, во второй модификации находится «коридор», границы, которого параллельны линии, построенной перпендикулярно линии, полученной по исходному уравнению регрессии. Расстояние между данными границами составляет $2k\sigma'_e$ (Рисунок 2.6).

Данные, которые выходят за границы рассматриваемого «коридора» исключаются из дальнейшего рассмотрения.

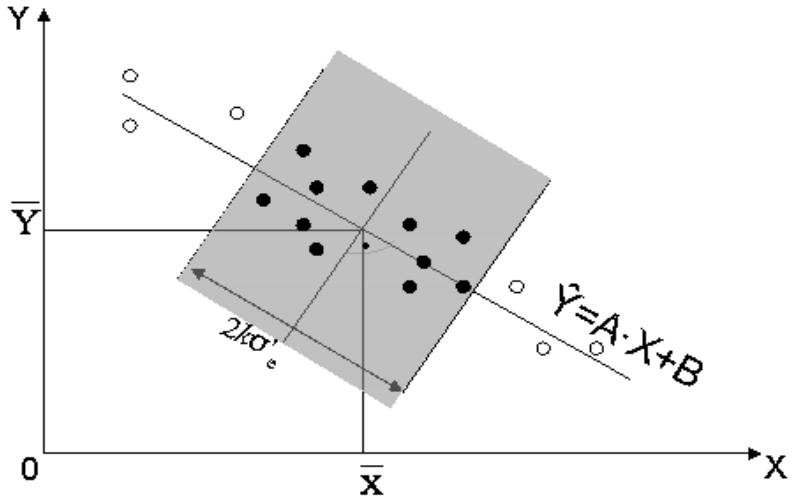


Рисунок 2.6 – Вторая модификация метода повышения качества модели, основанного на отбрасывании данных

Данную модификацию рекомендуется использовать в случаях, когда в исходных данных регрессора X существует достаточно большой размах между крайними его значениями и они удалены на значительное расстояние от остальных точек. В этом случае исследователю следует обратить особое внимание на эти значения, поскольку они могут оказаться выбросами. После исключения подозрительных данных из выборки величина коэффициента детерминации R^2 может даже уменьшиться. Это связано с тем, что крайние значения регрессора X являются определяющими для уравнения регрессии и оказывают на него большое влияние, поэтому при их устранении из выборки оно может значительно измениться. Но при этом, новое уравнение, полученное после исключения аномальных наблюдений из выборки, будет более точным и с его помощью можно будет построить более надёжные прогнозы. При применении для обработки статистических наблюдений данного (второго) упрощения предложенного метода, следует устранять не более 10-15% исходных значений, поскольку большее количество, может свидетельствовать о том, что данные имеют определенную закономерность и не будут являться аномальными.

Количество элементарных операций для второй модификации, предложенного метода повышения качества регрессионных моделей, с учётом

вычисления коэффициента детерминации, находится по формуле (2.27), а без учёта – согласно формуле (2.28).

$$K = 19 \cdot n + 9 \cdot m + 4; \quad (2.27)$$

$$K = 10 \cdot n + 6. \quad (2.28)$$

Алгоритм второй модификацию метода можно схематически изобразить, как показано на рисунке 2.7.

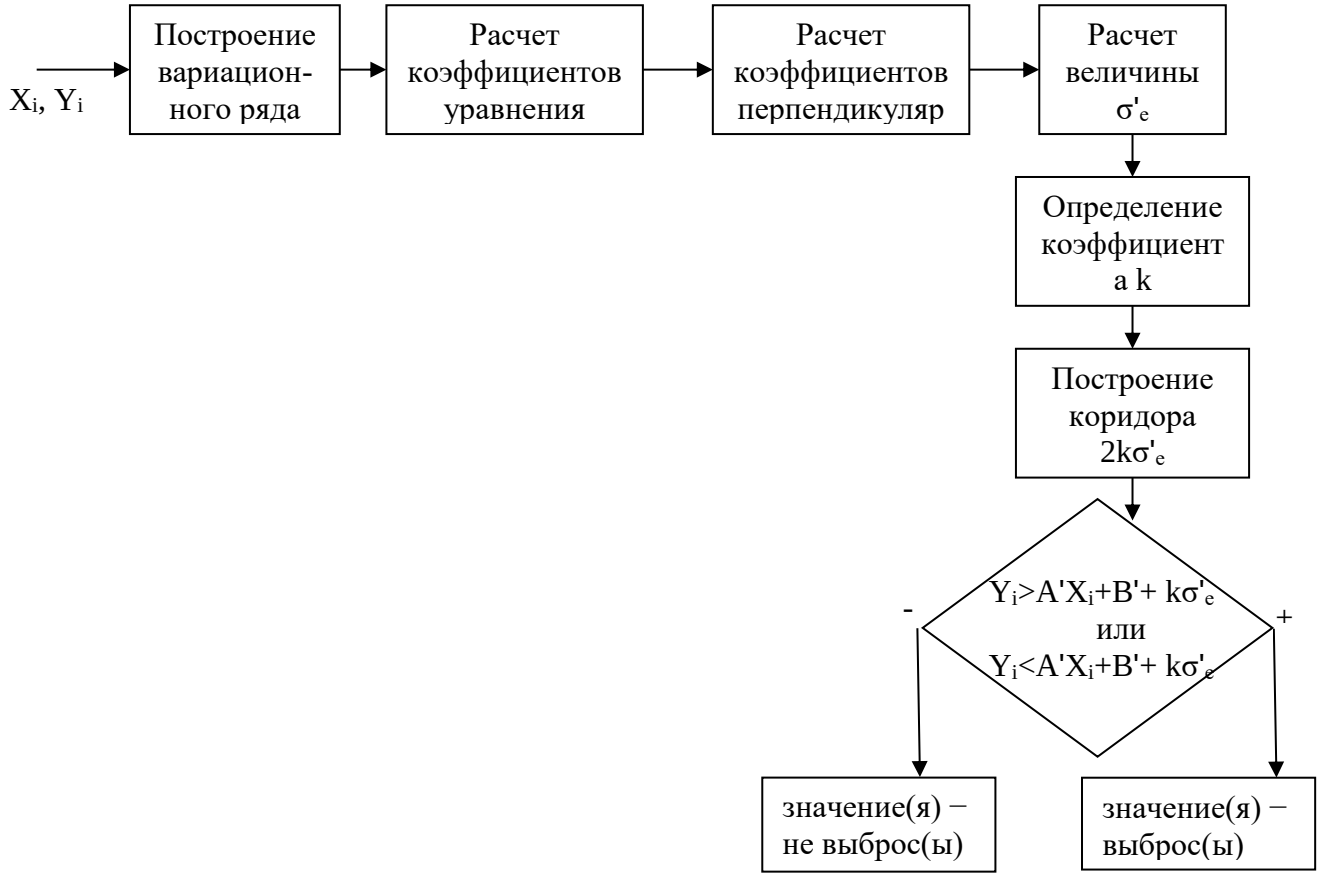


Рисунок 2.7 – Схематическое представление алгоритма второй модификации метода

Реализация описанных выше модификаций проще и быстрее, чем реализация предложенного метода повышения качества линейной регрессионной модели, основанного на обнаружении и устранении ненадёжных статистических данных, однако и первая, и вторая модификации имеют эффективность несколько ниже, по

сравнению с полученной от применения самого метода. В некоторых случаях применение полного метода отбрасывания данных по величине R^2 дает результаты хуже, чем первая его модификация, однако в случае использования метода исключаются аномальные значения и по независимой переменной X и по зависимой переменной Y одновременно, что даёт более точные результаты. Когда количество исходных данных большое (более 100) и рядом с крайними точками по X есть другие значения, коэффициент детерминации R^2 будет выше.

2.2 Метод повышения качества парной линейной регрессионной модели, основанный на переносе данных

2.2.1 Сущность метода повышения качества модели, основанного на переносе ненадёжных данных

Идея второго метода улучшения точности регрессионной модели состоит в том, что на первом этапе, также, как и в предыдущем, предложенном в данной работе методе, основанном на отбрасывании данных, находятся границы области попадания надёжных данных. Те данные, которые не попадают в найденную при заданной вероятности область, в отличие, от первого метода, не исключаются из выборки, а переносятся на границы данной области, т.е. изменяют свои значения (Рисунок 2.8) [82, 83, 84, 85].

При непопадании части измерений в область, определенную границами, которые параллельны линии исходного регрессионного уравнения, эти данные переносятся на уровень $A \cdot X_i + B \pm k \cdot \sigma_e$. При этом значения независимой переменной X_i остаются неизменными, а величины зависимой переменной Y_i меняются на соответствующие граничные значения.

При переносе найденных ненадёжных наблюдений на границы области, которые параллельны линии, являющейся перпендикулярной к линии регрессии, полученной по соответствующему уравнению, значения зависимой переменной Y_i

не требуют изменений, а по независимой переменной X_i корректируются в соответствии с заданной областью.

Таким образом, для определения новых значений X'_i, при перемещении на границы области, полученной как $A'\cdot X_i + B' + k\cdot\sigma'_e$, используется формула:

$$X_i^{'} = \frac{(A \cdot X_i + B) - k \cdot \sigma' - B'}{A'}. \qquad (2.29)$$

Новые значения X'_i при переносе на уровень $A'\cdot X_i + B' - k\cdot\sigma'_e$ находятся по формуле:

$$X_i^{'} = \frac{(A \cdot X_i + B) + k \cdot \sigma' - B'}{A'}. \qquad (2.30)$$

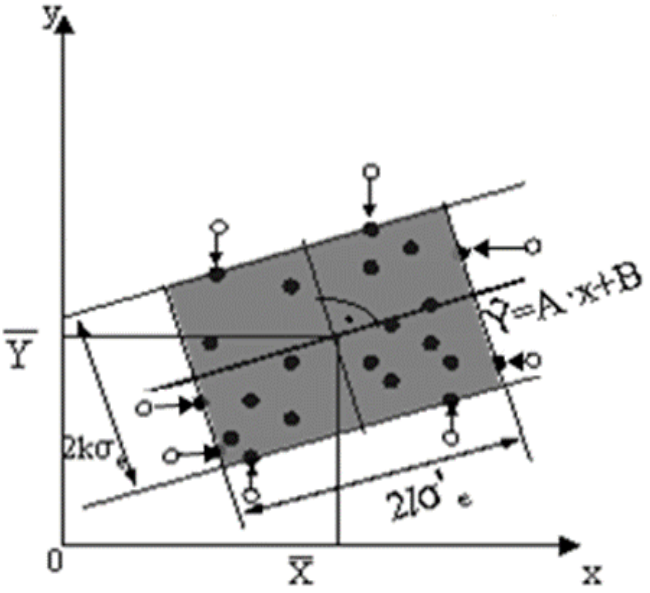


Рисунок 2.8 – Метод повышения точности регрессионной модели, основанный на переносе «аномальных» и ненадежных данных

Вычислительная сложность данного метода содержит, согласно формуле, следующее количество элементарных операций:

$$K = 35\cdot k + m' + 6\cdot l + 7, \qquad (2.31)$$

где m' − количество перенесенных точек на уровень $A\cdot x_i + B \pm k\cdot\sigma_e$;

l − количество перенесенных данных на уровень $A'\cdot x_i + B' \pm k\cdot\sigma'_e$.

Если же при определении вычислительной сложности не учитывать нахождение значений R^2, то она составит:

$$K = 17 \cdot n + m' + 10 \cdot l + 9. \quad (2.32)$$

Таким образом, как видно из формулы (2.32) количество элементарных операций, требующихся для реализации метода повышения качества регрессионной модели, основанного на корректировке данных больше, чем в методе, основанном на отбрасывании данных.

Реализацию алгоритма метода, повышения качества линейной регрессионной модели, основанного на переносе ненадёжных данных на границы области надёжности, можно изобразить в виде рисунка 2.9.

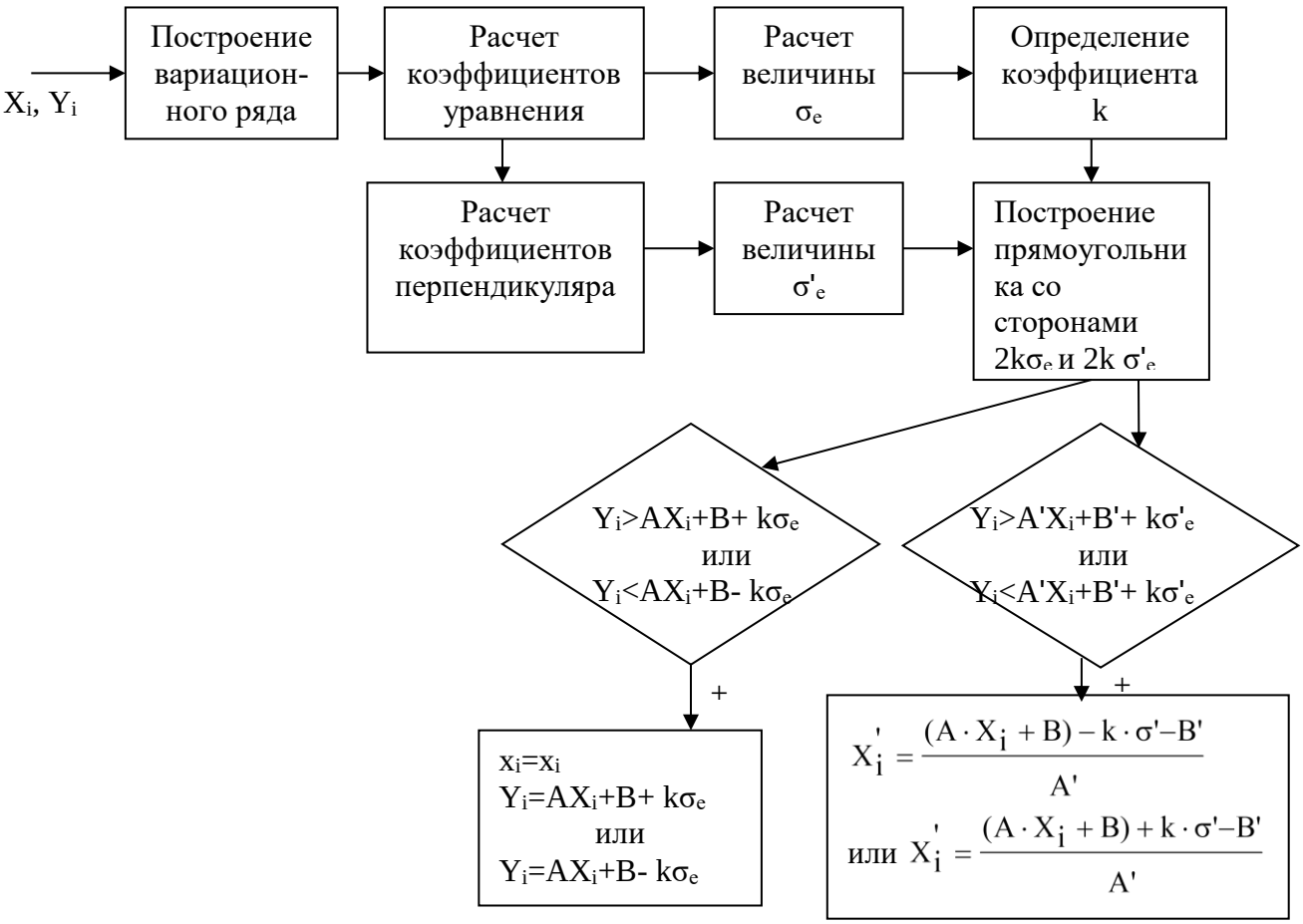


Рисунок 2.9 – Схематическое представление метода повышения качества модели, основанного на переносе данных

Для данного метода, аналогично методу повышения качества линейных регрессионных моделей, за счет отбрасывания ненадёжных наблюдений, могут быть предложены два упрощения (Рисунки 2.10, 2.11).

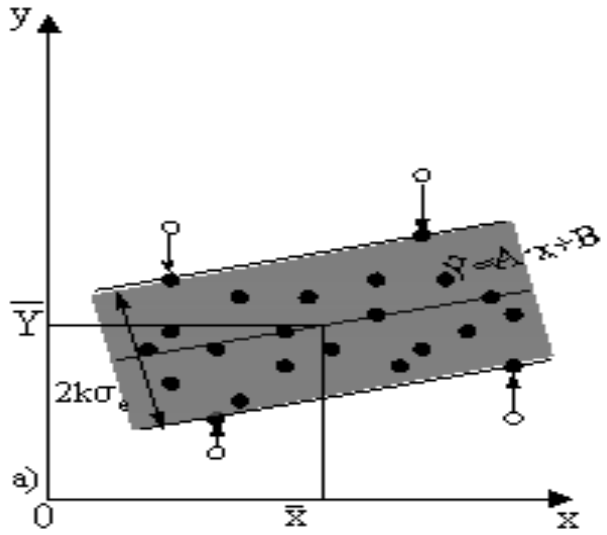


Рисунок 2.10 – Первая модификация метода, основанного на переносе данных

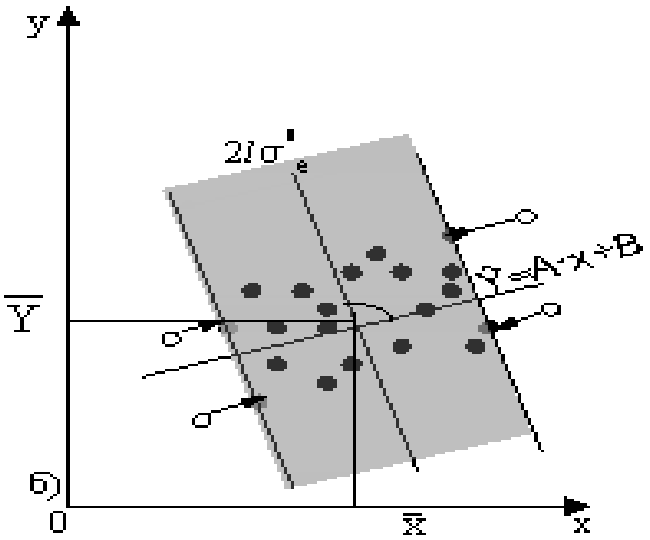


Рисунок 2.11 – Вторая модификация метода, основанного на переносе данных

Количество элементарных операций для первой модификации второго метода, с учётом расчета коэффициента детерминации, необходимых для реализации первой модификации, находится по формуле (2.33), а без учета – по формуле (2.34):

$$K = 25 \cdot n + m' + 1, \qquad (2.33)$$

$$K = 7 \cdot n + m' + 3. \quad (2.34)$$

Количество элементарных операций необходимых для реализации второй модификации, находятся аналогично:

$$K = 28 \cdot n + 6 \cdot l + 4, \quad (2.35)$$

$$K = 10 \cdot n + 6 \cdot l + 6. \quad (2.36)$$

Таким образом, вычислительная сложность модификаций метода, значительно ниже вычислительной сложности самого метода.

2.2.2 Статистические особенности метода улучшения точности модели, основанного на переносе данных

Метод повышения качества линейных регрессионных моделей, основанный на переносе данных, имеет определённые особенности, поскольку является нелинейным по отношению к исходным данным по которым построено регрессионное уравнение $\hat{Y} = A \cdot x + B$.

В результате переноса наблюдений, которые оказались ненадёжными, на уровни $\hat{Y} \pm u$, где u – граница области надёжности, происходит преобразование исходной нормальной плотности вероятностей ω(x') в новую плотность вероятностей невязок e_i :

$$\omega'(x') = \begin{cases} \omega(x'), если\ \ \hat{Y} - u < x' < \hat{Y} + u, \\ S_1 \cdot \delta(-x'), если\ \ x' = \hat{Y} + u, \\ S_2 \cdot \delta(x'), если\ \ x' = \hat{Y} - u. \end{cases} \quad (2.37)$$

где $\delta()$ − дельта-функция Дирака, которая определяется заданием правил интегрирования её произведений с непрерывными функциями. Эта функция сингулярная, т.е. она равна нулю везде, кроме точки x=0, где она обращается в бесконечность таким образом, чтобы её интеграл, содержащий точку x=0 был равен 1.

$S_1 = S_2 = \int_{u+\hat{Y}}^{\infty} \omega(x')dx'$, т.к. границы ± u симметричны относительно исходного уравнения $\hat{Y}$.

Преобразование исходного закона ω(x') в ω'(x') может быть осуществлено, если невязки e_i исходного регрессионного уравнения $\hat{Y}$ пропустить через двухсторонний безынерционный ограничитель с линейным участком [86]:

$$y = f(x) = \begin{cases} -u, & -\infty < x < -u, \\ x, & -u < x < u, \\ u, & u < x < \infty. \end{cases} \qquad (2.38)$$

При осуществлении нелинейного преобразования (2.38) исходные данные, которые находятся между уровнями ±u, не претерпевают изменений, а другие − привязываются к уровню +u или –u. Такое нелинейное преобразование изменяет спектральные (корреляционные) характеристики исходных данных, а данные, которые не претерпели изменений, характеризуются прежними вероятностными характеристиками.

Графически особенности данного метода, можно представить в виде рисунка 2.12.

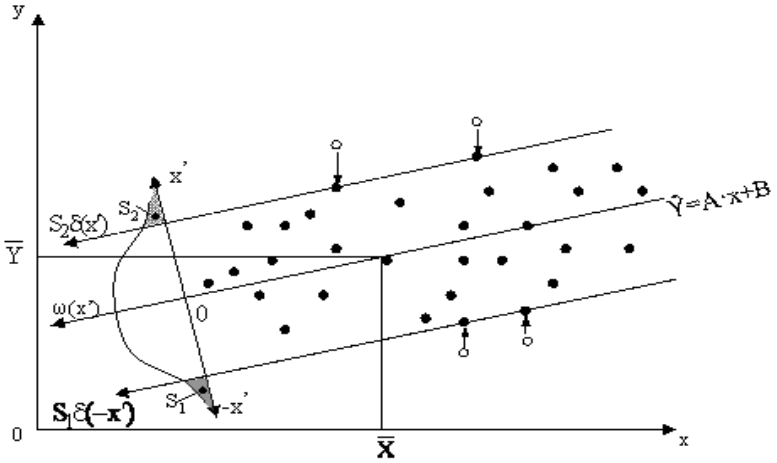


Рисунок 2.12 – Статистические особенности метода, основанного на переносе данных

Совершенно другими статистическими характеристиками описывается метод улучшения качества прогнозных регрессионных моделей, основанный на отбрасывании данных. Здесь осуществляется линейная операция отбрасывания аномальных и ненадежных исходных данных (Рисунок 2.13).

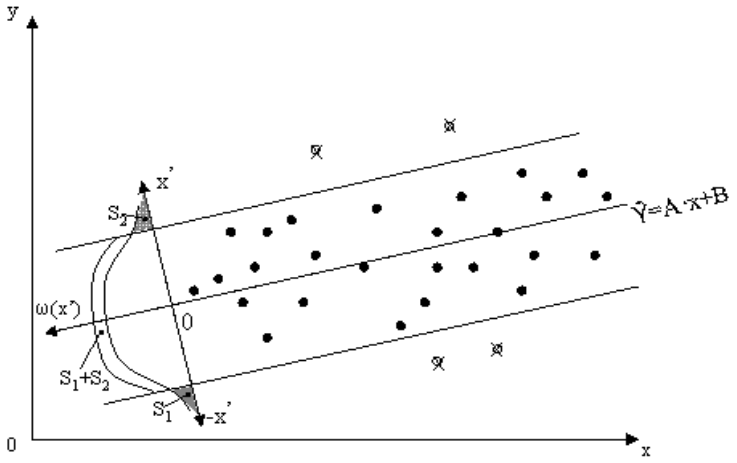


Рисунок 2.13 – Статистические характеристики метода, основанного на отбрасывании данных

Как видно из рисунка 2.13, при реализации отбрасывания аномальных и ненадежных данных происходит увеличение вероятности качественных исходных данных. При этом спектральные (корреляционные) характеристики в результате такого преобразования меняются в меньшей степени.

2.3 Парные нелинейные регрессионные зависимости

При решении практических задач очень часто встречаются парные нелинейные зависимости. Например, в экономике это могут быть:

– кривые Энгеля, которые отражают зависимость спроса (y) на товар от доходов населения (x) и представляются степенной функцией вида $Y = aX^b + \varepsilon$;

– кривые Филипса, которые описывают зависимость между нормативами безработицы (X) и процентом прироста заработной платы (Y). Модель представляет собой функцию гиперболы вида $Y = a/X + b + \varepsilon$;

– модель зависимости величины валового национального продукта (Y) от денежной массы (X) представляется функцией вида $Y = a + b*\ln(X) + \varepsilon$;

– модель изменения переменной y с постоянным темпом прироста во времени имеет вид $Y = ae^{bX} + \varepsilon$.

В общем случае парная нелинейная регрессия делится на 2 класса:

1) нелинейные по объясняющим переменным, но линейные по оцениваемым параметрам;

2) нелинейные по оцениваемым параметрам.

К первому классу относятся полиномы различных степеней и функция гиперболы. Параметры моделей данного класса определяются с использованием метода наименьших квадратов.

Для гиперболы вида $Y = a + b/X + \varepsilon$ при замене 1/X на z получаем линейное уравнение регрессии $Y = a + bz + \varepsilon$.

Второй класс делится на:

а) нелинейные модели внутренне линейные;

б) нелинейные модели внутренне нелинейные.

Внутренне линейные модели можно привести к линейному виду путём определённых преобразований (как правило, логарифмированием). К моделям данного типа относятся степенная, экспоненциальная, показательная и другие функции.

Модели с внутренней нелинейностью нельзя привести к линейному виду, в этом случае для оценки параметров применяются итеративные методы. Примером такой модели является модель вида: $Y = a + bX^c + \varepsilon$.

Для примера преобразования функции из нелинейного вида к линейному возьмём степенную функцию вида $Y = aX^b\varepsilon$. Сначала её необходимо прологарифмировать, тогда: $\ln Y = \ln a + b\ln X + \ln \varepsilon$. Затем произведём соответствующие замены: $y1 = \ln Y$, $a1 = \ln a$, $x1 = \ln X$, $\varepsilon 1 = \ln \varepsilon$. Тогда уравнение примет вид: $y1 = a1 + bx1 + \varepsilon 1$.

Аналогичным образом преобразуется функция вида $Y = ab^X\varepsilon$, а функция вида $Y = ae^{bX}\varepsilon$ после логарифмирования приобретает вид: $\ln Y = \ln a + b X + \ln \varepsilon$.

Коэффициент корреляции в случае нелинейных функций находится по формуле 2.39:

$$R = \sqrt{1 - \frac{\sum(Y_i - \widehat{Y_i})^2}{\sum(Y_i - \bar{Y})^2}}. \quad (2.39)$$

Однако, при расчете разностей и определении сумм, указанных в формуле необходимо использовать исходные нелинейные уравнения. Значения R для линейного и нелинейного уравнения регрессии будут одинаковы, только при условии, что преобразование уравнения не касалось зависимой переменной y (например, равносторонняя гипербола).

Таким образом, как показано выше большинство парных нелинейных моделей можно привести к линейному виду с помощью преобразований и производить все необходимые расчеты. Возможность приведения большинства нелинейных моделей к линейному виду, является ещё одним преимуществом линейных моделей. Поэтому, предложенный в данной работе подход для поиска аномалий и методы дальнейшей обработки данных применимы и для нелинейных моделей с внутренней линейностью.

2.4 Применение предложенного подхода для решения проблемы «квартета Энскомба»

«Квартет Энскомба» – четыре набора данных, которые подобрал английский математик Ф. Дж. Энскомб в 1973 году для того, чтобы показать важность применения графиков для статистического анализа и влияния аномальных измерений на свойства всего набора данных. Эти данные состоят из четырёх пар, основные статистические свойства которых идентичны: среднее значения, дисперсия, коэффициент корреляции, коэффициент детерминации (таблица 2.7). Модель линейной регрессии для всех вариантов описывается одинаковым уравнением. Значения данных каждой из четырёх пар «квартета Энскомба» представлены в таблице 2.8. Однако, при графическом отображении этих наборов

данных можно увидеть их существенные отличия. Графики изображены на рисунке 2.14.

Таблица 2.7 – Основные характеристики данных «квартета Энскомба»

Характеристика	**Значение**
Среднее значение x	9
Дисперсия x	10
Среднее значение *y*	7,5
Дисперсия *y*	3,75
Корреляция между x и y	0,82
Прямая линейной регрессии	$y = 3 + 0{,}5\,x$
Коэффициент детерминации линейной регрессии	0,67

Таблица 2.8 – Значения данных «квартета Энскомба»

I		**II**		**III**		**IV**	
x	y	x	y	x	y	x	y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

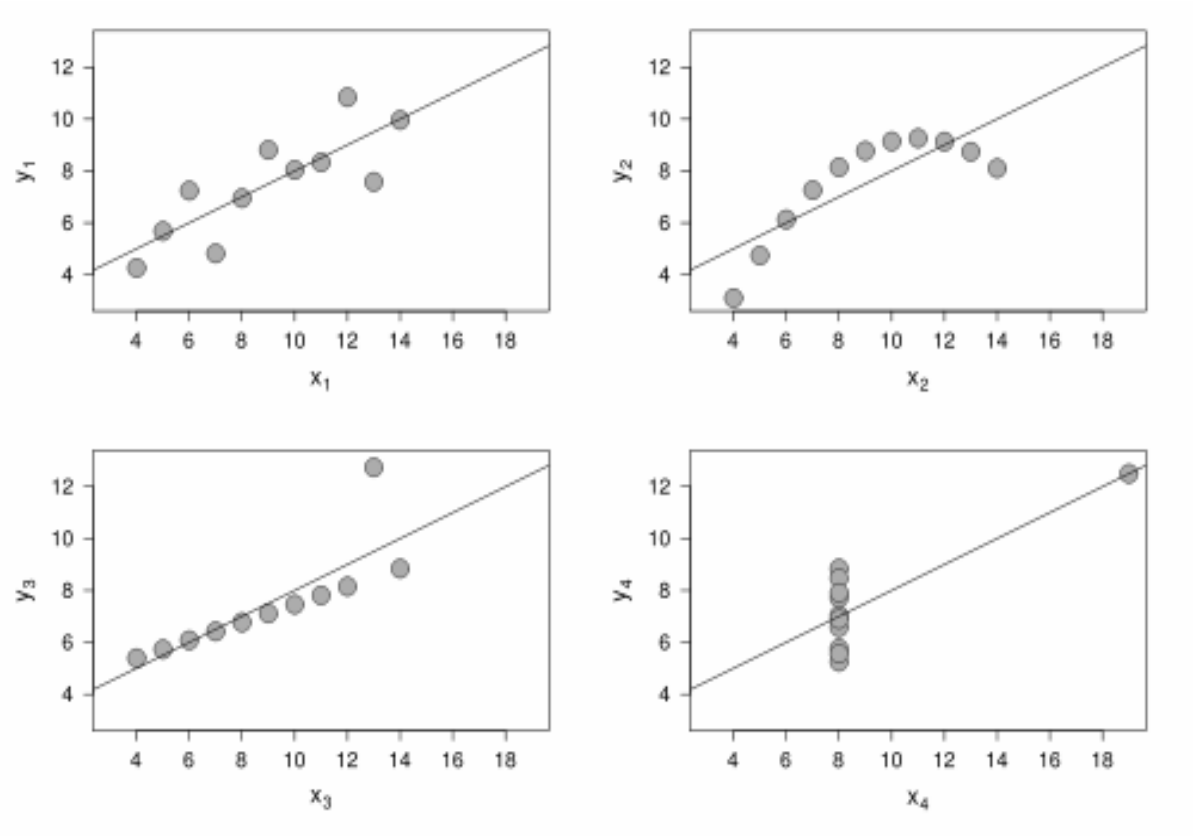


Рисунок 2.14 – График, изображающий данные «квартета Энскомба»

Как видно из рисунка, при одинаковых основных статистических свойствах, графики существенно отличаются. Однако, при большом объёме данных не всегда есть возможность их графического отображения. Воспользуемся предложенным в данной работе подходом для обнаружения и последующего отбрасывания аномальных измерений.

Для первого набора данных достаточно применить первую модификацию метода, поскольку среди значений независимой переменной X нет резко выделяющихся данных. Отбрасывание всего одного значения зависимой переменной Y=7,58 при значении X=13, позволяет повысить значение коэффициента детерминации на 11% и достичь величины 0,78.

Второй набор данных не подходит для применения предложенного метода, поскольку в нём ошибочно сделано исходное предположение о виде модели.

При использовании предложенного метода для третьего набора данных, достигается значение коэффициента детерминации практически равное 1 при

отбрасывании всего одного значения равного 12,74 при независимой переменной X=13. Этот же результат можно увидеть и на рисунке 2.14, поскольку там есть одно резко выделяющееся значение.

Четвёртый набор данных представляет наибольший интерес, с точки зрения результативности применения метода. Применение первой модификации подхода, как и большинство существующих на сегодняшний день методов, не дало бы положительных результатов, поскольку значения являющееся выбросом лежит непосредственно на линии. Однако, использование предложенного в данной работе метода позволило выявить аномальное значение. Поскольку, данное значение было значительно удалено от остальных и являлось определяющим для уравнения регрессии, то его удаление значительно изменило все основные параметры. Так коэффициент детерминации вместо 0,67 стал 0,09, тем самым давая исследователю понять, что исходное предположение о виде модели было ошибочным и требуется провести дополнительный анализ набора данных

Таким образом, как видно из полученных результатов, применение предложенного в работе метода даёт положительный эффект для трёх из четырёх наборов данных «квартета Энскомба». При этом для получения результатов не требуется дополнительного графического отображения исходных экспериментальных данных.

2.5 Пример использования первой модификации подхода для поиска аномальных данных при многомерной линейной регрессии

Для анализа эффективности применения первой модификации метода для поиска выбросов при многомерной линейной регрессии была взята выборка, представленная в таблице 2.9.

Таблица 2.9 – Исходные данные для многомерной регрессии

y	x1	x2	x3	x4	x5	x6
19	13,7	400	1,09	82	74,8	33,5
26,2	-0,8	710	1,01	66	100	32,8
18,1	9,6	1610	0,4	80	69,7	33,4
15,4	40	500	0,93	74	100	27,8
29	8,41	640	0,92	65	74	27,9
21,6	3,5	920	0,59	64	73,1	33,2
21,9	3	1890	0,63	82	52,3	30,8
18,9	7,1	3040	0,49	85	49,6	32,4
21,1	13	2730	0,71	78	71,2	29,2
23,8	10,7	1850	0,93	74	70,6	28,7
40,5	-16,2	2920	0,51	69	64,2	25,1
21,6	6,6	1070	0,8	85	58,3	35,9
25,4	21,9	160	0,74	69	100	31,4
19,7	17,8	380	0,44	83	72	30,1
38	-11,8	1140	0,81	54	100	34,1
30,1	7,5	690	1,05	65	100	30,5
24,8	3,7	1170	0,73	76	69,5	30
30,3	1,6	1280	0,65	67	81	32,4
19,5	8,4	2270	0,48	85	39,1	28,7
15,6	2,7	960	0,72	84	58,4	33,4
17,2	5,6	1710	0,62	84	42,4	29,9
18,4	12,7	1410	0,84	86	36,4	23,3
27,3	-4,8	200	0,73	66	99,8	27,5
19,2	16,5	960	0,45	74	90,6	29,5
16,8	15,2	11500	1	87	5,9	25,4
13,2	11,6	1380	0,63	85	44,2	28,8
29,7	4,9	530	0,54	70	100	33,1
19,8	1,1	370	0,98	75	52,6	30,8
27,7	3,8	440	0,46	48	100	28,4
20,5	19	1630	0,68	83	72,1	30,4

Уравнение регрессии для представленных выше данных имеет следующий вид: $Y = 31{,}266 – 0{,}39x_1+0{,}00075x_2+1{,}23x_3-0{,}083x_4+0{,}165x_5-0{,}419x_6$.

Коэффициент R^2 при этом равен 0,78.

Применяя первую модификацию предложенного подхода для поиска и отбрасывания аномальных данных, было получено, что при отбрасывании всего 20% данных (7 испытаний) значение коэффициента детерминации R^2 выросло до 0,91, что свидетельствует о повышении качества модели и применимости предложенного подхода.

2.6 Выводы по разделу 2

В данном разделе были предложены новые методы повышения качества линейных регрессионных моделей. Суть первого метода состоит в том, что среди статистических данных находятся те, которые не попадают в рассчитанную для заданной вероятности прямоугольную область надёжности. Эти данные признаются выбросами и исключаются из выборки, после чего находится новое регрессионное уравнение по оставшимся данным.

Второй предложенный метод заключается в том, что данные, не попавшие в область надёжности, не исключаются, а корректируются (переносятся) так, чтобы попасть на границы данной области.

Для каждого из перечисленных методов были рассмотрены по два упрощения, которые применяются в соответствующих случаях:

– если данные по независимой переменной X не вызывают сомнений у исследователя и отклонения между соседними значениями равномерны, то достаточным является применение первой модификации;

– если в исходных данных регрессора X наблюдается значительное отклонение крайних значений от остальных – вторая модификация.

Были отобраны критерии оценки эффективности, рассматриваемых методов:

– коэффициент детерминации R^2;

– величина остаточной дисперсии S^2;

– модуль величины смещения результата прогноза;

– доверительный интервал прогнозных значений $Y_{прогн}$;

– точность;

– вычислительная сложность.

Основные достоинства предложенных в данном разделе методов повышения качества линейных регрессионных моделей:

1) применение методов позволяет увеличивать значение коэффициента детерминации R^2. Это достигается за счет выявления и дальнейшей корректировки (устранения или преобразования) «аномальных» и ненадежных измерений;

2) т.к. невязки e_i не подчиняются нормальному закону распределения, использовались доверительные интервалы свободные от закона распределения;

3) данные методы рекомендуются для применения в пакетах программ уже существующих и новых вычислительных систем, а также в автоматизированных информационных системах различного назначения, поскольку они легко формализуемы и имеют небольшую вычислительную сложность;

4) предлагаемые методы можно применять для повышения качества нелинейных регрессионных прогнозных моделей с внутренней линейностью. При этом исходное нелинейное регрессионное уравнение путем специальных преобразований приводится к линейному виду. Производится отбрасывание части статистики и далее путем обратного преобразования возвращаются к исходному нелинейному уравнению;

5) первая модификация предложенного подхода подходит для поиска и устранения аномальных данных не только при парных линейных регрессионных моделях, но и при многомерных.

РАЗДЕЛ 3
РАЗРАБОТКА ПРОГРАММНЫХ МОДУЛЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОБНАРУЖЕНИЯ И ОБРАБОТКИ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ

3.1 Разработка программного приложения для методов повышения качества парных линейных регрессионных моделей

Всё многообразие прикладного программного обеспечения можно разделить на два больших класса: «расчетчики» и «анализаторы» [87].

Расчетные – задачи, в которых входные и выходные данные являются числовыми. Решение таких задач находится с использованием математических методов.

Боьшинство расчетных задач имеют следующую последовательность решения: ввод исходных данных, выполнение расчетов, вывод результатов. Часто их решение сводится к последовательному использованию нескольких алгоритмов.

Одной из важных задач является обработка данных. Решение такого рода задач осуществляется, в основном, специализированными (статистическими) пакетами прикладных программ. Однако, большинство этих приложений отличается высоким уровнем сложности, дороговизной, а также отсутствием достаточного внимания на анализ и устранение аномальных измерений, отсутствием механизма встраивания в состав других программных приложений в виде отдельного программного модуля.

Поскольку, предложенные во втором разделе методы анализа и обработки экспериментальных данных являются новыми и не реализованы в существующих статистических и математических программных пакетах общего назначения, целесообразной является разработка оригинального комплекса программ, позволяющего быстро и точно осуществлять поиск аномальных измерений, их корректировку, нахождение наилучшей регрессионной прогнозной модели и вывод результатов для последующего анализа.

Для разработки комплекса необходимо было решить следующие задачи:

1) формирование требований к программному приложению;
2) выбор среды разработки;
3) разработка архитектуры системы;
4) составление алгоритма работы программного приложения;
5) написание программного кода.

При разработке программного приложения были учтены следующие требования:

1) должно быть максимально доступным и понятным для пользователя;
2) приложение должно иметь небольшой размер;
3) приложение должно содержать полную реализацию математического аппарата, используемого для работы предложенных методов;
4) должна иметься возможность ввода произвольного количества исходных статистических данных;
5) должна быть реализована возможность графического представления данных и найденных областей надёжности;
6) результаты работы программного комплекса, должны быть наглядны и способствовать принятию верного решения при выборе конечной регрессионной модели, на основании которой, будут строиться дальнейшие прогнозы;
7) должна иметься возможность просмотра промежуточных результатов расчетов, для проведения, при необходимости, дополнительного анализа.

Программа должна выполнять следующие основные функции: ввод данных вручную или с помощью их экспорта из файла формата Microft Excel; предобработка данных; поиск аномальных измерений; отбрасывание или корректировка найденных аномальных значений; расчет коэффициентов эффективности; выбор наиболее точной модели; вывод результатов.

При вводе данных комплекс должен предусматривать изменение структуры данных, т.е. учитывать заданное пользователем количество наблюдений. Таблица с данными должна иметь следующий формат: первая строка содержит название переменных (X в первом столбце и Y во втором), последующие строки – исходные статистические данные. Помимо этого, в приложении необходимо организовать

контроль ошибок. Например, если статистические данные будут введены с ошибкой, то приложение должно сигнализировать об этом пользователю.

Таким образом, разрабатываемое программное приложение должно позволять человеку без дополнительной подготовки, использовать его для анализа и выдавать результаты, которые позволят максимально быстро и верно принять решение.

В результате был получен комплекс программ, состоящий из главного интерфейса, написанного на языке C# в среде Microsoft Visual Studio 2017 и макросов, разработанных с применением языка программирования Visual Basic for Applications для Microsoft Excel, поскольку при необходимости автоматизировать обработку данных в MS Excel, данный язык является наиболее удобным.

Выбор языка программирования C# связан с тем, что он объектно-ориентированный, относительно простой и достаточно популярный. C# является гибридом из нескольких языков программирования, таких как Java, C++, Visual Basic и т.д. Поэтому C# объединяет в себе лучшие идеи этих языков [88, 89].

VBA – это язык программирования, встроенный во множество отдельных программ и прикладных пакетов – от приложений Microsoft Office и до таких мощных пакетов, как AutoCAD, CorelDrow и Adobe Creative Suite, не говоря уже о многочисленных специализированных приложениях, предназначенных для управления производственными процессами, учета финансовых ресурсов или информационной поддержке клиентов [90].

3.2 Архитектура и общий алгоритм работы программного комплекса методов поиска и обработки аномальных данных

Интерфейс разрабатываемого программного комплекса должен быть дружественным, интуитивно понятным, реализовывать преимущества графического и табличного отображения.

Комплекс программ состоит из следующих модулей:

1) программный модуль для обнаружения и удаления аномальных данных;

2) программный модуль для обнаружения и корректировки выбросов;
3) программный модуль для графического отображения обнаруженных аномальных данных;
4) модули для реализации модификаций методов.

Модули содержат восемь логических блоков (Рисунок 3.1):

1) управление;
2) ввод исходных данных;
3) контроль введённых данных;
4) сортировка данных;
5) расчет всех необходимых показателей, согласно выбранного алгоритма;
6) вывод результатов;
7) графическое отображение;
8) сохранение данных.

Рисунок 3.1 – Логические блоки приложения

Блок «Управление» – блок с помощью которого пользователь осуществляет работу с приложением. Здесь пользователь производит запуск других блоков, корректировку введённых данных, в случае обнаружения в них ошибок при вводе.

Блок «Ввод исходных данных» позволяет пользователю вводить данные вручную или путём вставки исходных данных из стороннего файла формата Microsoft Excel. После ввода данных в блоке «Управление» пользователем инициируется работа приложения с использованием соответствующей кнопки, первым при этом вызывается блок «Контроль введённых данных». Он необходим для осуществления проверки корректности введённых измерений в блоке «Ввод исходных данных», остановки работы приложения и подачи сигнала в виде сообщения блоку «Управление», в случае обнаружения ошибки в вводимых данных. После получения сообщения об ошибке в введённых данных, пользователь может произвести необходимые исправления данных и вновь запустить приложение.

В блоке «Сортировка исходных данных» происходит сортировка измерений по возрастанию по независимой переменной X. После сортировки в соответствующих блоках автоматически происходит расчет всех необходимых показателей и вывод полученных результатов. Расчет числовых характеристик включает в себя вычисление коэффициентов уравнения регрессии, исходного коэффициента детерминации, математического ожидания, дисперсии, среднеквадратического отклонения, нахождение области надёжности и ряда других статистических величин, необходимых в дальнейшем для подбора наилучшей модели и вывода окончательных результатов.

Далее в блоке «Управление» пользователь может инициировать работу блока «Графическое отображения». Данный блок предназначен для графического представления оставшихся после отбрасывания или скорректированных данных, линий регрессии, построенных по исходным и изменённым данным, рассчитанных на предыдущих этапах областей надёжности.

Блок «Сохранение данных» позволяет при необходимости сохранить полученные результаты. Все промежуточные и итоговые результаты сохраняются

с использование соответствующей кнопки главного меню. Для удобства пользователя сохранение происходит в файле формата Microsoft Excel. Это позволяет, в случае необходимости, провести более детальный анализ.

Результатом работы приложения является вывод таблицы с показателями эффективности (значения коэффициентов детерминации R^2, величины доверительных интервалов, величины смещений, исходное количество наблюдений и после отбрасывания, точность) и возможными вариантами наилучшей модели.

Разработанный программный комплекс можно представить в виде UML диаграммы активности (деятельности) (Рисунок 3.2). Диаграмма активности отображает последовательность выполнения определённых действий, происходящих в системе. Одним из преимуществ диаграммы деятельности является то, что она наглядно описывает переходы от одной деятельности к другой. Графически она представляет собой ориентированный граф, в котором вершины – это действия или деятельности, а дуги – переходы между ними. В данном случае, на диаграмме представлены два объекта – пользователь и приложение. Пользователем осуществляется запуск программного комплекса, ввод исходных данных и выбор последующих действий в главном меню. При этом пользователю доступны такие пункты меню, как «Данные» и «Методы». В пункте меню «Данные» можно выбрать подпункты «Ввести вручную», «Из Excel файла», «Сохранить в Excel», а в меню «Методы» – «Метод отбрасывания» и «Метод корректировки», где можно выбрать либо сам метод, либо одну из его модификаций. Помимо этого, пользователь может перейти в файл формата .xlsx и вызвать автоматизированное построение графиков, после чего сохранить изменения в файле стандартными средствами Microsoft Excel. В приложении осуществляется проверка правильности введённых пользователем данных и в случае ввода данных не соответствующего формата, выдаёт сообщение об ошибке. Далее в приложении осуществляются все необходтмые расчёты и формируется таблица результатов.

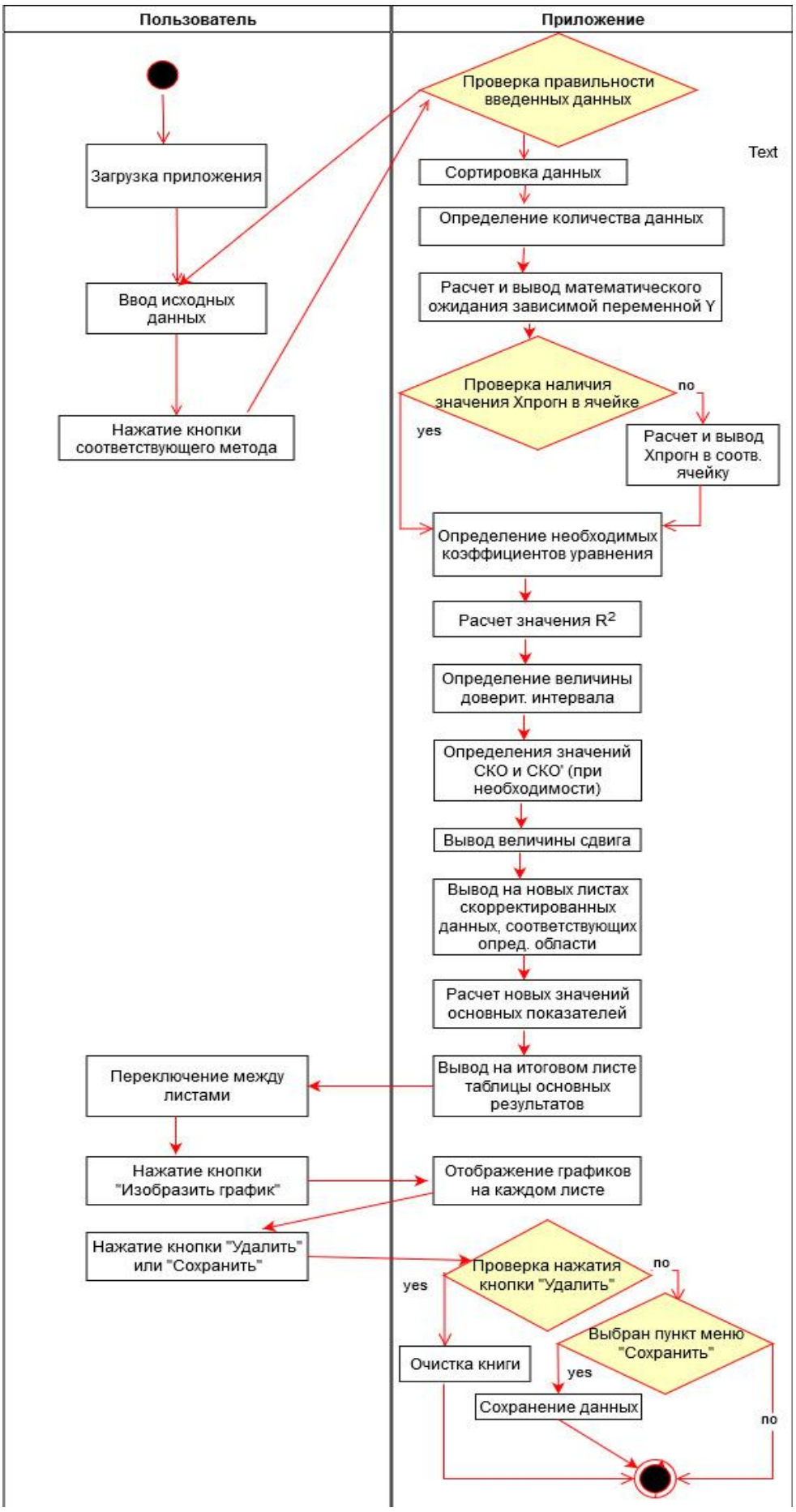


Рисунок 3.2 – Диаграмма активности

Работу отдельных блоков можно представить в виде UML диаграмм состояний.

Первая диаграмма (Рисунок 3.3) представляется собой работу блоков, отвечающих за ввод и контроль правильности исходных статистических данных.

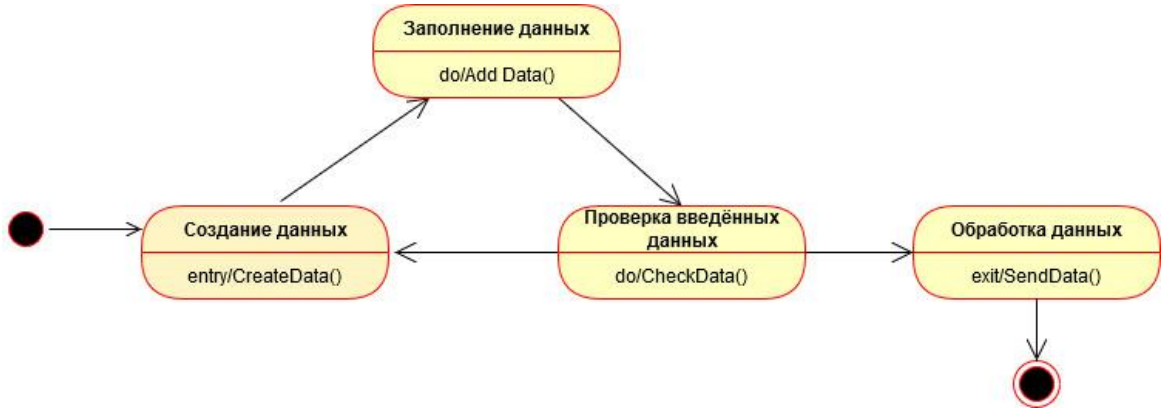


Рисунок 3.3 – Диаграмма состояний для отображения ввода и контроля данных

Далее, после проверки на корректность следует блок сортировки данных, в котором предварительно осуществляется подсчет количества исходных данных и дальнейшая сортировка данных. Соответствующая диаграмма состояний представлена на рисунке 3.4.

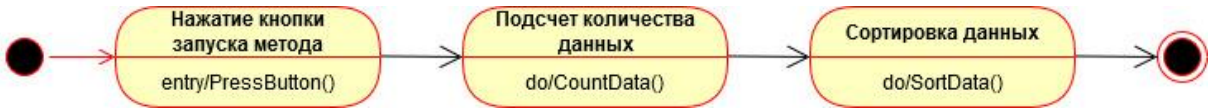


Рисунок 3.4 – Диаграмма состояний для отображения блока сортировки

Следующая диаграмма состояний (Рисунок 3.5) соответствует двум блокам – блоку расчета показателей и блоку вывода результатов.

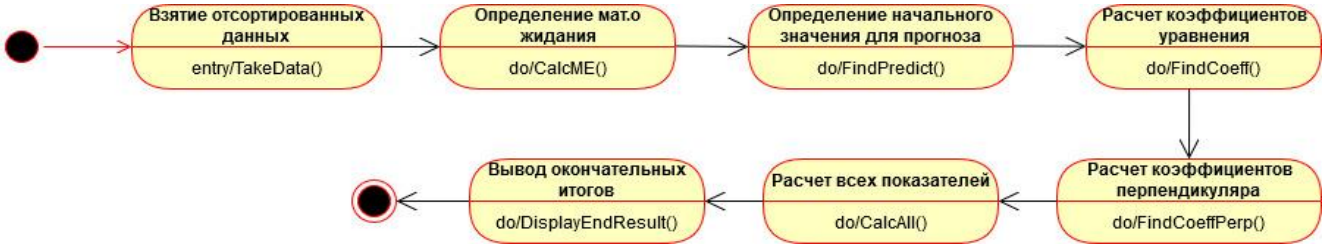


Рисунок 3.5 – Диаграмма состояний для отображения расчета показателей и вывода результатов

Заключительная диаграмма состояний представляет собой блок графического отображения необходимых данных (Рисунок 3.6).

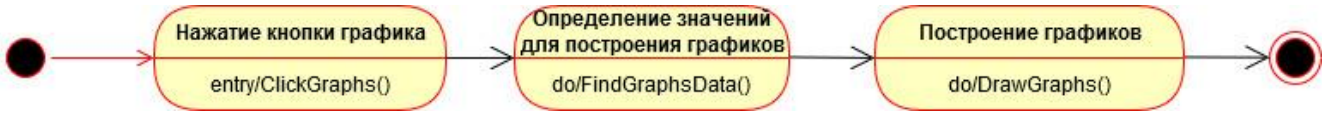


Рисунок 3.6 – Диаграмма состояний для изображения блока графического отображения

Выполнение графического отображения не является обязательным и вызывается пользователем при необходимости.

3.3 Описание работы разработанного программного приложения

Для каждого описанного в данной работе метода (метода, основанного на отбрасывании аномальных данных и метода, основанного на изменении данных) и их модификаций был разработан программный комплекс, которой состоит из взаимосвязанных приложений, написанных на языке C# и Visual Basic for Application для Microsoft Excel [91, 92, 93, 94]. Листинг фрагмента кода программного комплекса представлен в Приложении Б.

Исходное окно программы, выглядит как показано на рисунке 3.7.

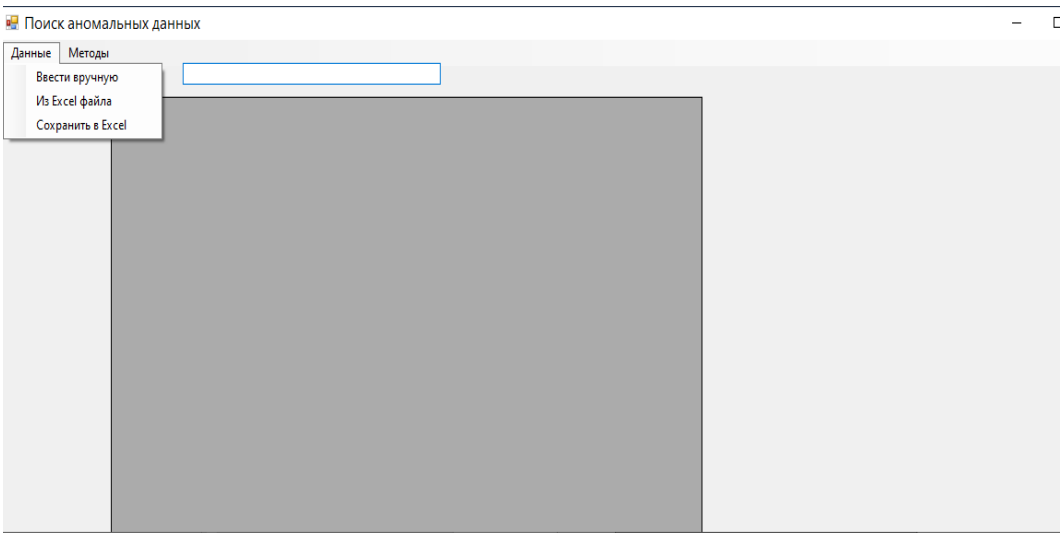


Рисунок 3.7 – Вид стартового окна программы

Пользователь может вводить значения вручную или из Excel файла. Окно с введёнными исходными данными представлено на рисунке 3.8.

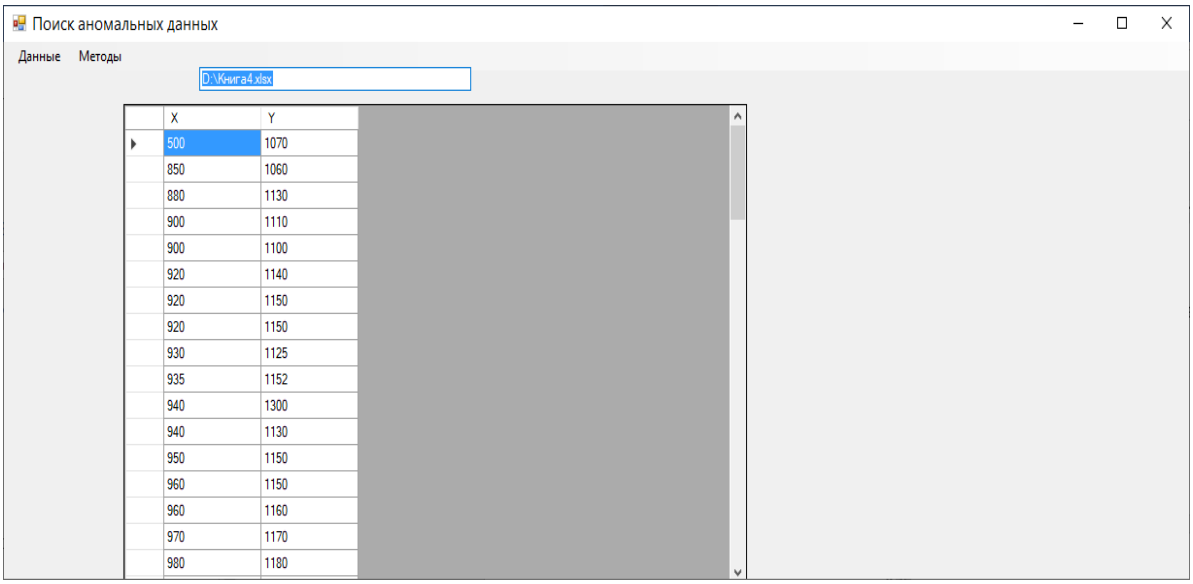


Рисунок 3.8 – Вид окна с исходными данными

После ввода исходных данных, пользователь определяется с тем, будет ли он использовать один из методов или выберет одну из его модификаций и нажимает соответствующую кнопку меню. Если данные введены некорректно, появляется соответствующее сообщение и пользователь может их изменить. После этого вновь нажимается соответствующая кнопка и, в случае корректного ввода данных это приводит к программному осуществлению всех необходимых расчетов и выводу итоговой таблицы (Рисунок 3.9), которая включает следующие параметры эффективности метода для каждого значения вероятности попадания в заданную область:

а) значения коэффициентов детерминации R^2;

б) величины доверительных интервалов;

в) величины смещений;

г) количество измерений (исходное и после отбрасывания);

г) точность;

д) коэффициенты нового линейного регрессионного уравнения.

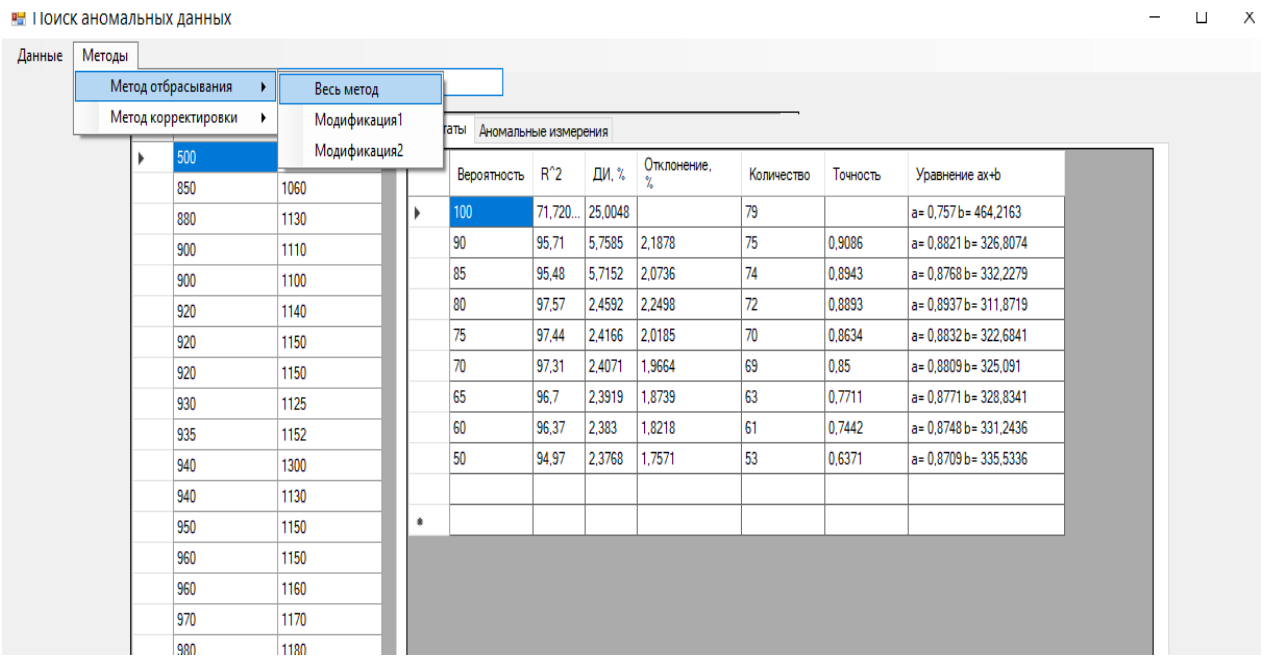


Рисунок 3.9 – Вид окна после выбора пункта меню «Метод отбрасывания-Весь метод»

При необходимости, можно простмотреть найденные в результате расчетов аномальные измерения для каждого значения вероятности, перейдя на соответствующую вкладку (Рисунок 3.10)

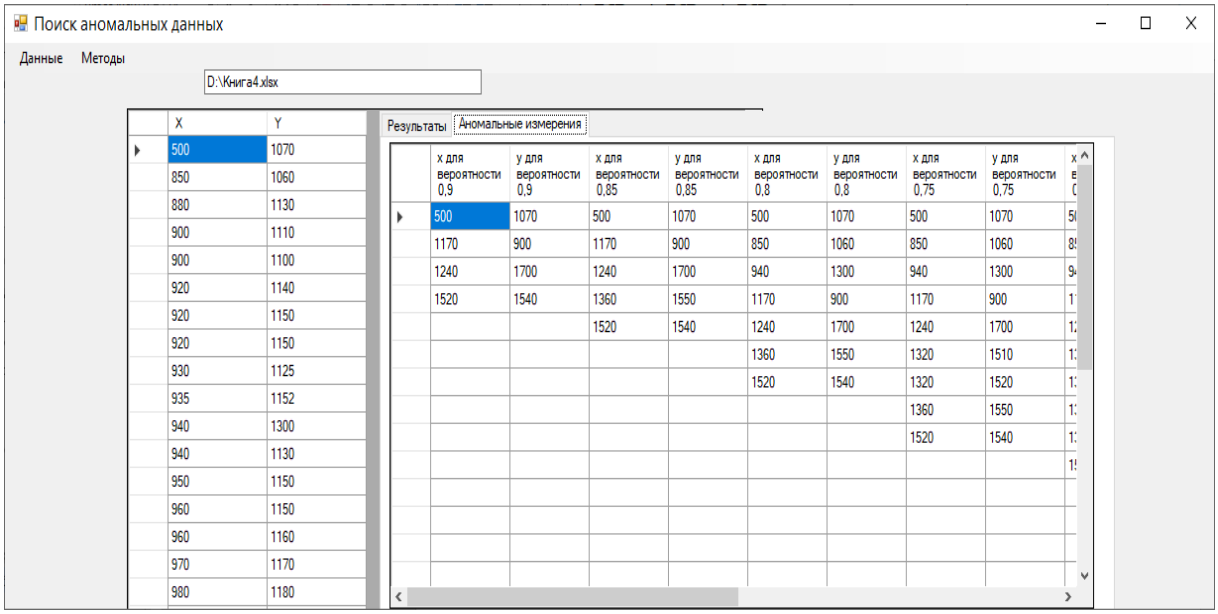
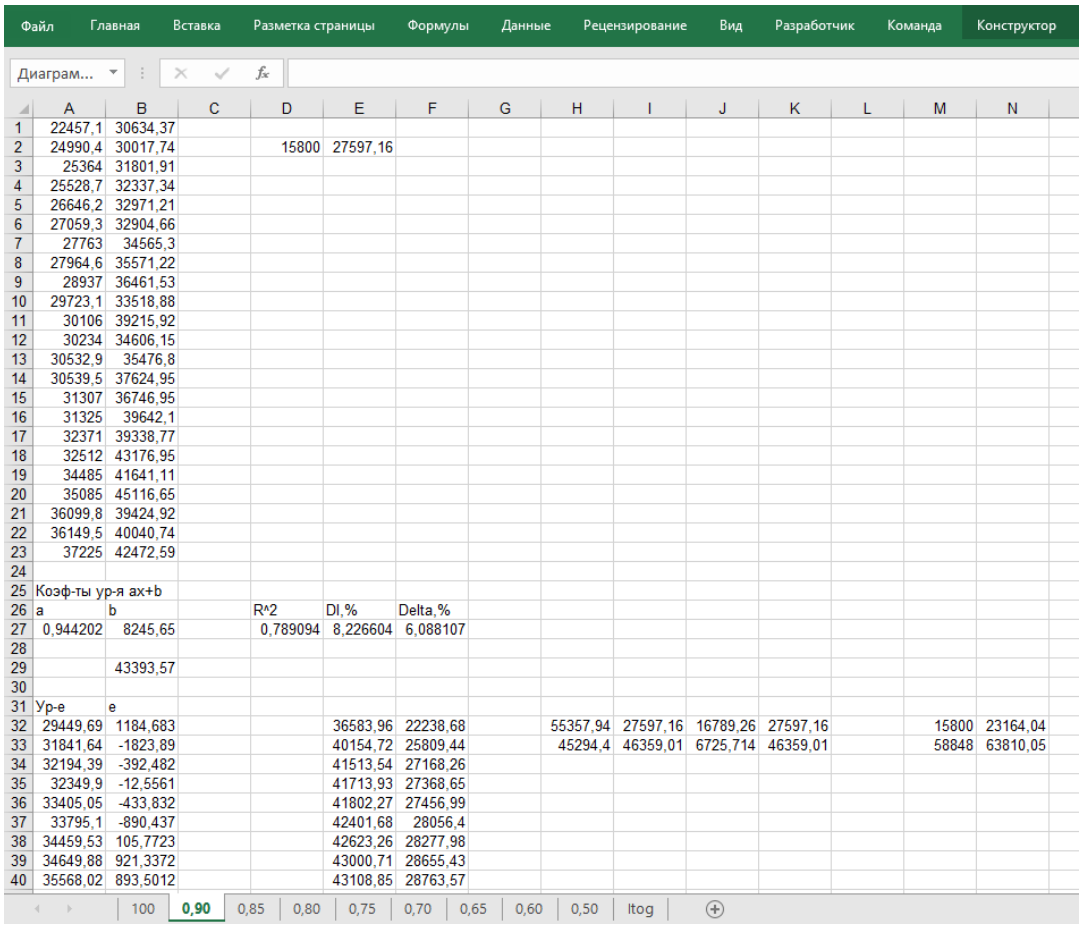


Рисунок 3.10 – Вид окна при выборе вкладки «Аномальные измерения»

Таким образом, все результаты отображаются в соответствующем окне программного комплекса, помимо этого результаты, а также промежуточные расчеты могут быть записаны в Excel-файл, который автоматически создаётся средствами C#. Кроме того, пользователь может не только рассчитывать показатели, но и заходить в редактор среды Microsoft Excel и работать непосредственно с ним (Рисунок 3.11).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	22457,1	30634,37												
2	24990,4	30017,74		15800	27597,16									
3	25364	31801,91												
4	25528,7	32337,34												
5	26646,2	32971,21												
6	27059,3	32904,66												
7	27763	34565,3												
8	27964,6	35571,22												
9	28937	36461,53												
10	29723,1	33518,88												
11	30106	39215,92												
12	30234	34606,15												
13	30532,9	35476,8												
14	30539,5	37624,95												
15	31307	36746,95												
16	31325	39642,1												
17	32371	39338,77												
18	32512	43176,95												
19	34485	41641,11												
20	35085	45116,65												
21	36099,8	39424,92												
22	36149,5	40040,74												
23	37225	42472,59												
24														
25	Коэф-ты ур-я ax+b													
26	a	b		R^2	DI,%	Delta,%								
27	0,944202	8245,65		0,789094	8,226604	6,088107								
28														
29		43393,57												
30														
31	Ур-е	e												
32	29449,69	1184,683			36583,96	22238,68		55357,94	27597,16	16789,26	27597,16		15800	23164,04
33	31841,64	-1823,89			40154,72	25809,44		45294,4	46359,01	6725,714	46359,01		58848	63810,05
34	32194,39	-392,482			41513,54	27168,26								
35	32349,9	-12,5561			41713,93	27368,65								
36	33405,05	-433,832			41802,27	27456,99								
37	33795,1	-890,437			42401,68	28056,4								
38	34459,53	105,7723			42623,26	28277,98								
39	34649,88	921,3372			43000,71	28655,43								
40	35568,02	893,5012			43108,85	28763,57								

Рисунок 3.11 – Вид рабочей области с промежуточными результатами расчетов

Для наглядного представления информации о данных, которые не попадают в заданную область на первом рабочем листе в поле «Изобразить график» (Рисунок 3.12) ставится галочка, в результате чего на листах «0,9»-«0,5» строятся соответствующие графики (Рисунки 3.13-3.14).

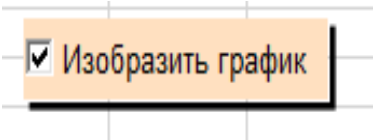


Рисунок 3.12 – Кнопка для построения графиков

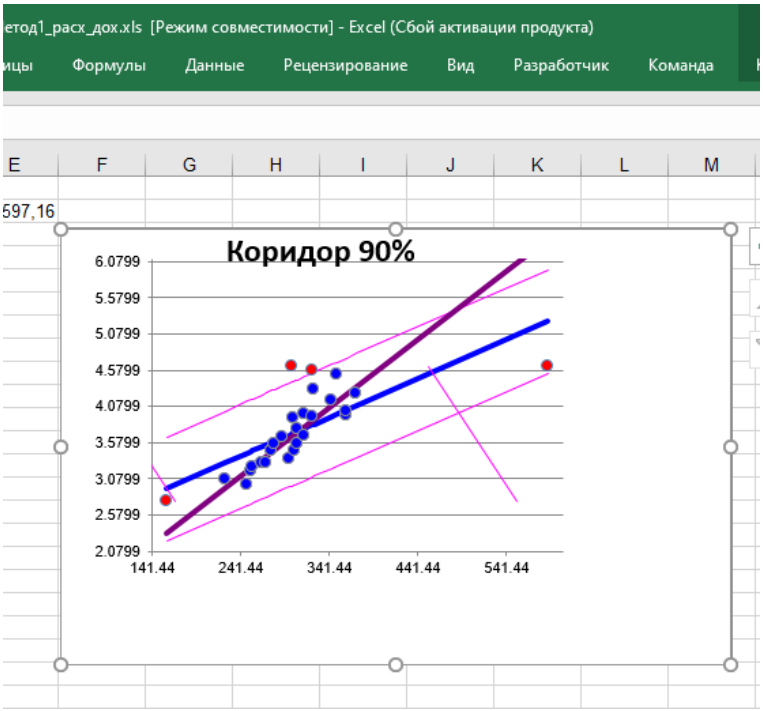


Рисунок 3.13 – График для доверительной области P=0,9

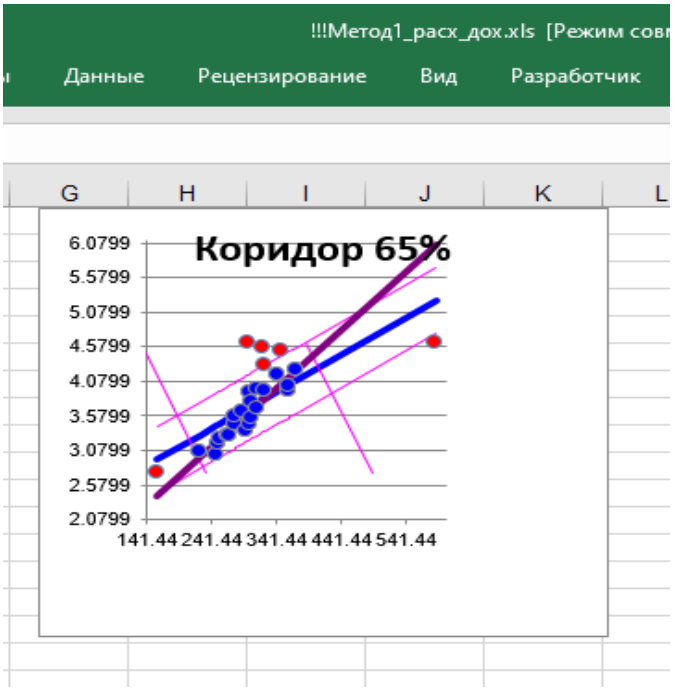


Рисунок 3.14 – График для доверительной области P=0,65

По окончанию построения графиков, полученные результаты можно сохранить с помощью стандартных команд Microsoft Excel или, при необходимости, все рабочие листы можно очистить, нажатием дополнительной кнопки «Delete», расположенной на первом листе.

3.4 Выводы по разделу 3

В данном разделе был описан разработанный для предлагаемых в диссертационной работе методов комплекс программ. Особенностью разработанного комплекса является то, что от пользователя не требуется специальных математических знаний и теоретических основ в области регрессионного анализа, а также информационных технологий. Основными достоинствами программного комплекса являются:

1) простота понимания и применения;
2) дружественный интерфейс;
3) небольшой размер приложения;
4) значительное сокращение времени расчетов по сравнению с ручной обработкой экспериментальных данных;
5) детальное отображение результатов расчетов, что позволяет просмотреть изменения, происходящие, в результате применения, выбранного метода;
6) наглядное графическое отображение областей, которые построены в соответствии с выбранными вероятностями и данных, представляющих собой аномальные измерения;
7) не требуется предустановка программного обеспечения. Файл с приложением может быть запущен на любом компьютере где установлен Microsoft Excel.

РАЗДЕЛ 4
СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРЕДЛОЖЕННЫХ МЕТОДОВ И РЕКОМЕНДАЦИИ ПО ИХ ПРИМЕНЕНИЮ

4.1 Сопоставление метода повышения качества линейных регрессионных моделей, основанного на отбрасывании данных и его модификаций

Сравнительный анализ метода повышения качества парных линейных регрессионных моделей, основанного на отбрасывании данных с его двумя модификациями, проводился на основании критериев эффективности, описанных выше (см. п. 2.1.4).

Для анализа использовались различные наборы экспериментальных данных (Таблицы В.1-В.19). В данном разделе будут рассмотрены два различных набора статистических данных с отличающимися исходными значениями коэффициента детерминации R^2. Со сравнительным анализом, полученным с использованием других статистических данных можно ознакомиться в Приложении В. Все расчеты и эксперименты, проводились с использованием разработанного в ходе данной работы программного комплекса, описанного в разделе 3.

Первый набор статистических данных представляет собой зависимость оборота розничной торговли непродовольственными товарами, млн. руб. от среднедушевого денежного дохода населения в Российской Федерации, руб./месяц (Таблица 4.1). Данные взяты поквартально с 2013 по 2019 гг.

Экспериментальные исследования проводились на основе данных, которые были взяты на официальном сайте Федеральной службы государственной статистики РФ [95].

Количество исходных статистических данных в рассматриваемом эксперименте равно 27. Значение коэффициента детерминации R^2 составляет 0,85. Это говорит о достаточно тесной линейной зависимости между переменными.

Таблица 4.1 – Исходные данные первой выборки

x	21800	24990,4	25528,7	30532,9	22457,1	27059,3	27964,6
y	2759716,2	3001774,3	3233734,4	3547679,7	3063437,2	3290465,9	3557121,9
x	32285	25364	29723,1	29945,5	36099,8	26646,2	30234
y	4064406,8	3180190,8	3351887,9	3635901,4	3942491,9	3297121,4	3460614,6
x	30539,5	36149,5	27763	31307	31325	37225	28937
y	3762494,9	4004074	3456530,3	3674695	3964210,3	4247259	3646152,5
x	32371	32512	38848	30106	34485	35085	-
y	3933877,4	4317694,5	4626216,3	3921591,6	4164110,8	4511665	-

Поскольку исходная модель достаточно хорошо описывает статистические данные, то для эксперимента добавим значения, которые будут отличаться от других. Для этого вначале отсортируем выборку по возрастанию доходов населения. Заменим при этом минимальное значение дохода с 21800 руб./месяц на 15800 руб./месяц, а максимальное с 38848 руб./месяц на 58848 руб./месяц. Далее в значениях оборота розничной торговли продовольственными товарами заменим 3635901,4 на 4635901,4, а 4064406,8 на 4564406,8 (Таблица 4.2).

Таблица 4.2 – Измененные данные исходной выборки

x	15800	22457,1	24990,4	25364	25528,7	26646,2	27059,3
y	2759716,2	3063437,2	3001774,3	3180190,8	3233734,4	3297121,4	3290465,9
x	27763	27964,6	28937	29723,1	29945,5	30106	30234
y	3456530,3	3557121,9	3646152,5	3351887,9	4635901,4	3921591,6	3460614,6
x	30532,9	30539,5	31307	31325	32285	32371	32512
y	3547679,7	3762494,9	3674695	3964210,3	4564406,8	3933877,4	4317694,5
x	34485	35085	36099,8	36149,5	37225	58848	-
y	4164110,8	4511665	3942491,9	4004074	4247259	4626216,3	-

В данном случае, исходное значение коэффициента детерминации значительно ниже, чем для первоначальных статистических данных и составляет всего 0,55. Поэтому целесообразно применить метод, обнаружения аномальных измерений и его модификации.

Результаты использования метода повышения качества парных линейных регрессионных моделей за счёт отбрасывания данных и его модификаций, представлены в таблицах 4.3, 4.4. В таблице 4.3, содержатся значения коэффициентов детерминации R^2 и точности T, рассчитанные при использовании модели, полученной после отбрасывания соответствующего количества статистических данных. Количество отбрасываемых наблюдений соответствует вероятности непопадания в заданную область. Величины смещения Δ (в %) предполагаемого прогнозного значения $Y_{прогн}$ и значения величины доверительного интервала отображены в таблице 4.4. В данном примере величина смещения находилась при значении $X_{прогн}$ равном 37225. Это наблюдение является предпоследним значением вариационного ряда и было выбрано, чтобы наглядно отобразить величину смещения, поскольку в этой точке величина смещения будет значительно больше, чем в точке равной среднему значению величины X. Последнее наблюдение вариационного ряда было решено не использовать, поскольку оно может оказаться аномальным и далеко отстоящим от других значений ряда. По этой же причине, во всех остальных экспериментах, для расчета величины смещения нового уравнения регрессии относительно исходного, было выбрано предпоследнее значение ряда.

Таблица 4.3 – Значения R^2 и точность для первой выборки

Вероятность попадания в область	Количество отброшенных точек			Метод, R^2	1-я модиф. R^2	2-я модиф. R^2	Метод. T	1-я модиф. T	2-я модиф. T
	Метод	1-я модиф.	2-я модиф.						
100	0	0	0	0,55	0,55	0,55			
90	2	2	2	0,69	0,69	0,61	0,64	0,64	0,561
85	4	4	2	0,789	0,81	0,61	0,67	0,68	0,561
80	4	5	2	0,789	0,85	0,61	0,67	0,69	0,561
70	5	6	3	0,79	0,854	0,57	0,64	0,66	0,51
65	7	6	5	0,83	0,854	0,5	0,61	0,65	0,44
60	8	7	7	0,8	0,876	0,4	0,56	0,65	0,3
50	9	8	8	0,77	0,87	0,36	0,51	0,61	0,24

На рисунке 4.1 изображены значения коэффициентов детерминации в соответствии с количеством отброшенных данных (в %), для метода и его модификаций.

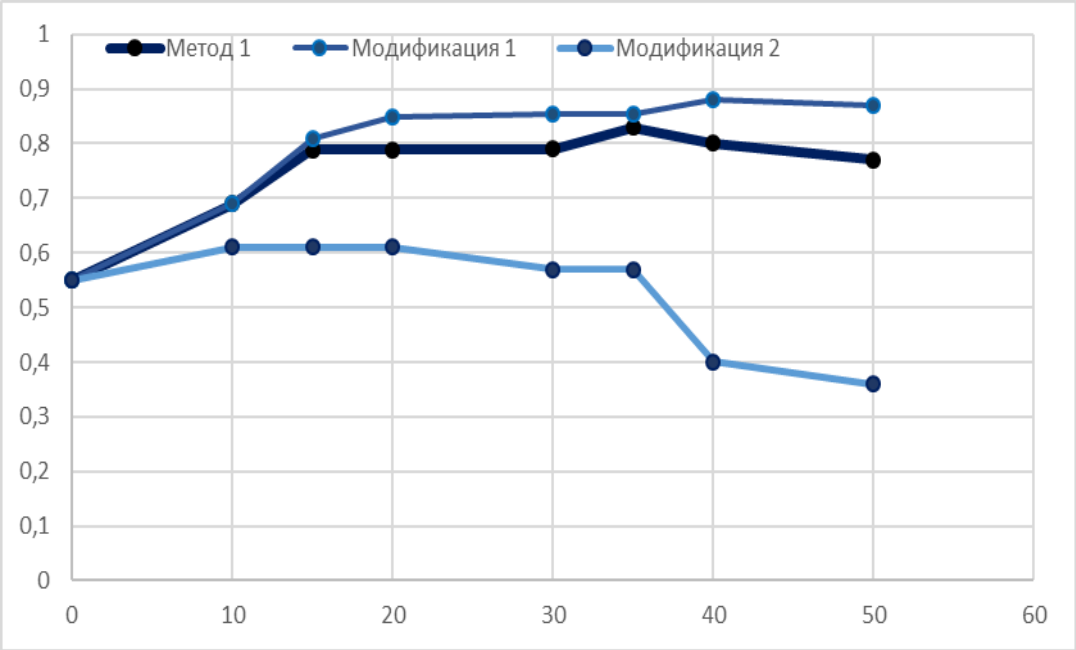


Рисунок 4.1 – Коэффициенты детерминации для метода и модификаций

Таблица 4.4 – Значения величины смещения Δ и доверительного интервала для первой выборки

Вероятность попадания в область	Метод. Δ,%	1-я модиф. Δ,%	2-я модиф. Δ,%	Метод. ДИ, %	1-я модиф. ДИ,%	2-я модиф. ДИ, %
0				20,82		
90	7	1,69	8,26	15,68	14,34	14,6
85	6	3,1	8,26	8,2	8,86	14,6
80	6	1,95	8,26	8,2	6,4	14,6
70	4,6	1,8	8,5	8,4	6,5	14,7
65	3,36	1,8	8	6,35	6,5	14,4
60	3,7	2,2	7	6,32	5,63	14,4
50	3,21	2,1	6,6	6,37	5,63	14,5

Таким образом, из таблицы видно, что при использовании предложенного метода обнаружения и устранения аномальных наблюдений были получены положительные результаты. Построение области надёжности при доверительной вероятности попадания данных в эту область равной 0,9 позволило обнаружить

аномальные измерения. При применении самого метода, в этом случае, было выявлено два аномальных наблюдения. Их отбрасывание позволило увеличить коэффициент детерминации с 0,55 до 0,69. Максимального значения коэффициент детерминации R^2 достиг при доверительной вероятности 0,65 (7 отброшенных точек) и составил 0,83. При этом отбрасывание всего 4-х значений позволило увеличить величину коэффициента детерминации до 0,79, что также свидетельствует об адекватности, найденной при этом линейной регрессионной модели. Хочется особо отметить, что данные 4 наблюдения – это те наблюдения, которые были изменены в исходной выборке для проведения эксперимента.

Максимальное значение R^2 при использовании первой модификации метода достигло 0,876 при отбрасывании 7 наблюдений. Однако, при этом аномальное значение независимой переменной X, равное 15800 руб./месяц не было обнаружено, поскольку достаточно близко находится к линии тренда. Поэтому, несмотря на более высокое значение коэффициента детерминации, в данном случае предпочтительным является применение самого метода повышения качества регрессионных моделей, основанного на отбрасывании данных.

Применение второй модификации метода, для данного эксперимента позволило повысить значение R^2 до 0,61, за счёт отбрасывания двух крайних аномальных наблюдение по независимой переменной X.

Во всех трёх случаях величина смещения не превышает 9%, а значение доверительного интервала, уменьшается по сравнению с исходным. Минимальной величина смещения является для первой модификации. Это объясняется тем, что одно из аномальных наблюдений по независимой переменной X в данном случае не обнаруживается. Однако из-за близкого расположения к уравнению регрессии оно оказывает значительное влияние на его вид, поэтому после исключения данного наблюдения (при использовании общего метода и второй модификации) из выборки, новое регрессионное уравнение значительно отличается от исходного.

Вычислительная сложность для данной выборке представлена в таблице 4.5.

Таблица 4.5 – Количество элементарных операций

Количество отброшенных точек, %	1-я выборка		
	Метод	1-я модиф.	2-я модиф.
0	702	432	513
10	934	658	742
15	925	649	733
20	916	640	724
30	907	631	715
35	889	613	697
40	880	604	688
50	871	595	679

Таким образом, как видно из таблицы вычислительная сложность является небольшой, а при применении компьютерных программ среднее время работы алгоритма будет минимальным, что позволит использовать, предложенный метод на любых персональных компьютерах.

Для второго эксперимента была выбрана зависимость урожайности яровой пшеницы (ц/га) от количества внесенных минеральных удобрений (кг/га) (на возрастающем участке) (Таблица 4.6) [96].

Таблица 4.6 – Зависимость урожайности яровой пшеницы от количества минеральных удобрений

x	20	22	24	26	28	30	32	34	36	36	38
y	15	15	15	16	16	17	18	17	18	17	18
x	40	42	44	46	48	50	52	54	56	58	60
y	20	20	20	20	22	22	22	22	24	18	24

Результаты применения первого предлагаемого метода повышения качества парных линейных регрессионных моделей за счёт отбрасывания данных и его модификаций, представлены в таблицах 4.7, 4.8. Исходное значение коэффициента детерминации составляет 0,8.

Таблица 4.7 – Значения R^2 и точность для второй выборки

Вероятность попадания в область	Количество отброшенных точек			Метод, R^2	1-я модиф. R^2	2-я модиф. R^2	Метод. T	1-я модиф. T	2-я модиф. T
	Метод	1-я модиф.	2-я модиф.						
100	0	0	0	0,8					
90	1	1	0	0,958	0,958	0,8	0,91	0,91	0,8
85	1	1	2	0,958	0,958	0,75	0,91	0,91	0,68
80	2	2	5	0,95	0,957	0,93	0,86	0,87	0,72
70	4	2	6	0,94	0,957	0,92	0,77	0,87	0,58
65	5	3	10	0,935	0,96	0,86	0,72	0,83	0,47
60	6	5	10	0,92	0,96	0,86	0,67	0,83	0,47
50	8	6	12	0,9	0,977	0,8	0,58	0,71	0,36

На рисунке 4.2 изображено графическое представление изменения значений коэффициентов детерминации в соответствии с количеством отброшенных данных (в %), для метода и его модификаций для второй выборки.

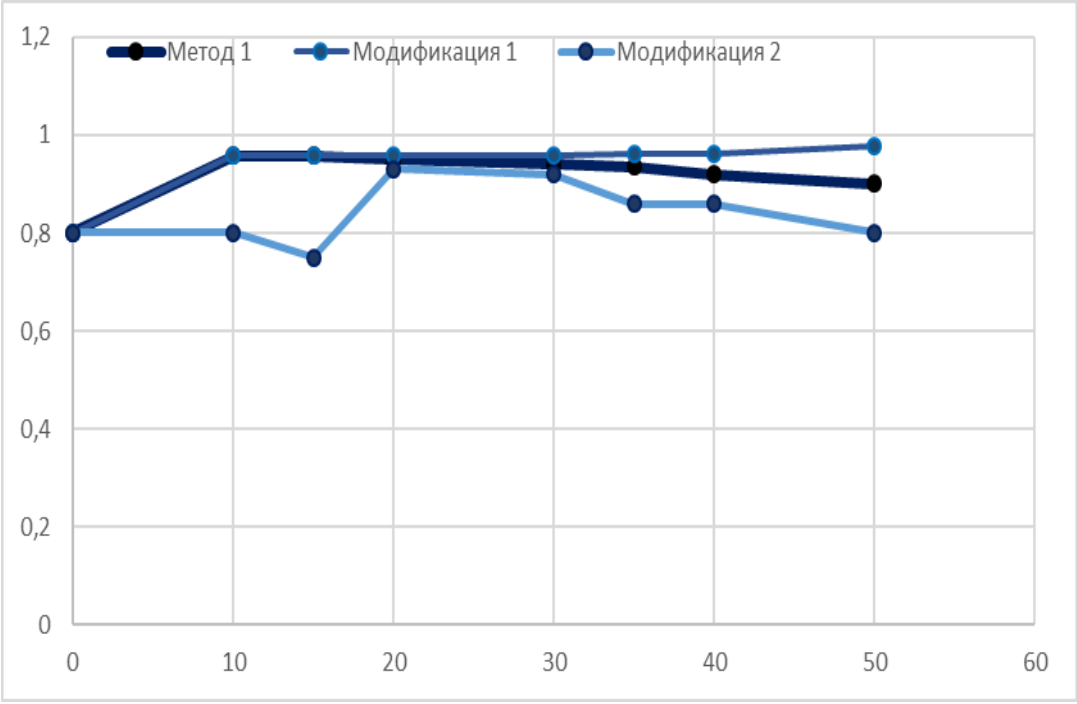


Рисунок 4.2 – Коэффициенты детерминации для метода и модификаций (вторая выборка)

Таблица 4.8 – Значения величины смещения Δ и доверительного интервала для второй выборки

Вероятность попадания в область	Метод. Δ,%	1-я модиф. Δ,%	2-я модиф. Δ,%	Метод. ДИ, %	1-я модиф. ДИ,%	2-я модиф. ДИ, %
0				14		
90	3,75	3,75	0	4,5	4,5	14
85	3,75	3,75	0,6	4,5	4,5	14
80	3,75	3,75	3,5	4,5	4,5	4,35
70	4,5	3	4,6	4,25	4,5	4,3
65	4,5	2,4	5,2	4,25	4,4	5,2
60	4,4	2,4	5,2	4,4	4,4	5,2
50	4,3	1,8	5,8	4,3	3,14	5,8

Поскольку исходные данные по независимой переменной X являются надёжными (количество минеральных удобрений, вносится согласно определенного плана) и не имеют значительных отклонений, достаточным является применение первой модификации, рассматриваемого метода. В этом случае при отбрасывании всего одного наблюдения (Рисунок 4.3) достигается высокое значение коэффициента детерминации равное 0,958. Такого же результата можно достичь, используя и сам метод, однако вычислительная сложность при этом будет несколько выше. Вторую модификацию метода, в данном случае, применять нецелесообразно, по причинам, описанным выше.

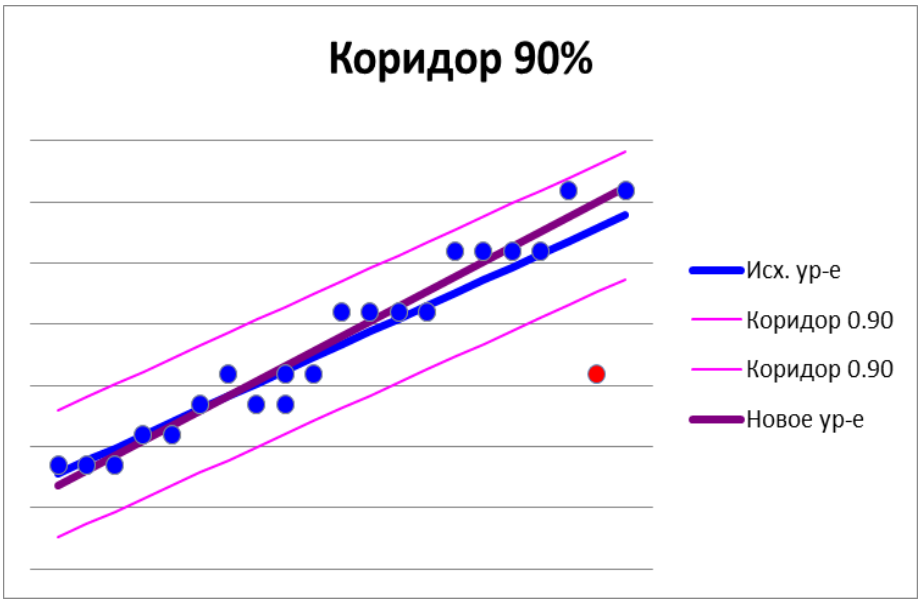


Рисунок 4.3 – Применение первой модификации доля области надёжности 90%

На основании проведённых экспериментов были получены следующие результаты:

- коэффициент детерминации R^2 увеличивается и достигает высоких значений (возрастание может доходить до 30%), что говорит о надёжности полученной при этом модели;

- величина доверительного интервала уменьшается до 3 раз;

- величина смещения прогнозного значения $Y_{прогн}$ не превышает 9%;

- число элементарных операций достаточно невысокое и является меньшим, чем при применении известных методов.

Таким образом, можно сказать, что применение предложенного метода повышения качества линейных регрессионных моделей, основанного на отбрасывании данных и его модификаций, даёт положительные результаты.

4.2 Сопоставление метода повышения качества линейной модели, основанного на перемещении данных и его модификаций

Анализ метода перемещения данных и его модификаций осуществлялся на основании тех же выборок и критериев эффективности, что и в разделе 4.1.

Результаты применения метода на основе переноса данных, а также его модификаций представлены в таблицах 4.9-4.12. В таблицах 4.9 и 4.11, содержится зависимость величины коэффициента детерминации R^2 от количества отбрасываемых точек в процентах для первой и второй выборки соответственно. Величины смещения прогнозного значения $Y_{прогн}$, рассчитанные в процентах и значения величины доверительного интервала, представлены в таблицах 4.10 и 4.12. В таблице 4.13 приведена информация о числе элементарных операций, необходимых при использовании метода.

Таблица 4.9 – Значения R^2 для первой выборки

Вероятность попадания в область	Метод, R^2	1-я модиф. R^2	2-я модиф. R^2
0	0,55		
90	0,71	0,61	0,69
85	0,75	0,65	0,7
80	0,77	0,69	0,69
70	0,79	0,74	0,66
65	0,791	0,77	0,62
60	0,792	0,79	0,6
50	0,78	0,82	0,5

Таблица 4.10 – Значения величины смещения Δ и доверительного интервала для первой выборки

Вероятность попадания в область	Метод Δ,%	1-я модиф. Δ,%	2-я модиф. Δ,%	Метод. ДИ, %	1-я модиф. ДИ,%	2-я модиф. ДИ,%
0				19		
90	5	0,34	7	12,2	15	14,4
85	5,9	0,18	8	11	13,7	14,7
80	6,7	0,134	9	10,5	12,5	14,9
70	7,6	0,14	10,3	10	10,9	15,7
65	8	0,14	10,7	9,7	10	17
60	8,2	0,13	11	9,44	9,4	17,8
50	8,5	0,1	12,6	9,8	8,2	19,4

Таким образом, из таблицы 4.9 видно, что коэффициент детерминации R^2, при использовании предложенного метода повышения качества регрессионных моделей, основанного на изменении значений аномальных наблюдений, возрастает с 0,55 до 0,79. Эта величина достигается путем изменения значений пяти наблюдений (Рисунок 4.4). Значение среднедушевого денежного дохода равное 15800 руб./месяц меняется на 22228,43 руб./месяц при обороте розничной торговли непродовольственными товарами 2759716.2 млн. руб., а значение 58848 руб./месяц – на 39907,18 руб./месяц при 4626216,3 млн. руб. По обороту розничной торговли непродовольственными товарами значения изменились с 4635901,4 млн. руб. на 4214832 млн. руб. при среднедушевом денежном доходе 29945,5 руб./месяц, с 4564406,8 млн. руб. на 4340319 млн. руб. при доходе 32285 руб./месяц и с 4511665

млн. руб. на 4490506 млн. руб. при 35085 руб./месяц. При этом смещение составляет менее 8%, а величина доверительного интервала уменьшилась в 2 раза.

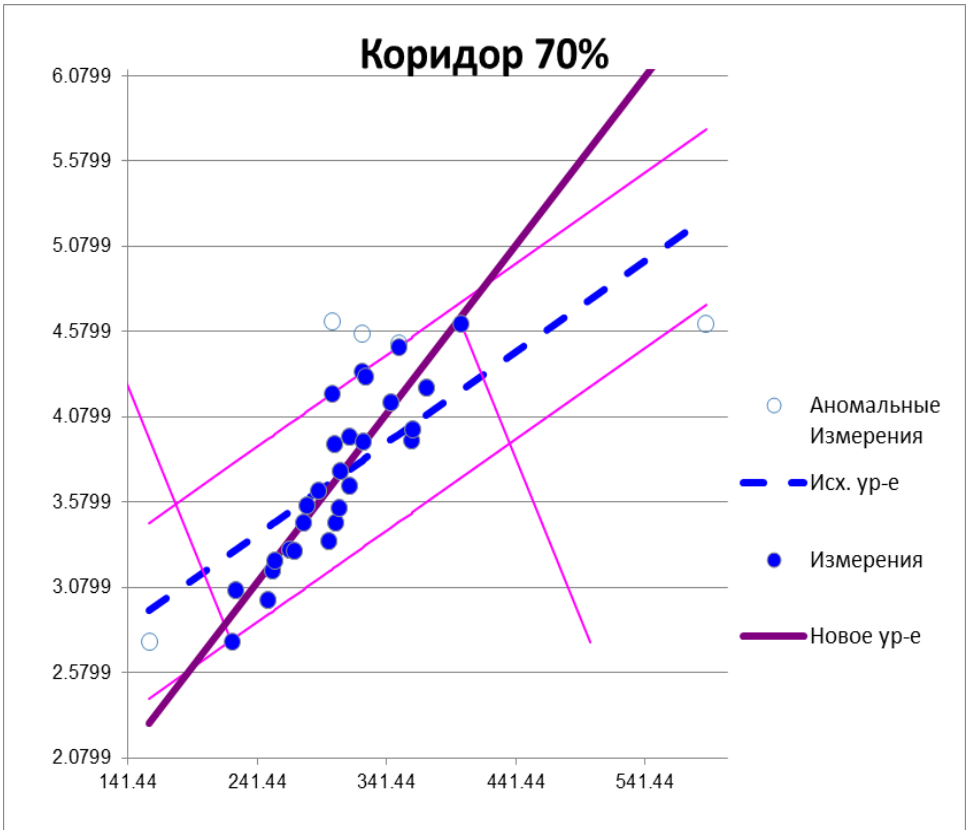


Рисунок 4.4 – Вид нового уравнения при изменении исходных данных для вероятности надёжности 0,7

Как видно из рисунка, новое уравнение отличается от исходного. Это связано с изменением значений по независимой переменной X.

Использование первой и второй модификации также дало положительные результаты, однако их применение для данного набора экспериментальных данных является недостаточным, поскольку не позволяет обнаружить все аномальные наблюдения.

Теперь рассмотрим результаты применение метода, основанного на изменении ненадёжных и аномальных измерений и его модификаций для второй выборки (зависимость урожайности яровой пшеницы от количества минеральных удобрений, данные таблицы 4.6).

Таблица 4.11 – Значения R^2 и величина доверительного интервала для второй выборки

Вероятность попадания в область	Метод, R^2	1-я модиф. R^2	2-я модиф. R^2
0	0,8		
90	0,9	0,91	0,8
80	0,915	0,93	0,8
75	0,918	0,933	0,78
70	0,92	0,94	0,8
65	0,921	0,943	0,79
60	0,919	0,95	0,79
50	0,91	0,96	0,77

Таблица 4.12 – Значения величины смещения Δ и доверительного интервала для второй выборки

Вероятность попадания в область	Метод Δ, %	1-я модиф. Δ, %	2-я модиф. Δ, %	Метод. ДИ, %	1-я модиф. ДИ,%	2-я модиф. ДИ,%
0				14		
90	1,4	1,7	0	9,3	8,4	14
85	1,6	1,9	0,4	8,7	7,8	14
80	1,8	2	1,45	8,3	7,3	13,6
70	2,5	2,1	3,7	7,7	6,3	13,4
65	3	2	5,5	7,4	5,7	13,2
60	3,3	2	7	7,8	5,4	13
50	4	1,9	10	8,5	4,7	12,75

В данном примере при использовании предложенного метода, основанного на корректировке аномальных данных за счёт их изменения, коэффициент детерминации вырос на 11,5% от исходного и достиг значения 0,915 при замене всего одного аномального наблюдения. Значение урожайности яровой пшеницы изменилось с 18 ц/га на 20,62 ц/га при количестве минеральных удобрений 58 кг/га, величина доверительного интервала при этом уменьшилась почти в 2 раза, а смещение не превысило 2%. Однако, как было выявлено выше, для данного набора данных наилучшим вариантом является использование первой модификации, рассматриваемого метода. Изменение всего двух значений урожайности пшеницы

с 18 ц/га на 21,35 ц/га и с 24 ц/га на 23,5 ц/га даёт рост коэффициента детерминации с 0,8 до 0,94, что является очень хорошим результатом.

Вычислительная сложность метода повышения качества линейных регрессионных моделей, основанного на изменении значений аномальных данных для 27 исходных данных представлено ниже (Таблица 4.13).

Таблица 4.13 – Количество элементарных операций

Количество измененных точек, %	1-я выборка		
	Метод	1-я модиф.	2-я модиф.
0			
10	965	678	772
15	966	679	772
20	966	680	772
30	967	681	778
40	968	682	790
50	974	683	802

На основании проведённых экспериментов по применению предложенного метода, основанного на перемещении аномальных наблюдений, были получены следующие результаты:

– обнаружение всех аномальных и ненадёжных измерений;

– рост коэффициента детерминации R^2 достигает 15-30%;

– величина доверительного интервала уменьшается до 2 раз;

– построение нового линейного регрессионного уравнения, которое является надёжнее, чем исходное, что позволяет построить более точный прогноз;

– число элементарных операций менее 1000, для исходного объема выборки равного 27 значениям, что является меньше, чем при применении известных методов.

Таким образом, можно утверждать, что использование метода поиска аномалий и последующей их корректировки за счёт перемещения данных и его модификаций, как и использование метода, основанного на отбрасывании аномальных наблюдений, дало хорошие результаты.

4.3 Сопоставление методов

4.3.1 Сравнение эффективности метода, основанного на отбрасывании данных и метода, основанного на перемещения данных

Для сопоставления двух методов была взята первая выборка данного раздела – зависимость оборота розничной торговли непродовольственными товарами от среднедушевого денежного дохода населения в РФ. Поскольку реальные данные, взятые из официального источника, были намерено скорректированы для проведения эксперимента, то сравнение будет проводиться между моделью, полученной по исходными официальным данным (Таблица 4.1) и моделями, полученными в результате применения, предложенных в данной работе методов, использованных для измененных исходных данных, имеющих аномальные значения (Таблица 4.2).

Перед сравнением методов между собой необходимо отметить следующий момент. При величине коэффициента линейной регрессии a > 5, следует значения переменных приводить в сопоставимый вид, путём деления одной из них на определённое число (например 10, 100, 1000…), т.е. другими словами произвести замену переменной. Это необходимо, поскольку из-за большой разницы между значениями X и Y, угол наклона уравнения стремится к 90° и при построении перпендикулярной линии, относительно исходного уравнения и соответствующей области надёжности, отбрасываться или корректироваться будут не те данные. После нахождения коэффициентов линейного регрессионного уравнения, путём обратной замены, находятся коэффициенты исходного регрессионного уравнения. В рассматриваемом примере, для нахождения коэффициентов регрессионного уравнения была произведена соответствующая замена z = y/100.

Исходное значение коэффициента детерминации R^2 для реальных данных достаточно высокое и составляет 0,85. Парная линейная регрессионная модель, в данном случае, имеет вид:

$$\hat{Y} = 99{,}67X + 672579{,}65. \quad (4.1)$$

Для скорректированной исходной выборки коэффициент детерминации был равен 0,55, а регрессионная модель имеет вид:

$$\hat{Y} = 53{,}64X + 2093646{,}375. \quad (4.2)$$

После применения первого метода повышения качества регрессионных моделей, основанного на отбрасывании данных, коэффициент детерминации вырос до 0,79 при отбрасывании всего четырёх аномальных значений, которые являлись именно теми значения, что были изменены в исходной выборке, основанной на реальных статистических данных. Регрессионная модели при этом изменилась и стала:

$$\hat{Y} = 94{,}42X + 824565{,}7. \quad (4.3)$$

Как видно из формул 4.1 и 4.3, соответствующие коэффициенты имеют очень близкие значения. Смещение уравнения, полученного по формуле 4.2 относительно уравнения, построенного по реальным данным (Рисунок 4.5) при максимальном значении переменной X, равном 38848 составляло 8%, а смещение нового уравнения, полученного после отбрасывания данных и исходного (Формула 4.1) составило всего 1.14% (Рисунок 4.6).

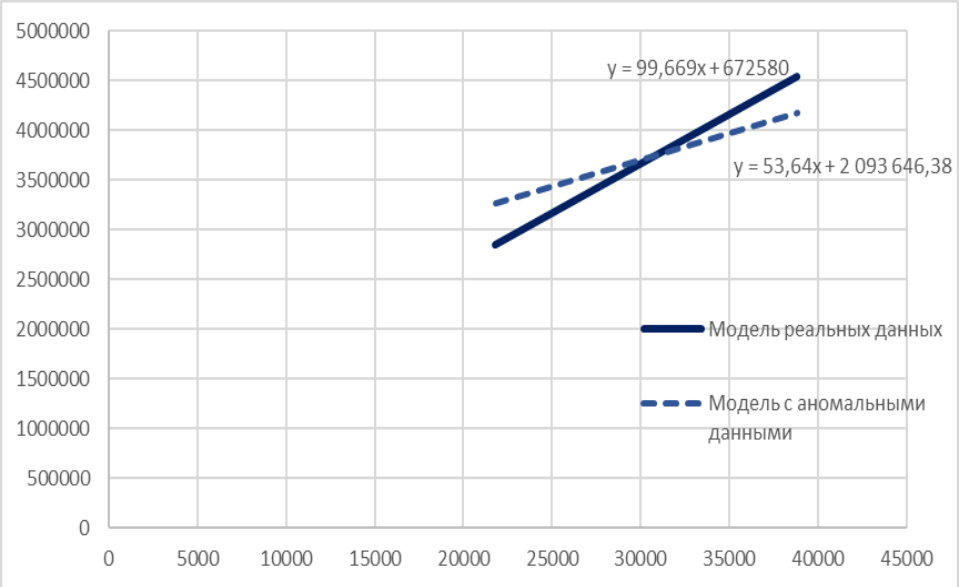


Рисунок 4.5 – Модель реальных данных и модель с аномальными данными

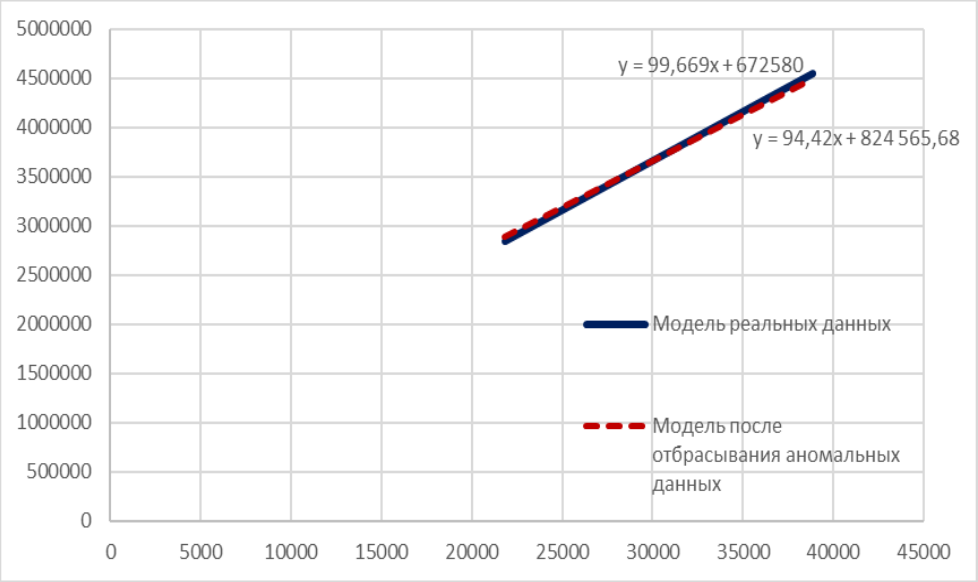


Рисунок 4.6 – Модель реальных данных и модель после отбрасывания аномальных данных

Таким образом, можно сделать вывод, что применение метода, основанного на отбрасывании данных, привело к получению модели идентичной реальной регрессионной модели, а значит она может быть использована для прогнозирования, поскольку результаты, полученные по данной модели будут точными и надёжными.

Перейдём к анализу второго метода, который заключается в корректировке значений аномальных наблюдений. В результате использования данного метода, коэффициент детерминации, так же, как и в предыдущем методе вырос до 0,79, при этом были изменены пять значений статистических данных. Максимальное и минимальное значение независимой переменной X изменились с 15800 и 58848 на 22228,43, и 39907,18 соответственно. Если сравнить их с реальными данными (21800 и 38848), то можно заметить, что они стали очень близки. Также, были скорректированы и три значения по зависимой переменной Y. Линейное регрессионное уравнение приняло следующий вид:

$$\hat{Y} = 98{,}88X + 721767{,}8. \qquad (4.4)$$

Разница между значением, полученным по реальной модели в точке 38848 и по новому регрессионному уравнению (Формула 4.4) составила всего 0,41 % (Рисунок 4.7).

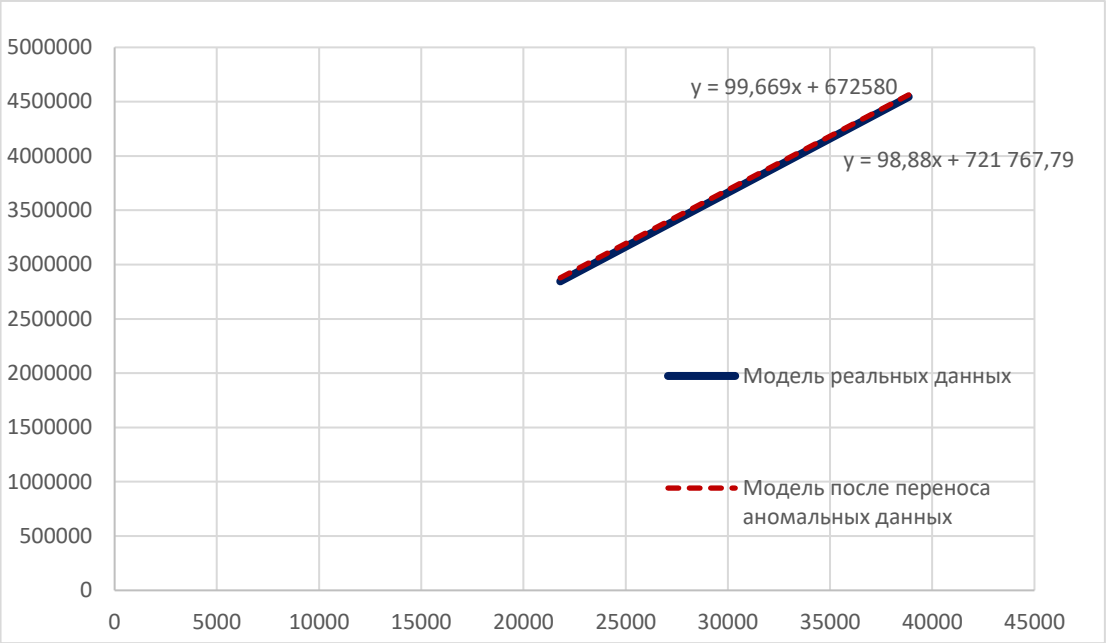


Рисунок 4.7 – Модель реальных данных и модель после изменения аномальных данных

Таким образом, можно сделать вывод, что оба метода дают положительные результаты. Вследствие применения методов повышения качества парных линейных регрессионных моделей были получены модели, которые полностью соответствуют уравнению, полученному по реальным исходным данным. Это свидетельствует о том, что предложенные методы способствуют повышению надёжности прогнозов, полученных по улучшенной новой модели регрессионного уравнения.

Для проверки качества полученных моделей построим по ним прогнозы оборотов розничной торговли непродовольственными товарами в четвёртых кварталах 2019 и 2020 гг., после чего сравним их с реальными данными (Таблица 4.14).

Таблица 4.14 – Прогнозные значения оборотов розничной торговли непродовольственными товарами

Период прогноза	X	Исходная модель	Модель с аномалиями	Модель после отбрасывания аномалий	Модель после корректировки аномалий	Реальные данные
4 квартал 2019г.	41328	4791741,41	4310480,3	4726755,44	4808280,44	4873283,6
4 квартал 2020г.	42543	4912840,46	4375652,9	4841475,74	4928419,64	5048843,7
Отклонение, % для 2019 г.	-	1,67	11,55	3	1,33	-
Отклонение, % для 2020 г.	-	2,7	13,33	4	2,4	-

Как видно из таблицы, прогнозные значения, построенные по регрессионным моделям, которые получены с использованием предложенных в данной работе методов очень близки к реальным данным. Максимальное отклонение прогнозных значений от реальных при этом составляет 4%, а минимальное 1,33%, при этом отклонение прогнозных значений, полученных по модели с аномальными данными составляет более 13%. Таким образом, можно говорить о положительном эффекте от применения предложенных в работе методов.

4.3.2 Применение предложенных методов на нелинейных моделях с внутренней линейностью

Как сказано в п. 2.3 данной работы, парные нелинейные модели с внутренней линейностью можно привести к линейному виду с помощью определенных преобразований.

Для проведения эксперимента рассмотрим два набора статистических данных. Данные для первого эксперимента представлены в таблице 4.15. Количество исходных статистических данных составляет 50.

Таблица 4.15 – Первый набор экспериментальных данных для нелинейной модели

x	43,1	19,9	19,2	17,7	18,1	20,3	21,5	16,9	15,5	18,5	27,2	41,5	28,0
y	1985	3365	3535	3445	3205	2830	3245	4360	4054	3940	3190	2144	2625
x	24,0	20,2	20,5	28,0	34,7	36,1	35,7	20,2	23,9	29,9	30,4	36,0	17,6
y	2865	3570	3155	2678	2215	1800	1915	2965	3420	2380	3250	1980	3465
x	32,9	38,0	24,2	38,1	39,4	25,4	31,3	34,1	34,0	31,0	27,4	22,3	-
y	2615	1965	2930	1968	2070	2900	2542	1985	2395	2720	2670	2890	-
x	27,2	39,1	28,0	27,5	33,7	64,0	34,4	20,6	46,6	23,7	22,6	36,4	-
y	2490	1755	2605	2560	2210	1850	3465	3380	2110	2420	2800	2950	-

Как видно из расположения точек (Рисунок 4.8) зависимость между X и Y близка к степенной.

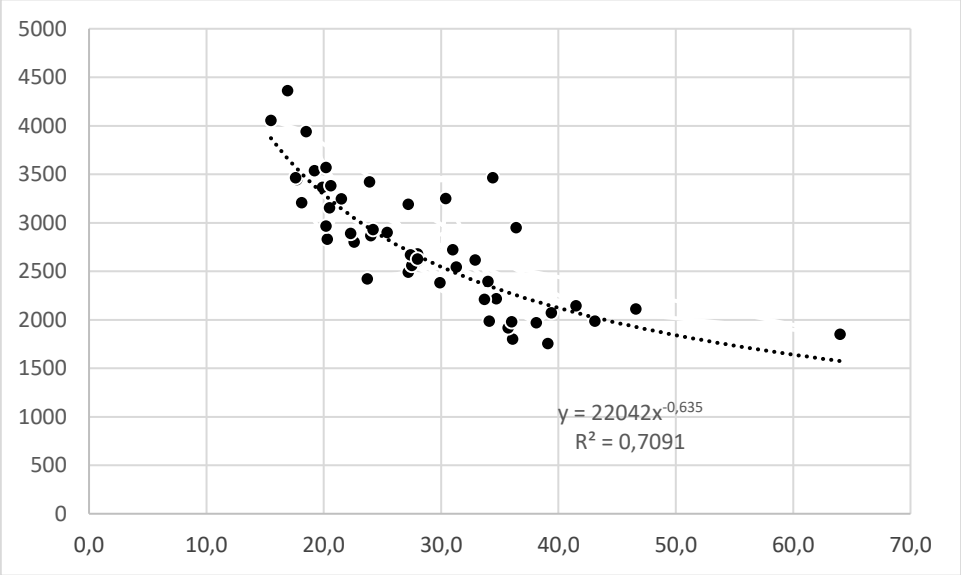


Рисунок 4.8 – Графическое изображение экспериментальных данных

Для рассматриваемого примера уравнение регрессии с вычисленными коэффициентами, которое будет выражать зависимость между переменными имеет следующий вид: $Y=22042X^{-0,635}$. Коэффициент детерминации R^2 равен 0,71. Чтобы воспользоваться предложенными в данной диссертационной работе методами повышения качества регрессионной модели, необходимо нелинейную модель вида $Y = aX^b$ привести к линейному виду $Y = aX + b$. Для этого воспользуемся следующими заменами: $Y1 = \ln(Y)$, $X1 = \ln(X)$, $B1 = \ln(B)$. Тогда уравнение примет следующий вид $Y1=AX1+B1$.

В таблице 4.16 представлены данные преобразованные согласно вышеизложенным формулам.

Таблица 4.16 – Преобразованные данные первого набора

x	3,764	2,991	2,955	2,874	2,896	3,011	3,068	2,827	2,741	2,918	3,3	3,73	3,44
y	7,593	8,121	8,170	8,145	8,072	7,948	8,085	8,380	8,307	8,279	8,07	7,67	7,84
x	3,303	3,666	3,332	3,178	3,006	3,020	3,332	3,547	3,586	3,575	3	3,17	3,53
y	7,820	7,470	7,865	7,960	8,180	8,057	7,893	7,703	7,496	7,557	7,99	8,14	7,59
x	3,584	3,118	3,595	3,314	3,517	4,159	3,493	3,638	3,186	3,640	3,67	3,24	
y	7,591	7,937	7,990	7,848	7,701	7,523	7,869	7,583	7,983	7,585	7,64	7,97	
x	3,526	3,434	3,311	3,105	3,332	2,868	3,538	3,025	3,842	3,165	3,4	3,41	
y	7,781	7,908	7,890	7,969	7,873	8,150	8,150	8,126	7,654	7,792	7,78	8,09	

Применяя предложенные в работе методы к преобразованным данным, получаем результаты, представленные в таблицах 4.17 и 4.18.

Таблица 4.17 – Результаты использования метода, основанного на отбрасывании данных для первого набора наблюдений

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,71	4,13		50	
90	0,84	2,52	1,00	46	0,77
85	0,82	2,47	0,70	43	0,70
80	0,82	2,45	0,51	42	0,69
75	0,84	2,17	0,75	40	0,67
70	0,83	2,15	0,71	38	0,63
65	0,81	2,09	0,40	34	0,55
60	0,84	2,00	0,65	32	0,54
50	0,81	1,10	0,14	23	0,37

Как видно из таблицы 4.17, значение коэффициента детерминации возрастает на 13% и достигает значения 0,84. Для этого отбрасывается 4 пары преобразованных экспериментальных данных: x=4,16 и y=7,52; x=3,59 и y=7,989; x=3,538 и y=8,15; x=3,4 и y=8,086. Что соответствует следующим парам исходных данных: x=64 и y=1850; x=36,4 и y=2950; x=34,4 и y=3465; x=30,4 и y=3250. Коэффициенты линейного уравнения регрессии при этом равны: A=-0,729, B1=10,288. Чтобы записать уравнение степенной функции, необходимо найти значение коэффициента B из выражения: $B=e^{B1}$. Таким образом, уравнение степенной функции, описывающее наилучшую модель для рассматриваемых данных будет иметь вид: $Y=29378X^{-0,729}$.

Таблица 4.18 – Результаты использования метода, основанного на переносе данных для первого набора наблюдений

Процент данных	R^2	DI,%	Delta,%
100	0,7091	4,13	
90	0,7570	3,18	0,35
85	0,7695	2,97	0,47
80	0,7706	2,88	0,56
75	0,7706	2,86	0,61
70	0,7700	2,97	0,69
65	0,7658	3,09	0,80
60	0,7589	3,27	0,92
50	0,7368	3,62	1,27

В данном случае значение коэффициента детерминации увеличивается на 6% и достигает значения 0,77. На рисунке 4.9 представлено наглядное изображение данных, которые изменяются.

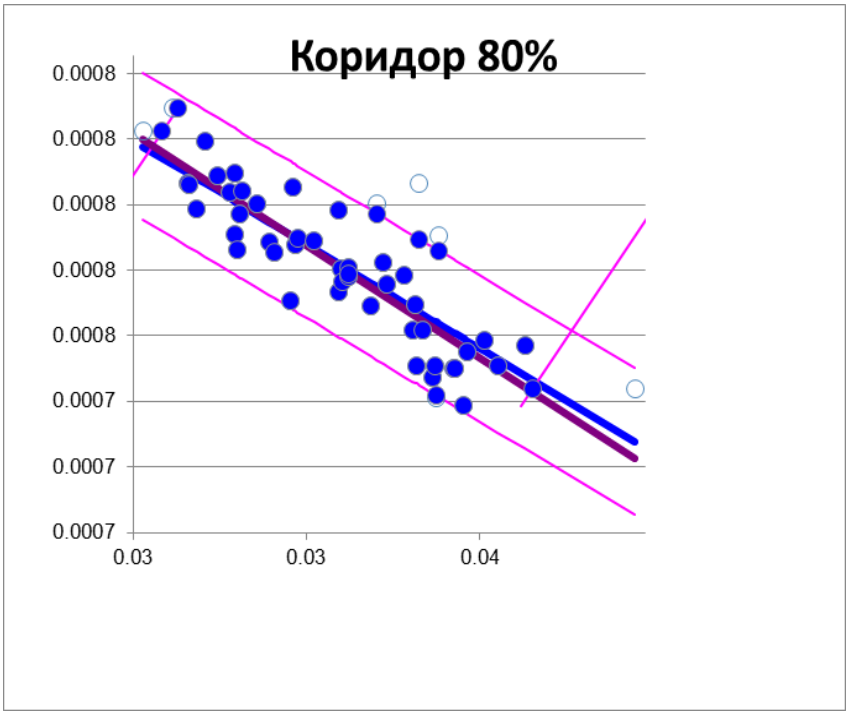


Рисунок 4.9 – Изображение изменившихся данных эксперимента

При этом коэффициенты линейного уравнения равны A=-0,6875 и B1=10,1685. Тогда после обратных преобразований нелинейное уравнение примет вид: $Y=26069X^{-0,6875}$.

Рассмотрим данные второго эксперимента (Таблица 4.19). Количество статистических наблюдений составляет 21.

Таблица 4.19 – Второй набор экспериментальных данных для нелинейной модели

X	120	112,3	107,25	107,25	127,25	112,2	113,85	122,1	122,1	132,1	137,05
Y	33,66	28,56	20,4	27,5	48,8	11	22,5	43,9	48	72,42	79,56
X	127,05	128,7	150	133,65	128,6	140,25	130,25	130,25	131,9	136,9	-
Y	55,08	90	77,52	81,6	56,1	77,52	51	66,3	64,26	66,3	-

Приведенные выше данные можно описать экспоненциальной моделью вида $Y = be^{ax}$. Чтобы привести данную экспоненциальную модель к линейному виду, необходимо произвести следующие замены – Y2 = ln(Y), b2 = ln(b). Тогда уравнение примет следующий вид: Y2 = ax + b2.

В таблице 4.20 представлены преобразованные наблюдения, к которым применяются предложенные в данной диссертационной работе методы поиска и обработки аномальных данных.

Таблица 4.20 – Преобразованные данные второго экспериментального набора

X	120	112,3	107,25	107,25	127,25	112,2	113,85	122,1	122,1	132,1	137,05
Y	3,516	3,352	3,016	3,314	3,888	2,398	3,114	3,782	3,871	4,282	4,377
X	127,05	128,7	150	133,65	128,6	140,25	130,25	130,25	131,9	136,9	-
Y	4,009	4,500	4,351	4,402	4,027	4,351	3,932	4,194	4,163	4,194	-

Найденное по данным из таблицы 4.20 линейное регрессионное уравнение имеет вид: Y = 0,0422X -1,4642. Коэффициент детрминации при этом равен 0,73.

С использованием разработанного в работе программного комплекса получаем результаты, представленные в таблицах 4.21 и 4.22 для метода обработки аномальных наблюдений, основанного на отбрасывании данных и метода, основанного на корректировке аномалий соответственно.

Таблица 4.21 – Результаты применения метода, основанного на отбрасывании данных для второго набора наблюдений

Процент данных	R^2	DI,%	Отклонение,%	Количество точек	Точность
100	0,73	15,802		21	
90	0,854	8,3357	1,7494	19	0,773
85	0,91	5,2992	0,8912	18	0,7777
80	0,88	4,6538	1,356	16	0,6716
75	0,88	4,6538	1,356	16	0,6716
70	0,88	4,6538	1,356	16	0,6716
65	0,88	4,6538	1,356	16	0,6716
60	0,88	4,6538	1,356	16	0,6716
50	0,80	3,8061	1,4084	13	0,4955

Из таблицы 4.21 видно, что коэффициент детерминации возрастает на 18% и достигает значения 0,91 при отбрасывании трёх аномальных наблюдений. Это следующие пары значений: x = 112,2 и y = 2,398; x = 128,7 и y = 4,5; x = 150 и y = 4,351. Линейная регрессионная модель при этом имеет вид Y = 0,0417X - 1,3568. Чтобы вернутся к экспоненциальной модели, необходимо определить коэффициент b из соотношения b2 = ln(b). Отсюда, получаем $b = e^{b2} = 0,257$. Таким образом, уравнение экспоненциальной функции, описывающее наилучшую модель для данных второго примера, будет иметь вид: $Y = 0,257e^{0,0417x}$.

Таблица 4.22 – Результаты применения метода, основанного на переносе данных для второго набора наблюдений

Процент данных	R^2	DI,%	Delta,%
100	0,73	15,80	
90	0,78	12,77	0,13
85	0,80	11,83	0,46
80	0,82	10,65	0,76
75	0,84	9,70	1,11
70	0,86	8,77	1,46
65	0,87	7,86	1,79
60	0,88	6,96	2,09
50	0,90	5,57	3,02

Как видно из таблицы, значение коэффициента детерминации R^2 увеличивается и достигает оптимального значения 0,84 при изменении всего 4 экспериментальных значений. При этом линейное регрессионное уравнение принимает вид Y = 0,042X – 1,424, а после преобразований, описанных выше, исходное экспоненциальное – $Y = 0{,}241e^{0{,}042X}$.

Как видно из полученных результатов, оба метода позволяют достичь значительного повышения коэффициента детерминации и улучшения качества модели, за счёт обработки аномальных наблюдений, что даёт основания рекомендовать их применение с целью повышения качества нелинейных моделей [97].

4.3.3 Сравнение метода Кука с первыми модификациями предложенных методов поиска и обработки выбросов

В связи с тем, что метод Кука позволяет отбрасывать сразу несколько исходных статистических данных одновременно, сравним данный метод с первыми модификациями, методов повышения качества регрессионных моделей. Для их взаимного тестирования был использован пример [30] о взаимосвязи возраста ребенка в месяцах, в котором он произнес свое первое слово (X), и оценок способностей тех же детей по результатам адаптивного теста Геселля (Y). Тесты заключаются в оценке четырех основных областей поведения маленького ребенка: моторной, адаптивной, речевой и личностно-социальной. Они в значительной степени основаны на простом наблюдении за поведением младенца. Для наглядности, на рисунке 4.10 изображены данные этого примера с нумерацией экспериментальных измерений. При использовании 100% исходных данных линейное регрессионное уравнение имеет вид:

$$\hat{Y}=109{,}87-1{,}127\cdot X, \qquad (4.5)$$

коэффициент детерминации R^2 равен при этом 0,41. При $X_{прогн}$=14,381, $Y_{прогн}$=93,67. Доверительный прогнозный интервал составляет 24,5% от величины $Y_{прогн}$=93,67 при $P_{дов}$=1.

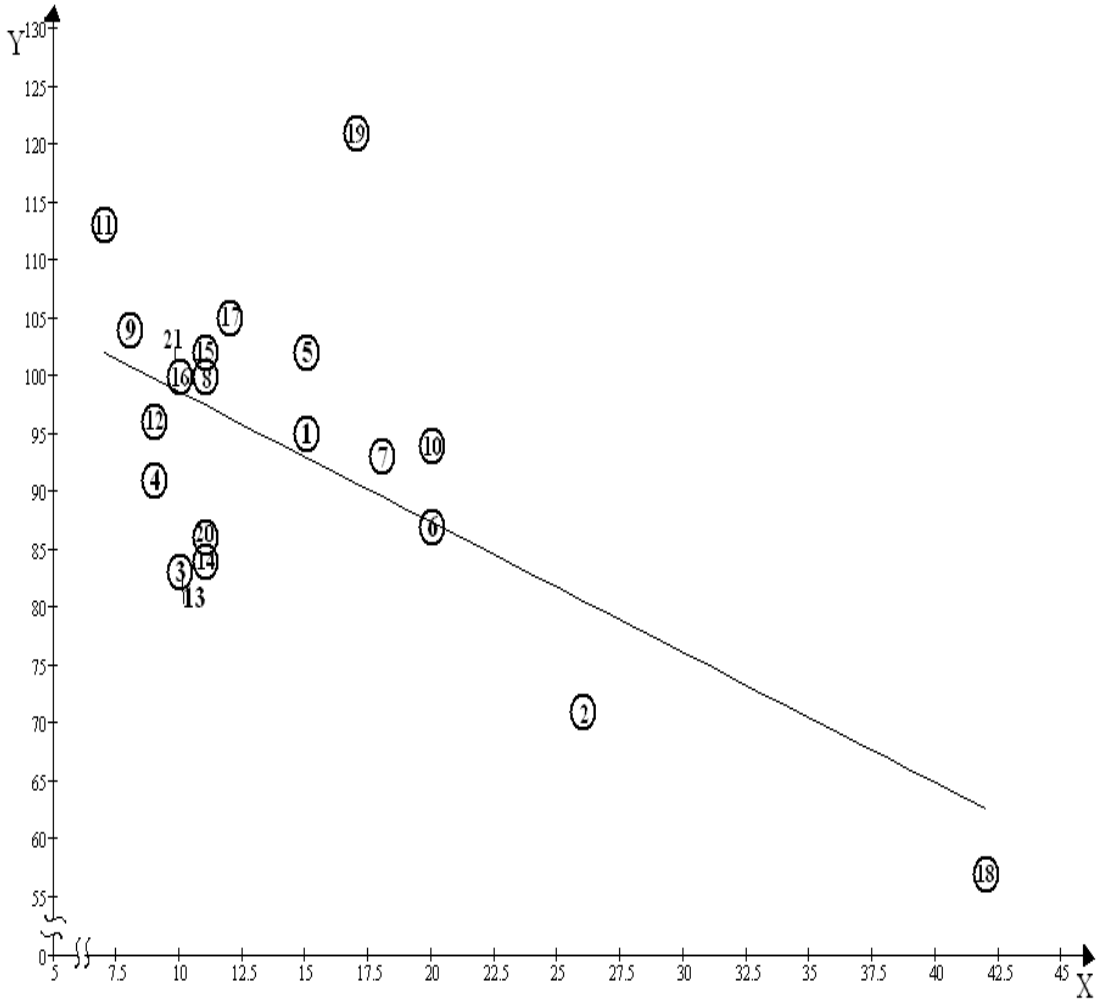


Рисунок 4.10 – Пример регрессионной модели для тестирования

Основная идея метода Кука, заключается в поиске наиболее влияющих на уравнение регрессий наблюдений. В случае вычисления статистики Кука для пар, троек и т.д. наблюдений, при отбрасывании данных, необходимо обнаружить уравнивающие наблюдения, которые будут компенсировать воздействия друг друга.

Результаты сопоставительного анализа приведены в таблице 4.23.

Таблица 4.23 – Результаты сопоставительного анализа

Используемый метод	Параметры метода	Номера исключенных или преобразованных наблюдений	Критерий качества метода			
			R^2	Δ^2, %	Величина доверительного интервала, %	Количество элементарных операций
Метод Кука	D_k=0,015	**2, 3, 19**	0,58	0,9		~ $0,4 \cdot 10^6$
	D_k=0,002	**2, 3, 11, 19**	0,57	0,58		~ $7 \cdot 10^6$
	D_k=0,002	**2, 3, 11, 14, 19**	0,64	0,2		~ $7 \cdot 10^6$
Метод, основанный на отбрасывании данных (1-я модификация)	k=1,27 (P=0,8) (D_k=0,17)	**19, 3, 13**	0,7	2,4	15	~ $0,5 \cdot 10^3$
	k=1,15 (D_k=0,7)	**3, 13, 14, 19**	0,77	2,03	13,7	~ $0,5 \cdot 10^3$
	k=1 (D_k=0,74)	**3, 13, 14, 20, 19**	0,83	1,5	12	~ $0,5 \cdot 10^3$
Метод основанный на переносе данных (1-я модификация)	k=1,27 (D_k=0,17)	**19, 3, 13**	0,53	1,4	18	~ $0,5 \cdot 10^3$
	k=1,15 (D_k=0,7)	**3, 13, 14, 19**	0,6	1,48	16	~ $0,5 \cdot 10^3$
	k=1 (D_k=0,74)	**3, 13, 14, 20, 19**	0,62	1,53	15	~ $0,5 \cdot 10^3$

Сопоставление результатов (Таблица 4.23) методов повышения качества прогнозных регрессионных моделей для рассмотренного примера дало следующие результаты:

− по критерию R^2 наиболее предпочтителен метод, основанный на отбрасывании данных, ему более, чем на 10% уступает метод Кука;

− по величине модуля смещения результата прогноза лучшим оказался метод Кука. Это объясняется тем, что метод Кука основан на поиске уравновешивающих значений, исключение, которых из выборки приводит к наименьшему смещению нового уравнения регрессии;

− по величине доверительного интервала прогноза метод, основанный на преобразовании исходных данных уступает методу отбрасывания данных на 3%, но опережает метод Кука на 30%.

− по количеству элементарных операций на ЭВМ, которые необходимы для реализации сравниваемых методов, метод-прототип полностью проигрывает

предложенным и не может быть рекомендован к использованию в современных компьютерных технологиях и в современных автоматизированных информационных системах для ведения оперативного управления (проигрыш достигает $2 \cdot 10^5$ раз для данного примера при числе отбрасывания точек k=5).

При использовании метода-прототипа Кука при исключении большого количества исходных статистических данных, количество возможных вариантов отбрасывания, по крайней мере, составляет $C_n^2 + C_n^3 + \ldots + C_n^{n/2}$, где C_n^m – количество сочетаний из n элементов по m. В этой ситуации отдельные варианты отбрасывания, при достаточно малых значениях D_k, являются явно ошибочными. При отбрасывании таких наборов измерений резко снижается величина R^2 (а критерий Кука об этом умалчивает). Это связано с сущностью критерия Кука, который направлен на слепое (в варианте использования ЭВМ) отслеживание минимума величины суммарного Δ_i. В машинном варианте критерий Кука возможно применять только при совместном использовании с критерием R^2. При ручной обработке данных эту проблему решает исследователь путем отбраковки некоторых вариантов, полученных по Куку.

Таким образом, методы, предложенные в данной работе, дают лучшие результаты практически по всем критериям и достаточно просты при применении в современных компьютерных технологиях, а метод Кука трудно формализовать из-за большой вычислительной сложности, и отсутствия формального критерия для выявления влияния конкретных наблюдений.

4.3.4 Сравнение метода «ящик с усами» с предложенным методом поиска аномалий

Для построения диаграммы размаха («ящик с усами») используется 5 основных показателей: медиана, первый и третий квартили, наименьшее и наибольшее значения в выборке.

Первый квартиль (нижняя граница ящика) – это величина, превышающая 25% значений ряда, а третий квартиль (верхняя граница) – величина, превышающая

75% значений. Отрезки («усы»), которые отходят вверх и вниз от середины прямоугольника – строятся на основе интерквартильного размаха и обозначают верхнюю и нижнюю границу значимой части данных, исключая выбросы.

Для сравнения методов были взяты данные из таблицы 4.2. При использовании метода «ящик с усами» аномальных данных по зависимой переменной Y выявлено не было (Рисунок 4.11), однако при применении предложенного в работе метода такие данные были обнаружены (Рисунок 4.12), что позволило увеличить значение коэффициента детерминации R^2.

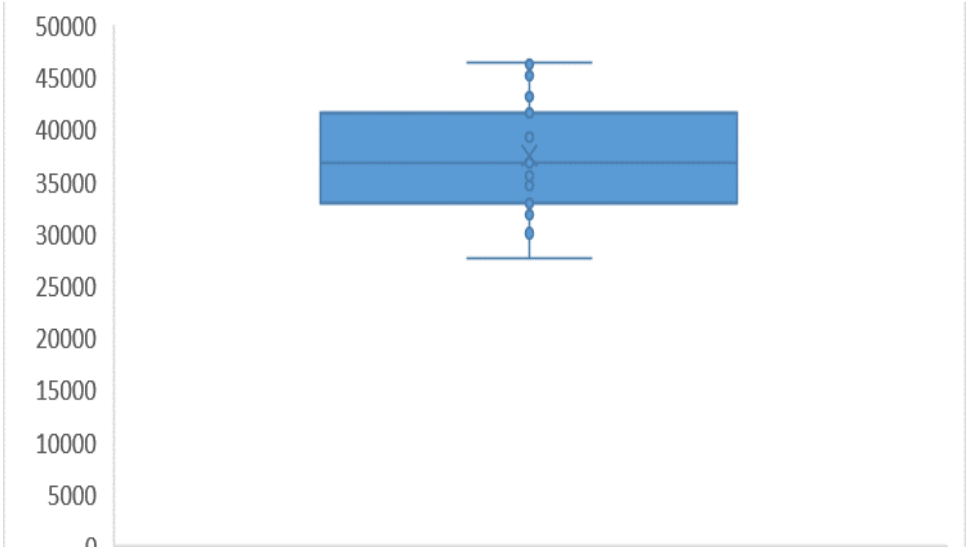


Рисунок 4.11 – Метод «ящик с усами» для исходных данных

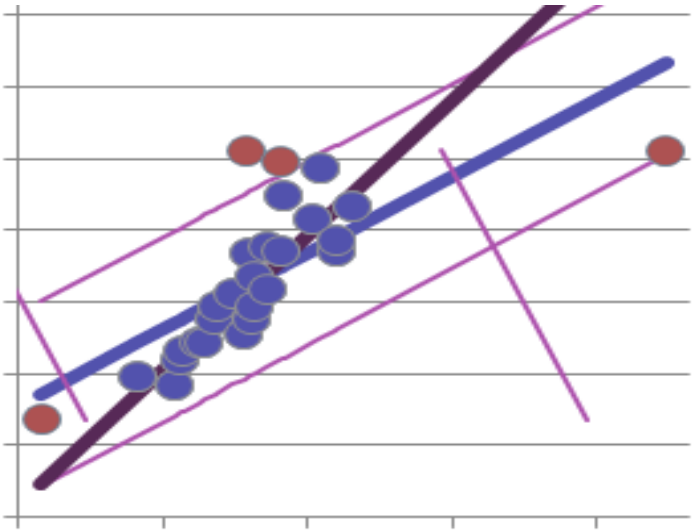


Рисунок 4.12 – Метод поиска аномалий, предложенный в работе

Основным недостатком метода «ящик с усами» является то, что для поиска аномальных данных используется сравнение с максимальным и минимальным значением выборки, однако при применении регрессионных моделей, выбросом может быть значение, не выходящее за рамки экстремумов, а отстоящее на определённое значение от уравнения регрессии, что не учитывается в данном методе.

4.3.5 Сравнение основных методов поиска аномалий для линейных регрессионных моделей и предложенного в работе метода

Для сравнения эффективности основных методов поиска аномалий в регрессии (Эктона, Титьена-Мура-Бэкмана, Прескотта-Лунда, Кука) и предложенного в работе метода использовались данные из таблицы 4.2.

При поиске аномальных значений методом Эктона было найдено подозрительное значение Y=46359 при X=29946, как наибольшее отклонение исходных измерений от расчетных данных. После этого, по формуле (1.37) было расчитано значение V, которое сравнивалось с критическим. Т.к. в данном случае расчетное значение оказалось больше критического, значение признаётся выбросом. Данный метод предназначен для проверки одного подозрительного значения, однако в качестве эксперимента были проверены ещё 2 значения. Одно из них оказалось выбросом: Y=45644,07 при X=32285.

Далее для поиска аномального значения использовался метод Титьена-Мура-Бэкмана. Используя формулу (1.39) было выявлено подозрительное значение. Величина, полученная по критерию, составила 2,71, однако она оказалась меньше критического значения для уровня значимости 0,1, поэтому подозрительное значение выбросом согласно данного метода не является.

Третий метод – метод Прескотта-Лунда. С помощью данного метода было получено значение R*=2,69, что является меньше критического равного 2,72. Следовательно, данное значение не признаётся выбросом.

Метод Кука является наиболее трудоёмким для ручного просчёта, поэтому поиск аномальных данных осуществлялся с использованием статистического пакета R. Были выявлены 3 выброса (Рисунок 4.13) – Y=46262,16 при X=58848, Y=46359 при X=29946 и Y=45644,068 при X=32285.

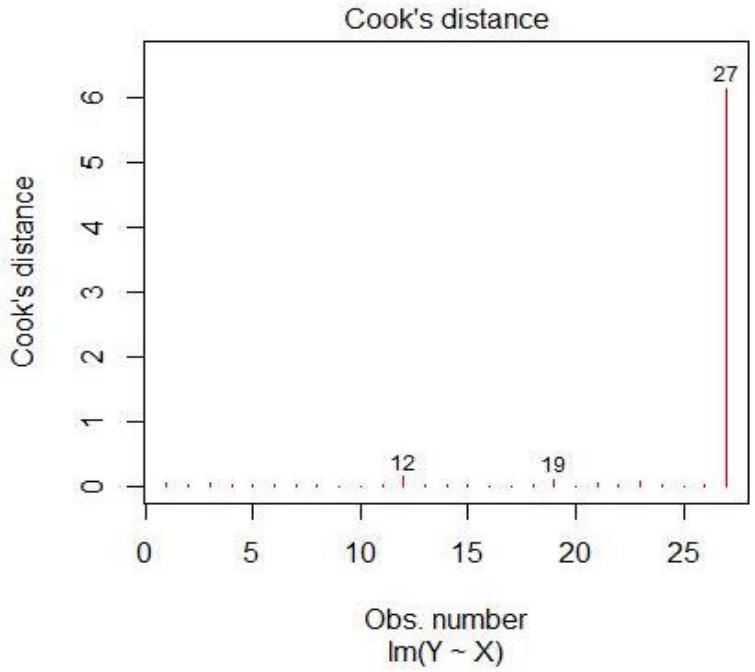


Рисунок 4.13 – Расстояние Кука

Результаты сравнения методов представлены в таблице 4.24.

Таблица 4.24 – Сравнительные данные методов поиска аномалий

Метод	Аномалии	Количество элементарных операций
Эктона	X=29946 Y=46359, X=32285 Y=45644,07	5n+5 (для поиска 1-го выброса) Для данной задачи: 280
Титьена-Мура-Бэкмана	не выявлено	12n+5 (для поиска 1-го выброса) Для данной задачи: 329
Прескотта-Лунда	не выявлено	5n+2 (для поиска 1-го выброса) Для данной задачи: 137
Кука	X=58848 Y=46262,16, X=29946 Y=46359, X=32285 Y=45644,068	10n (для поиска 1-го выброса), 8n (для последующих). Для данной задачи 702
Метод, предложенный в работе	X=58848 Y=46262,16, X=29946 Y=46359, X=15800 Y=27597,16, X=32285 Y=45644,068	17n+9 (для поиска всех выбросов, независимо от их количества) Для данной задачи: 468

Как видно из таблицы, методом поиска аномалий, предложенным в данной работе было выявлено большее число аномалий, чем другими методами. Этот результат был достигнут за 468 элементарных операций. Методом Эктона было обнаружено два аномальных наблюдения, а методами Титьена-Мура-Бэкманэ и Прескотта-Лунда не было выявлене ни одного. Ближе всего по результативности к предложенному методу оказался метод Кука, однако данным методом было выявлено 3 аномалии, помимо этого количество элементарных операций возрастает при увеличении количества проверяемых подозрительных значений, а также отсутствует конкретный критерий того, какие из подозрительных значеиий признавать аномальными. Таким образом, можно сделать вывод, что предложенный в данной работе метод поиска аномалий является наиболее эффективным и быстродейственным.

4.4 Рекомендации по выбору метода повышения качества регрессионной модели

На основе проведенных исследований можно выделить следующие рекомендации по выбору одного или другого метода:

1. Метод повышения качества прогнозной модели, основанный на отбрасывании исходных статистических данных следует применять в случае, если прогнозное значение $Y_{прогн}$ близко к среднему $\overline{Y}$, т.к. чем дальше от среднего значения $\overline{Y}$ отстоит $Y_{прогн}$, тем большего значение достигает смещение при использовании данного метода и наоборот. Метод, основанный на перемещении данных, применяется в обратной ситуации. Помимо этого, для выборок большого объема наиболее эффективна стратегия исключения данных, не попавших в коридор надежности, а для выборок малого объема – стратегия изменения данных, что позволяет сохранить исходное количество данных.

2. После того как выбран метод, с помощью которого будет производиться улучшение регрессионной модели, необходимо определить достаточно ли применить для повышения качества модели одну из модификаций метода. В случае

если у исследователя не возникает сомнений в исходных значениях независимой переменной X (например, значения X_i даны заранее и в них не может быть случайных ошибок), достаточно применить первую модификацию. Если подозрения вызывают только крайние значения независимой переменной X, то рекомендуется использовать вторую модификацию. Однако следует помнить, что в этом случае отсекаются (или изменяются) крайние точки, которые являются определяющими для уравнения регрессии, и если они находятся на значительном расстоянии от остальных данных, то оказывается большое влияние на исходное уравнение. При использовании второй модификации следует отбрасывать не более 10% исходных статистических данных

Если же абсолютной уверенности в исходных статистических данных нет, то предпочтительнее использовать сам метод, а не одну из его модификаций, т.к. в этом случае аномальные значения обрабатываются как по независимой переменной X, так и по зависимой переменной Y. Таким образом, картина является более точной.

3. Важным моментом при использовании предложенных методов, является критерий остановки, т.к. невозможно отбрасывать или изменять все исходные статистические данные. В качестве критерия остановки была выбрана точность T, рассчитанная по формуле 2.16. Как только величина T, становится меньше порогового значения 0,5, следует остановить применение, выбранного метода и выбрать наилучший вариант. Помимо этого, если коэффициент детерминации достиг достаточно высокого уровня (например, более 0,8), это также свидетельствует о том, что дальнейшее исключение(корректировку) данных можно не проводить, поскольку была найдена адекватная модель, а дальнейшее исключение данных может приводить к уменьшению точности. Также следует отметить, что в случае применения метода, основанного на отбрасывании данных, не следует исключать более чем 20% исходных данных. Это значение рекомендовано в литературных источниках [98], а также получено экспериментальным путём (Таблица 4.25)

Таблица 4.25 – Данные по количеству отбрасываемых измерений

Доверительная область	Процент отброшенных данных	R^2	T
1 выборка			
100		0,55	0,55
90	7	0,69	0,64
85	**15**	**0,789**	0,67
80	15	0,789	0,67
70	20	0,790	0,64
2 выборка			
100		0,8	0,8
90	**5**	**0,958**	0,91
85	5	0,958	0,91
80	9	0,95	0,86
70	18	0,94	0,77
3 выборка			
100		0,71	0,71
90	8	0,8	0,73
85	**16**	**0,81**	0,68
80	20	0,76	0,6
70	25	0,84	0,63
4 выборка			
100		0,72	0,72
90	5	0,95	0,9
85	6	0,95	0,89
80	**9**	**0,97**	0,89
70	13	0,97	0,85

Как видно из таблицы, оптимальных значений коэффициент детерминации и точность достигают при отбрасывании от 5-16% аномальных данных. Это говорит о том, что отбрасывание большего количества данных нецелесообразно.

4.5 Выводы по разделу 4

В данной главе был произведён сравнительный анализ методов, предложенных в диссертационной работе. На основании анализа можно сделать следующие выводы:

1. В результате, проведённых экспериментов было выявлено, что применение обоих методов повышения качества парных линейных регрессионных моделей дают хорошие результаты, благодаря обнаружению и устранению/корректировки всех ненадёжных статистических данных, что приводит к построению надёжной линейной регрессионной модели, на основании которой, можно получить достоверный прогноз.
2. Рост коэффициента детерминации достигает 30%.
3. Были рассмотрены, на конкретных примерах, ситуации, в которых достаточным является применение одной из упрощенных модификаций предложенных методов.
4. Даны рекомендации по выбору одного из методов или модификаций, а также по числу, отбрасываемых значений:

- при выборе одного из методов повышения качества прогнозной модели следует учитывать объём выборки. В случае большого объёма предпочтительнее метод основанный на отбрасывании данных, в противном случае – метод корректировки аномальных данных;

- в случае уверенности в исходных данных по одной из переменных можно использовать одно из упрощений методов;

- при применении методов следует учитывать критерий точности, поскольку данный критерий может являться индикатором для прекращения обработки исходных данных. Помимо этого, не рекомендуется отбрасывать более 20% исходных измерений.

ЗАКЛЮЧЕНИЕ

В диссертационной работе дано теоретическое обоснование и приведено решение актуальной научно-технической задачи совершенствования методов обработки исходных статистических данных с целью выявления и дальнейшей корректировки аномальных измерений, что позволяет повысить точность парных регрессионных моделей, используемых для прогнозирования в различных областях науки и техники.

Результаты диссертационного исследования могут быть сформулированы следующим образом:

1. Выполненный анализ наиболее часто используемых в рамках регрессионного анализа методов обнаружения и устранения аномальных данных показал наличие ряда недостатков, основным из которых является большая трудоёмкость вычислений. Исходя из этого, разработка новых методов обнаружении и устранении аномальных измерений для повышения качества парных регрессионных моделей, основанных на использовании доверительной вероятности, соответствующего коэффициента и учитывающих наклон уравнения регрессии, является актуальной научно-практической задачей.

2. Разработаны алгоритмы функционирования методов обнаружения и корректировки экспериментальных данных, базирующиеся на основных математических и статистических принципах, позволяющие улучшить качество регрессионных моделей, для дальнейшего их использования при прогнозировании и проектировании. Использование предложенных в работе методов ведёт к повышению качества прогнозов, полученных по линейным регрессионным моделям (от 10%), за счет предварительной обработки исходных данных. Возрастание коэффициента детерминации при этом может достигать от 10% до 30%.

3. Разработана архитектура оригинального комплекса программ для реализация предложенного алгоритма поиска и обработки аномальных данных, включающая следующие модули: программные модули на языке C# и Visual Basic

for Application, встроенном в MS Excel для обнаружения и удаления аномальных данных, программный модуль для корректировки выбросов, программный модуль для графического отображения, обнаруженных аномальных данных, модули для модификаций методов.

4. Проведённые численные эксперименты позволили сформулировать рекомендации по выбору одного из методов, наиболее подходящих в определенных ситуациях.

5. Обосновано, что предложенные в работе методы повышения качества парных регрессионных моделей одинаково эффективно используются как для линейных, так и для нелинейных регрессионных прогнозных уравнений с внутренней линейностью. Для этого исходное нелинейное регрессионное уравнение путём специальных преобразований приводится к линейному виду.

6. Показано, что предложенные методы обнаружения аномальных значений в исходных статистических данных позволяют наиболее быстро и точно проводить процедуру анализа данных на наличие грубых выбросов. Сокращение временных затрат на поиск и обработку аномальных данных достигается за счет уменьшения количества вычислительных операций (при этом сокращение может быть от $4*n$ до $2*10^s$, где n – количество исходных данных, s – число аномальных измерений).

7. Применение метода, состоящего в обнаружении и дальнейшем изменении значений аномальных измерений, позволяет обеспечить его реализацию для выборок малого объёма, поскольку в отличие от метода, где аномальные данные отбрасываются, в данном случае сохраняется исходное количество данных.

8. Представлены результаты применения методов для построения модели зависимости оборота розничной торговли непродовольственными товарами от среднедушевого денежного дохода населения в РФ, а также зависимостей из других предметных областей, подтверждающие эффективность предложенных в работе методов.

СПИСОК ЛИТЕРАТУРЫ

1. Edgeworth, F. Y. On discordant observations// The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. – 1887. – Vol. 23, no.5. – Pp.364–375

2. Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. ACM Comput. Surv. – 41, 3, Article 15 (July 2009) – 58 pages.

3. ГОСТ 8.736-2011 "Государственная система обеспечения единства измерений. Измерения прямые многократные. Методы обработки результатов измерений. Основные положения".

4. ГОСТ Р ИСО 16269-4-2017 Статистические методы. Статистическое представление данных Часть 4. Выявление и обработка выбросов

5. Тарасик, В.П. Математическое моделирование технических систем: учебник /В.П. Тарасик. — Минск: Новое знание, 2013. — 584 с.

6. Амосов А.А., Дубинский Ю.А., Копченова Н.В. Вычислительные методы для инженеров: Учеб. пособие. — М.: Высш. шк., 1994. — 544 с.

7. Рейзлин, В. И. Математическое моделирование. Учебное пособие/В.И. Рейзлин. - М.: Юрайт, 2016. - 128 с.

8. Семакин И.Г. и др. Программирование, численные методы и математическое моделирование - М.: КноРус, 2016. - 304 с.

9. Федоткин, И. М. Математическое моделирование технологических процессов- М.: Ленанд, 2015. - 416 с.

10. Юдин, С. В. Математика и экономико-математические модели. Учебник / С.В. Юдин. - М.: Инфра-М, РИОР, 2016. - 376 с.

11. Величко Е.Н. Экспериментальные методы исследования: основы теории эксперимента : [учебное пособие] – М-во образования и науки Рос. Федерации, С.-Петерб. политехн. ун-т Петра Великого. – СПб.: Издательство Политехнического университета, 2015. - 116 с.

12. Костин В. Н., Паничев В. В. Теория эксперимента: [учеб. пособие для вузов по направлениям подготовки 230100.68 «Информатика и вычисл. техника» и 231000.68 «Программная инженерия»] - Оренбург: Университет, 2014. - 212 с

13. Моргунов А. П., Ревина И. В. Планирование и анализ результатов эксперимента: [учебное пособие] - Минобрнауки России, Омский гос. техн. ун-т. - Омск : Издательство ОмГТУ, 2014. - 347 с.

14. Чернышева Е. В., Серых И. Р. Основы научных исследований, планирование и организация эксперимента: [учебное пособие для направления подготовки 27.04.01 «Стандартизация и метрология»] - Белгород: БГТУ, 2018. - 142 с.

15. Фельдман Л.П., Петренко А.І., Дмитрієва О.А. Чисельні методи в інформатиці. – К.: Видавнича група BHV, 2006. – 480 с.

16. Кун Макс, Джонсон Кьелл Предиктивное моделирование на практике – СПб. [и др.] : Питер, 2019. - 637 с.

17. Куприенко Н. В., Пономарева О. А., Тихонов Д. В. Статистика. Временные ряды. Анализ тенденций и прогнозирование: учебное пособие - М-во образования и науки Рос. Федерации, СПб. политехн. ун-т Петра Великого. – СПб.: Издательство Политехнического университета, 2015. - 122 с

18. Kuhn M., Johnson K. Applied predictive modeling. Springer, 2018. 600 p.

19. Елисеева, И. И. Эконометрика : учебник / И. И. Елисеева [и др.]; под ред. И. И. Елисеевой. – М. : Юрайт, 2016. – 449 с.

20. Носков С.И., Базилевский М.П. Построение регрессионных моделей с использованием аппарата линейно-булевого программирования: монография – Иркутск: ИрГУПС, 2018 – 176 с.

21. Bertsimas D. OR forum – An algorithmic approach to linear regression/D. Bertsimas, A.King//Operations Research. – 2016. – 64(1). – P. 2-16.

22. Faraway, Julian James Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models – Chapman & Hall/CRC, 2016 – 411 p.

23. Gatica G.N. A Simple Introduction to the MixedFinite Element Methods. Theory and Applications. – Heidelberg:Springer Briefs in Mathematics,2014.

24. Harrell Jr., Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer Series in Statistics, 2015. 582 p.

25. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction Springer, 2009. Corrected 12th printing, 2017.

26. Henning Best, Christof Wolf. The SAGE Handbook of Regression Analysis and Causal Inference - SAGE Publications Ltd, 2014 – 424 p.

27. Левин, Дэвид М. и др. Статистика для менеджеров с использованием Microsoft Excel, 4-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2004. – 1312 с.

28. Эндрю Ф.Сигел Практическая бизнес-статистика.: Пер. с англ. – М.: Вильямс, 2002. – 1056 с.

29. Вучков И. и др. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987. – 239 с.

30. Себер Дж. Линейный регрессионный анализ. – М.: Мир, 1980. – 455 с.

31. Э. Фёрстер, Б. Рёнц Методы корреляционного и регрессионного анализа. – Москва: Финансыи статистика, 1983. – 302 с.

32. Норман Р. Дрейпер, Гарри Смит Прикладной регрессионный анализ, 3-е изд.: Пер. с англ. – М.: Вильямс, 2007. – 912 с.

33. Четыркин Е.М. Статистические методы прогнозирования. Изд. 2-е, перераб. и доп. М.: Статистика, 1977. – 199 с.

34. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. – М.: Финансы и статистика, 1982. – 320 с.

35. Доугерти К. Введение в эконометрику: Учебник. 3-е изд./Пер. с англ. – М.: ИНФРА-М, 2009. – XIV, 465 с.

36. John Fox Applied Regression Analysis and Generalized Linear Models – SAGE Publ.,2015 – 816 p.

37. Marc S. Paolella Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH - Wiley Series in Probability and Statistics, 2018 – 897 p.

38. Richard B. Darlington , Andrew F Hayes. Regression Analysis and Linear Models: Concepts, Applications, and Implementation, 2016 – 689 p.

39. Wilson J. Holton , Keating Barry P. , Beal Mary Regression Analysis: Understanding and Building Business and Economic Models Using Excel, 2nd Edition. — New York, USA, Business Expert Press, LLC, 2016. — 205 p.

40. Wu C., Yu J. Z. Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting // Atmospheric Measurement Techniques. 2018. Vol. 11. P. 1233–1250

41. Jacek Welc, Pedro J. Rodriguez Esquerdo (auth.) Applied Regression Analysis for Business: Tools, Traps and Applications – Springer International Publishing, 2018 – 294 p.

42. Бородич С. А. Вводный курс эконометрики: Учебное пособие − Мн.: БГУ, 2000. − 354 с.

43. ГОСТ Р 8.736-2011 ГСИ. Измерения прямые многократные. Методы обработки результатов измерений. Основные положения

44. Сергеев А.Г. Метрология, стандартизация и сертификация: учебник / А.Г. Сергеев, В.В Терегеря. – М.: Издательство Юрайт, 2011. -820 с.

45. Заляжных В.В, Чупакова А.А. Сравнительный анализ критериев грубых ошибок. Ломоносовские научные чтения студентов, аспирантов и молодых ученых - 2013: Сборник материалов конференции: - Т1, Архангельск, КИРА. С. 166-169.

46. Ван дер Варден Б.Л. Математическая статистика, перевод с немецкого. — М.: Изд-во Иностранной литературы, 1960. - 436 с.

47. Бондарчук С.С., Бондарчук И.С. Статобработка экспериментальных данных в MS Excel: учебное пособие. – Томск: ИздательствоТомского государственного педагогического университета, 2018. –433 с.

48. Попукайло В.С. (2017), Поддержка принятия решений по пассивным выборкам малого объёма, Дисс. ... доктора информатики, Кишинёв, АН Республики Молдова, Институт математики и информатики, http://www.cnaa.md/files/theses/2017/52347/vladimir_popukaylo_thesis.pdf

49. Смирнов, А.В., Рычка, О.В. Метод повышения качества прогнозных регрессионных моделей // Наукові праці Донецького національного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка. - 2010. - Вип. 12. - С. 141-147.

50. Лемешко Б. Ю, Лемешко С. Б. Расширение области применения критериев типа Граббса, используемых при отбраковке аномальных измерений// Измерительная техника. – 2005. – № 6 – С.13-20

51. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006. – 816 с.

52. Костин В.Н., Тишина Н.А. Статистические методы и модели: Учебное пособие. – Оренбург: ГОУ ОГУ, 2004. – 138 с.

53. Третьяк Л.Н. Обработка результатов наблюдений: Учебное пособие. – Оренбург: ГОУ ОГУ, 2004. – 171 с.

54. Грибовский С.В., Баринов Н.П., Анисимова И.Н. О повышении достоверности оценки рыночной стоимости методом сравнительного анализа.

55. Дж. Тейлор. Введение в теорию ошибок. Пер. с англ. – М.: Мир, 1985. – 272 с.

56. David H.A. Revised upper percentage points of the extreme studentized deviate from the sample mean// Biometrika. – 1956. – V. 43. P. 450-461.

57. Казакавичус К.А. Приближенные формулы для статистической обработки результатов механических испытаний// Заводская лаборатория. – 1988. – Т.54, №12. С. 82-85.

58. Hoaglin D.C., Iglewicz B. Fine-tuning some resistant rules for outlier labeling // JASA. – 1987. – V.82, №400. P.1147-1149.

59. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и её инженерные приложения. Учеб. пособие для втузов. – 2-е изд. – М.: Высш. шк., 2000. – 480 с.

60. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере – 3-е изд., перераб. и доп. – М: ИНФРА-М, 2003. – 544 с.

61. Суслов, И. П. Общая теория статистики: Учеб. пособие. – М.: Статистика, 1970. - 376 с.

62. Acton F.S. Analysis of straight line data. – N.Y.: Welly, 1959 – P.261.

63. Tietjen G.L., Moore R.H., Beekman R.J. Testing for a single outlier in simple linear regression // Technometrics. – 1973. – V.15, № 4. P. 717-721.

64. Попукайло В.С. Исследование критериев грубых ошибок применительно к выборкам малого объема // Радіоелектронні і комп'ютерні системи. – 2015г. – № 3(73). – С. 39-44.

65. Попукайло В.С. Обнаружение аномальных измерений при обработке данных малого объема // Технология и конструирование в электронной аппаратуре. – 2016 г. – № 4-5. – С. 42-46.

66. Рычка О.В. Улучшение прогнозных значений с использованием метода отбрасывания данных. Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2018): сборник научных трудов II Международной научно-практической конференции, Том. 1. 14-18 ноября 2018 г. – Донецк, ГОУВПО «Донецкий национальный технический университет», 2018. – С. 44-51

67. Айвазян С.А. и др. Прикладная статистика: Исследование зависимостей: Справ. изд. / С. А. Айвазян, И.С. Енюков, Л. Д. Мешалкин – М.: Финансы и статистика, 1985 – 487 с.

68. Скляров Ю. С. Эконометрика. Краткий курс: учебное пособие. 2-е изд., испр. / Ю. С. Скляров; ГУАП. – СПб., 2007. – 140 с.

69. Рычка, О.В. Разработка и анализ метода повышения точности прогнозных регрессионных моделей и его модификаций // Питання прикладної математики і математичного моделювання: зб. наук. пр. / ред. кол....О. М. Кісельова (голов. ред.) та ін. –Д.: Вид-во Дніпропетр. нац. ун-ту, 2011. – С. 200-212

70. Рычка, О.В. Исследование эффективности применения метода повышения качества прогнозных регрессионных моделей и его модификаций //

Наукові праці Донецького національного технічного університету. Серія «Проблеми моделювання та автоматизації проектування» (МАП-11). Випуск 9 (179): -- Донецьк: ДонНТУ. – 2011. – С. 72-79

71. Григорьев Ю.Д. Методы оптимального планирования эксперимента: линейные модели : учебное пособие - СПб [и др.] : Лань, 2015. - 319 с.

72. Горидько Н. П., Нижегородцев Р. М. Кривые Филлипса для современных макросистем: регрессионный анализ и моделирование - Москва: Восход-А, 2015. - 159 с.

73. Дубровский С. А., Дудина В. А., Садыева Я. В. Методы обработки и анализа экспериментальных данных: учебное пособие – М-во образования и науки Рос. Федерации, Липец. гос. техн. ун-т. - Липецк: Издательство Липецкого государственного технического университета, 2015. - 61 с.

74. Кириченко А. В. (и др.) Математические модели и методы анализа и прогнозирования: предварительная обработка результатов эксперимента, проверка статистических гипотез, корреляционный анализ, парный регрессионный анализ : учебное пособие для студентов всех специальностей – Министерство образования и науки Российской Федерации, Саратовский государственный технический университет имени Гагарина Ю. А. - Саратов : КУБиК, 2019. - 259 с.

75. Кисляков А. Н. Методы и инструменты анализа данных в экономике и управлении : учебно-методическое пособие – Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Владимирский филиал. - Владимир: Владимирский филиал РАНХИГС, 2019. - 161 с.

76. Круценюк К. Ю. Корреляционно-регрессионный анализ в эконометрических моделях: учебное пособие - Министерство науки и высшего образования Российской Федерации, Норильский государственный индустриальный институт. - Норильск: НГИИ, 2018. - 108 с.

77. Ларионова И. А. Статистика. Введение в регрессионный анализ. Временные ряды: учебное пособие - М-во образования и науки РФ, Нац. исслед. технол. ун-т «МИСИС», каф. пром. менеджмента. - Москва: МИСИС, 2016. - 73 с.

78. Мельников, Р. М. Эконометрика: учеб. пособие / Р. М. Мельников. – М.: Проспект, 2014. – 288 с.

79. G. David Garson Partial Least Squares. Regression and Structural Equation Models - Statistical Publishing Associates, 2016 – 261 p.

80. Справочник по прикладной статистике. В 2-х т. Т.1: Пер. с англ./ Под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. − М.: Финансы и статистика, 1989. − 510 с.

81. Справочник по специальным функциям с формулами, графиками и математическими таблицами./ Под. ред. М.Абрамовица и Н. Стигана: Пер. с англ. под ред. В.А. Диткина и Л.И. Кармазиной. − М.: Наука, 1979. − 830 с.

82. Новые методы повышения точности прогнозных регрессионных моделей: монография / О.В. Рычка – LAP Lambert Academic Publishing, 2014. – 61 с.

83. Рычка, О.В. Методы обработки данных для повышения качества прогнозирования при использовании регрессионных моделей // Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: Тези доповідей XIX міжнародної науково-практичної конференції, Ч.IV – м.Харків – 01-03 червня 2011 р. – С.76-79

84. Рычка, О.В. Анализ эффективности новых методов для повышения точности регрессионных прогнозных моделей // Проблеми інформатики і моделювання. Тезиси одинадцятої міжнародної науково-технічної конференції. – Харків: НТУ «ХПІ», 2011. – С.6

85. Рычка, О.В. Повышение эффективности прогнозирования при использовании линейных регрессионных моделей // Моделирование и компьютерная графика – 2011: материалы 4-й международной научно-технической конференции. – г. Донецк – 5-8 октября 2011 г. – С. 213-217

86. Тихонов В.Н. Статистическая радиотехника. Изд. 2-е, перераб. и доп. – М.: Радио и связь, 1982. – 624 с.

87. Технология разработки прикладного программного обеспечения: учебное пособие/ С. В. Соловьев, Л. С. Гринкруг, Р. И. Цой; М-во образования и

науки Российской Федерации, Федеральное гос. бюджетное образовательное учреждение высш. проф. образования «Дальневосточная гос. социально-гуманитарная акад.», Каф. технических дисциплин. - Москва: Акад. естествознания, 2011. - 407 с.

88. Троелсен Эндрю, Джепикс Филипп. Язык программирования C# 7 и платформы .NET и .NET Core, 8-е изд.: Пер. с англ. – СПб.: ООО “Диалектика”, 2018 — 1328 с.:

89. Гриффитс Иэн. Программируем на C# 8.0. Разработка приложений. – СПб.: Питер, 2021. – 944 с.

90. Слепцова Л.Д. Программирование на VBA в Microsoft Office 2010. – М.: Вильямс, 2010. – 432 с.

91. Рычка, О.В. Описание и программная реализации методов обработки данных для повышения точности прогнозирования // Научный журнал «Информатика и кибернетика», № 1(3). –Донецк, ДонНТУ, 2016. – С.92-98.

92. Ричка, О.В. Метод збільшення точності прогнозних регресійних моделей з можливістю застосування в сучасних комп'ютерних технологіях // Донбас-2020: перспективи розвитку очима молодих вчених: матеріали V науково-практичної конференції — м. Донецьк, 25-27 травня 2010 р. – Донецьк, ДонНТУ, 2010 – С.476-479

93. Григорьев А.В., Рычка О.В. Программная реализация алгоритмов методов поиска и обработки аномальных измерений. Современные тенденции развития и перспективы внедрения инновационных технологий в машиностроении, образовании и экономике: материалы и доклады VIII Международной научно-практической конференции, Т7. № 1 (6). 26-29 мая 2021 г. – Азов, 2021. – С. 111-116.

94. Рычка, О.В. Разработка алгоритма реализации методов повышения качества регрессионных моделей, используемых при проектировании технических систем. // Научный журнал «Информатика и кибернетика» № 3 (21), 2020, Донецк, ДонНТУ. - С.42-48.

95. Федеральная служба государственной статистики [Электронный ресурс]: официальный сайт – Режим доступа: http://www.gks.ru

96. Кочегарова О.С., Лажаунинкас Ю.В. Статистическое исследование зависимости урожайности яровой пшеницы от количества внесенных минеральных удобрений с использованием информационных технологий // Информационные технологии естественных и математических наук / Сборник научных трудов по итогам международной научно-практической конференции. № 2. г. Ростов-на-Дону, 2015. – С. 21-24.

97. Рычка, О.В. Анализ эффективности усовершенствованных методов поиска и обработки аномалий для нелинейных моделей с внутренней линейностью // Международный рецензируемый научно-теоретический журнал «Проблемы искусственного интеллекта». – Донецк, 2020. – Выпуск №3(18) – С. 101-110.

98. И.Я. Чуракова. Направления использования методик выявления аномальных наблюдений при решении задач операционного менеджмента. Научный доклад № 13 (R)–2010. СПб.: ВШМ СПбГУ, 2010.

ПРИЛОЖЕНИЕ А

Таблицы критических значений критериев

Таблица А.1 – Критические значения $T_1(\alpha)$ и $T_2(\alpha)$ статистик Граббса

n	Доверительная вероятность α								
	0,90			0,95			0,99		
	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3
3	1,497	1,406	1,818	1,738	1,412	2,121	2,215	1,414	2,712
4	1,696	1,645	1,943	1,941	1,689	2,234	2,431	1,710	2,806
5	1,835	1,791	2,036	2,080	1,869	2,319	2,574	1,917	2,877
6	1,939	1,894	2,111	2,184	1,996	2,386	2,679	2,067	2,934
7	2,022	1,974	2,172	2,267	2,093	2,442	2,761	2,182	2,981
8	2,091	2,041	2,224	2,334	2,172	2,490	2,828	2,273	3,022
9	2,150	2,097	2,269	2,392	2,237	2,531	2,884	2,349	3,057
10	2,200	2,146	2,309	2,441	2,294	2,568	2,931	2,414	3,089
11	2,245	2,190	2,344	2,484	2,343	2,601	2,973	2,470	3,117
12	2,284	2,229	2,376	2,523	2,387	2,630	3,010	2,519	3,143
13	2,320	2,264	2,406	2,557	2,426	2,657	0,043	2,562	3,166
14	2,352	2,297	2,432	2,589	2,461	2,682	3,072	2,602	3,187
15	2,382	2,326	2,457	2,617	2,493	2,705	3,099	2,638	3,207
16	2,409	2,354	2,480	2,644	2,523	2,726	3,124	2,670	3,226
17	2,434	2,380	2,502	2,668	2,551	2,746	3,147	2,701	3,243
18	2,458	2,404	2,522	2,691	2,577	2,765	3,168	2,728	3,259
19	2,480	2,426	2,541	2,712	2,600	2,783	3,188	2,754	3,275
20	2,500	2,447	2,559	2,732	2,623	2,799	3,207	2,778	3,289
21	2,529	2,467	2,576	2,750	2,644	2,815	3,224	2,801	3,303
22	2,538	2,486	2,592	2,768	2,664	2,830	3,240	2,823	3,316
23	2,555	2,504	2,607	2,784	12,683	2,844	3,255	2,843	3,328
24	2,571	2,520	2,621	2,800	2,701	2,857	3,269	2,862	3,340
25	2,587	2,537	2,635	2,815	2,717	2,870	3,282	2,880	3,351

Таблица А.2 – Критические значения $\tau(\alpha)$ критерия наибольшего абсолютного отклонения

n	Доверительная вероятность α			n	Доверительная вероятность α		
	0,90	0,95	0,99		0,90	0,95	0,99
3	1,412	1,414	1,414	14	2,461	2,602	2,859
4	1,689	1,710	1,728	16	2,523	2,670	2,946
5	1,869	1,917	1,972	18	2,577	2,718	3,017
6	1,996	2,067	2,161	20	2,623	2,779	3,079
7	2,093	2,182	2,310	22	2,664	2,823	3,132
8	2,172	2,273	2,431	24	2,701	2,862	3,179
9	2,238	2,349	2,532	25	2,734	2,897	3,220
10	2,294	2,414	2,616	28	2,764	2,929	3,258
12	2,387	2,519	2,753	30	2,792	2,958	3,291

Таблица А.3 – Критические значения статистики Дэвида

n	m								
	4	5	6	7	8	9	10	11	13
Доверительная вероятность $\alpha = 0{,}90$									
10	1,68	1,92	2,09	2,23	2,33	2,42	2,50	2,56	2,68
11	1,66	1,90	2,07	2,20	2,30	2,39	2,46	2,53	2,64
12	1,65	1,88	2,05	2,17	2,28	2,36	2,44	2,50	2,61
13	1,63	1,86	2,03	2,16	2,26	2,34	2,41	2,47	2,58
14	1,62	1,85	2,01	2,14	2,24	2,32	2,39	2,45	2,56
15	1,61	1,84	2,00	2,12	2,22	2,31	2,38	2,44	2,54
16	1,61	1,83	1,99	2,11	2,21	2,29	2,36	2,42	2,52
17	1,60	1,82	1,98	2,10	2,20	2,28	2,35	2,41	2,51
18	1,59	1,82	1,97	2,09	2,19	2,26	2,34	2,39	2,49
19	1,59	1,81	1,96	2,08	2,18	2,26	2,33	2,38	2,48
20	1,58	1,80	1,96	2,08	2,17	2,25	2,32	2,37	2,47
24	1,57	1,78	1,94	2,05	2,15	2,22	2,29	2,34	2,44
30	1,55	1,77	1,92	2,03	2,12	2,20	2,26	2,32	2,41
40	1,54	1,75	1,90	2,01	2,10	2,17	2,23	2,29	2,38
60	1,52	1,73	1,87	1,98	2,07	2,14	2,20	2,26	2,35
120	1,51	1,71	1,85	1,96	2,05	2,12	2,18	2,23	2,32
∞	1,50	1,70	1,83	1,94	2,02	2,09	2,15	2,20	2,28
Доверительная вероятность $\alpha = 0{,}95$									
10	2,01	2,27	2,46	2,60	2,72	2,81	2,89	2,96	3,08
11	1,98	2,24	2,42	2,56	2,67	2,76	2,84	2,91	3,03
12	1,96	2,21	2,39	2,52	2,63	2,72	2,80	2,87	2,98
13	1,94	2,19	2,35	2,50	2,60	2,69	2,76	2,83	2,94
14	1,93	2,17	2,34	2,47	2,57	2,66	2,74	2,80	2,91
15	1,91	2,15	2,32	2,45	2,55	2,64	2,71	2,77	2,88
16	1,90	2,14	2,31	2,43	2,53	2,62	2,69	2,75	2,86
17	1,89	2,13	2,29	2,42	2,52	2,60	2,67	2,73	2,84
18	1,88	2,11	2,28	2,40	2,50	2,58	2,65	2,71	2,82
19	1,87	2,11	2,27	2,39	2,49	2,57	2,64	2,70	2,80
20	1,87	2,10	2,26	2,38	2,47	2,56	2,63	2,68	2,78
24	1,84	2,07	2,23	2,34	2,44	2,52	2,58	2,64	2,74
30	1,82	2,04	2,20	2,31	2,40	2,48	2,54	2,60	2,69
40	1,80	2,02	2,17	2,28	2,37	2,44	2,50	2,56	2,65
60	1,78	1,99	2,14	2,25	2,33	2,41	2,47	2,52	2,61
120	1,76	1,96	2,11	2,22	2,30	2,37	2,43	2,48	2,57
∞	1,74	1,94	2,08	2,18	2,27	2,33	2,39	2,44	2,52
Доверительная вероятность $\alpha = 0{,}99$									
10	2,78	3,10	3,32	3,48	3,62	3,73	3,82	3,90	4,04
11	2,72	3,02	3,24	3,39	3,52	3,63	3,72	3,79	3,93
12	2,67	2,96	3,17	3,32	3,45	3,55	3,64	3,71	3,84
13	2,63	2,92	3,12	3,27	3,38	3,48	3,57	3,64	3,76
14	2,60	2,88	3,07	3,22	3,33	3,43	3,51	3,58	3,70
15	2,57	2,84	3,03	3,17	3,29	3,38	3,46	3,53	3,65
16	2,54	2,81	3,00	3,14	3,25	3,34	3,42	3,49	3,60
17	2,52	2,79	2,97	3,11	3,22	3,31	3,38	3,45	3,56
18	2,50	2,77	2,95	3,08	3,19	3,28	3,35	3,42	3,53
19	2,49	2,75	2,93	3,06	3,16	3,25	3,33	3,39	3,50
20	2,47	2,73	2,91	3,04	3,14	3,23	3,30	3,37	3,47
24	2,42	2,68	2,84	2,97	3,07	3,16	3,23	3,29	3,38
30	2,38	2,62	2,79	2,91	3,01	3,08	3,15	3,21	3,30
40	2,34	2,57	2,73	2,85	2,94	3,02	3,08	3,13	3,22
60	2,29	2,52	2,68	2,79	2,88	2,95	3,01	3,06	3,15
120	2,25	2,48	2,62	2,73	2,82	2,89	2,95	3,00	3,08
∞	2,22	2,43	2,57	2,68	2,76	2,83	2,88	2,93	3,01

Таблица А.4 – Критические значения критерия Ирвина

n	Доверительная вероятность α			n	Доверительная вероятность α		
	0,90	0,95	0,99		0,90	0,95	0,99
2	2,33	2,77	3,64	70	0,84	1,05	1,53
3	1,79	2,17	2,90	80	0,83	1,04	1,50
10	1,18	1,46	2,03	90	0,82	1,03	1,50
20	1,03	1,27	1,80	100	0,81	1,02	1,47
30	0,96	1,20	1,70	200	0,75	0,95	1,38
40	0,91	1,15	1,63	300	0,72	0,91	1,32
50	0,88	1,11	1,60	500	0,68	0,87	1,28
60	0,86	1,08	1,57	1000	0,65	0,83	1,22

Таблица А.5 – Критические значения статистики Диксона

n	Доверительная вероятность α			n	Доверительная вероятность α		
	0,90	0,95	0,99		0,90	0,95	0,99
6	0,482	0,560	0,698	13	0,515	0,570	0,670
	0,609	0,689	0,805	14	0,294	0,349	0,450
	0,745	0,824	0,925		0,336	0,395	0,502
	0,670	0,736	0,836		0,369	0,432	0,542
	0,821	0,872	0,951		0,395	0,445	0,538
	0,965	0,983	0,995		0,448	0,501	0,593
7	0,434	0,507	0,637		0,492	0,546	0,641
	0,530	0,610	0,740	15	0,285	0,338	0,438
	0,636	0,712	0,836		0,323	0,381	0,486
	0,596	0,661	0,778		0,354	0,416	0,523
	0,725	0,780	0,885		0,382	0,430	0,522
	0,850	0,881	0,945		0,431	0,483	0,574
8	0,399	0,468	0,590		0,454	0,507	0,595
	0,479	0,554	0,683	16	0,277	0,320	0,426
	0,557	0,632	0,760		0,313	0,359	0,472
	0,545	0,607	0,710		0,341	0,388	0,508
	0,650	0,710	0,829		0,370	0,406	0,508
	0,745	0,803	0,890		0,416	0,433	0,557
9	0,370	0,437	0,555		0,454	0,490	0,595
	0,441	0,512	0,635	17	0,269	0,313	0,416
	0,504	0,580	0,701		0,303	0,349	0,460
	0,505	0,565	0,667		0,330	0,377	0,493
	0,594	0,657	0,776		0,359	0,397	0,495
	0,676	0,737	0,840		0,403	0,440	0,542
10	0,349	0,412	0,527		0,438	0,475	0,577
	0,409	0,477	0,597	18	0,263	0,306	0,407
	0,454	0,537	0,655		0,295	0,341	0,449
	0,474	0,531	0,632		0,320	0,367	0,480
	0,551	0,612	0,726		0,350	0,379	0,484
	0,620	0,682	0,791		0,391	0,428	0,529
11	0,332	0,392	0,502		0,424	0,462	0,561
	0,385	0,450	0,566	19	0,258	0,306	0,398
	0,431	0,502	0,619		0,288	0,341	0,439
	0,449	0,504	0,603		0,311	0,367	0,469
	0,517	0,576	0,679		0,341	0,379	0,473
	0,578	0,637	0,745		0,380	0,428	0,517
12	0,318	0,376	0,482		0,412	0,462	0,547
	0,367	0,428	0,541	20	0,252	0,300	0,391
	0,406	0,473	0,590		0,282	0,334	0,430
	0,429	0,481	0,579		0,303	0,358	0,458
	0,490	0,546	0,642		0,333	0,372	0,464
	0,543	0,600	0,704		0,371	0,419	0,506
13	0,305	0,361	0,465		0,401	0,450	0,535
	0,350	0,410	0,520	21	0,247	0,295	0,384
	0,387	0,451	0,554		0,276	0,327	0,421
	0,411	0,461	0,557		0,296	0,349	0,449
	0,467	0,521	0,615		0,326	0,365	0,455

Таблица А.6 – Критические значения статистики Хоглина-Иглевича

n	Доверительная вероятность α						n	Доверительная вероятность α					
	0,90			0,95				0,90			0,95		
	l_1	l_2	l_3	l_1	l_2	l_3		l_1	l_2	l_3	l_1	l_2	l_3
7	2,3	1,7	1,5	3,0	2,3	2,0	17	2,2	1,8	1,7	2,6	2,1	2,0
8	1,8	1,6	1,4	2,2	2,1	1,8	18	2,0	1,9	1,7	2,3	2,1	2,4
9	2,7	1,7	1,4	3,3	2,1	1,8	19	2,2	1,9	1,8	2,6	2,3	2,2
10	2,0	1,8	1,5	2,4	0,2	1,8	20	1,9	1,8	2,2	2,3	2,1	2,1
11	2,2	1,8	1,7	2,7	2,2	2,1	30	2,0	1,9	1,9	2,2	2,2	2,1
12	1,8	1,8	1,6	2,2	2,1	2,0	40	2,0	2,0	1,9	2,2	2,2	2,2
13	2,3	1,8	1,6	2,8	2,2	1,9	50	2,0	2,0	1,9	2,2	2,2	2,2
14	2,0	1,9	1,7	2,3	2,2	2,0	75	2,1	2,0	2,0	2,3	2,2	2,2
15	2,1	1,8	1,7	2,5	2,2	2,1	100	2,1	2,0	2,0	2,2	2,2	2,2
16	1,9	1,9	1,7	2,3	2,2	2,1	200	2,2	2,2	2,2	2,4	2,4	2,4

Таблица А.7 – Критические значения статистики Титьена-Мура

n	k									
	1	2	3	4	5	6	7	8	9	10
Доверительная вероятность $\alpha = 90$										
3	0,003									
4	0,005	0,002								
5	0,127	0,022								
6	0,204	0,056	0,009							
7	0,268	0,094	0,027							
8	0,328	0,137	0,053	0,016						
9	0,377	0,175	0,080	0,032						
10	0,420	0,214	0,108	0,052	0,022					
11	0,449	0,250	0,138	0,073	0,036					
12	0,485	0,278	0,162	0,094	0,052	0,026				
13	0,510	0,309	0,189	0,116	0,068	0,038				
14	0,538	0,337	0,216	0,138	0,085	0,052	0,029			
15	0,558	0,360	0,240	0,160	0,105	0,067	0,040			
16	0,578	0,384	0,263	0,182	0,122	0,082	0,053	0,032		
17	0,594	0,406	0,284	0,198	0,140	0,095	0,064	0,042		
18	0,610	0,424	0,304	0,217	0,156	0,110	0,076	0,051	0,034	
19	0,629	0,442	0,322	0,234	0,172	0,124	0,089	0,062	0,042	
20	0,644	0,460	0,336	0,252	0,188	0,138	0,102	0,072	0,051	0,035
25	0,693	0,528	0,417	0,331	0,264	0,210	0,168	0,132	0,103	0,080
30	0,730	0,582	0,475	0,391	0,325	0,270	0,224	0,186	0,154	0,126
35	0,763	0,624	0,523	0,443	0,379	0,324	0,276	0,236	0,202	0,172
40	0,784	0,657	0,562	0,486	0,422	0,367	0,320	0,278	0,243	0,212
45	0,803	0,684	0,593	0,522	0,459	0,406	0,360	0,320	0,284	0,252
50	0,820	0,708	0,622	0,552	0,492	0,440	0,396	0,355	0,319	0,287

Таблица А.8 – Критические значения статистики Роснера

n	k	i	τ^*_{1i}	n	k	i	τ^*_{1i}	n	k	i	τ^*_{1i}	n	k	i	τ^*_{1i}	n	k	i	τ^*_{1i}
Доверительная вероятность $\alpha = 90$																			
20	2	1	2,69	30	4	2	2,65	40	5	4	2,46	60	3	3	2,64	80	5	1	3,44
20	2	2	2,41	30	4	3	2,48	40	5	5	2,39	60	4	1	3,31	80	5	2	2,98
20	3	1	2,76	30	4	4	2,39	50	2	1	3,10	60	4	2	2,85	80	5	3	2,77
20	3	2	2,47	30	5	1	3,05	50	2	2	2,72	60	4	3	2,67	80	5	4	2,63
20	3	3	2,34	30	5	2	2,67	50	3	1	3,18	60	4	4	2,54	80	5	5	2,54
20	4	1	2,81	30	5	3	2,51	50	3	2	2,76	60	5	1	3,34	100	2	1	3,34
20	4	2	2,51	30	5	4	2,42	50	3	3	2,58	60	5	2	2,77	100	2	2	2,92
20	4	3	2,38	30	5	5	2,35	50	4	1	3,24	60	5	3	2,68	100	3	1	3,44
20	4	4	2,29	40	2	1	3,01	50	4	2	2,81	60	5	4	2,56	100	3	2	2,97
20	5	1	2,85	40	2	2	2,72	50	4	3	2,62	60	5	5	2,48	100	3	3	2,77
20	5	2	2,55	40	3	1	3,07	50	4	4	2,50	80	2	1	3,28	100	4	1	3,47
20	5	3	2,40	40	3	2	2,69	50	5	1	3,28	80	2	2	2,85	100	4	2	3,00
20	5	4	2,33	40	3	3	2,52	50	5	2	2,84	80	3	1	3,32	100	4	3	2,79
20	5	5	2,27	40	4	1	3,14	50	5	3	2,65	80	3	2	2,90	100	4	4	2,66
30	2	1	2,89	40	4	2	2,74	50	5	4	2,52	80	3	3	2,71	100	5	1	3,54
30	2	2	2,55	40	4	3	2,57	50	5	5	2,44	80	4	1	3,40	100	5	2	3,04
30	3	1	2,97	40	4	4	2,45	60	2	1	3,15	80	4	2	2,93	100	5	3	2,81
30	3	2	2,61	40	5	1	3,16	60	2	2	2,77	80	4	3	2,74	100	5	4	2,68
30	3	3	2,44	40	5	2	2,76	60	3	1	3,26	80	4	4	2,61	100	5	5	2,59
30	4	1	3,02	40	5	3	2,59	60	3	2	2,83								

Таблица А.9 – Критические значения критерия Эктона

n	α		n	α		n	α	
	0,95	0,99		0,95	0,99		0,95	0,99
3	123	31,4	7	3,98	5,88	15	3,34	4,22
4	7,17	16,27	8	3,77	5,33	20	3,28	4,02
5	5,05	9	9	3,63	4,98	25	3,26	3,94
6	4,34	6,85	10	3,54	4,75			

Таблица А.10 – Критические значения критерия Титьена-Мура-Бэкмана

n	Доверительная вероятность α			n	Доверительная вероятность α		
	0,90	0,95	0,99		0,90	0,95	0,99
4	1,41	1,41	1,41	16	2,50	2,64	2,92
5	1,69	1,71	1,73	18	2,56	2,71	2,99
6	1,88	1,92	1,97	20	2,60	2,76	3,06
7	2,01	2,07	2,16	24	2,69	2,85	3,17
8	2,10	2,19	2,31	30	2,79	2,97	3,28
9	2,18	2,28	2,43	36	2,86	3,03	3,35
10	2,24	2,35	2,53	48	2,97	3,15	3,41
11	2,30	2,43	2,64	60	3,04	3,21	3,50
12	2,35	2,48	2,70	100	3,22	3,40	3,75
14	2,43	2,57	2,80				

Таблица А.11 – Критическое значение критерия Прескотта-Лунда

n	k							
	1	2	3	4	5	6	8	10
Доверительная вероятность $\alpha = 0{,}90$								
5	1,87							
6	2,00	1,89						
7	2,10	2,02	1,90					
8	2,18	2,12	2,03	1,91				
9	2,24	2,20	2,13	2,05	1,92			
10	2,30	2,26	2,21	2,15	2,06	1,92		
12	2,39	2,37	2,33	2,29	2,24	2,17	1,93	
14	2,47	2,45	2,42	2,39	2,36	2,31	2,19	1,94
16	2,53	2,51	2,50	2,47	2,45	2,42	2,34	2,23
18	2,58	2,57	2,56	2,54	2,52	2,50	2,44	2,35
20	2,63	2,62	2,61	2,59	2,58	2,56	2,52	2,46
25	2,72	2,72	2,71	2,70	2,69	2,68	2,66	2,63
30	2,80	2,79	2,77	2,78	2,77	2,77	2,75	2,73
35	2,86	2,85	2,85	2,85	2,84	2,84	2,82	2,81
40	2,91	2,91	2,90	2,90	2,90	2,89	2,88	2,87
45	2,95	2,95	2,95	2,95	2,94	2,94	2,93	2,93
50	2,99	2,99	2,99	2,99	2,98	2,98	2,98	2,97
60	3,06	3,06	3,05	3,05	3,05	3,05	3,05	3,04
70	3,11	3,11	3,11	3,11	3,11	3,11	3,10	3,10
80	3,16	3,16	3,15	3,15	3,15	3,15	3,15	3,15
90	3,20	3,20	3,19	3,19	3,19	3,19	3,19	3,19
100	3,23	3,23	3,23	3,23	3,23	3,23	3,23	3,22

ПРИЛОЖЕНИЕ Б

Листинг программы

Листинг программы на C# для ввода данных из файла .xls

```
private void
изExcelФайлаToolStripMenuItem_Click(object
sender, EventArgs e)
        {
            OpenFileDialog ofd using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;
using System.Reflection;
using ExcelObj =
Microsoft.Office.Interop.Excel;
using System.Threading;
namespace WindowsFormsApp1
{
    public partial class Form1 : Form
    {
        public Form1()
        {
            InitializeComponent();
        }
  private void
//выбор пункта меню «Из Excel файла»
изExcelФайлаToolStripMenuItem_Click(object
sender, EventArgs e)
        {
            OpenFileDialog ofd = new
OpenFileDialog();
//Задаем расширение имени файла по умолчанию
            ofd.DefaultExt = "*.xls;*.xlsx";
//Задаем строку фильтра имен файлов
            ofd.Filter = "Excel
Sheet(*.xlsx)|*.xlsx;*.xls";
 //Задаем заголовок диалогового окна.
            ofd.Title = "Выберите документ
для загрузки данных";
            ExcelObj.Application app = new
ExcelObj.Application();
            ExcelObj.Workbook workbook;
            ExcelObj.Worksheet NwSheet;
            ExcelObj.Range ShtRange;
            DataTable dt = new DataTable();
            if (ofd.ShowDialog() ==
DialogResult.OK)
            {
                textBox1.Text =
ofd.FileName;

                workbook =
app.Workbooks.Open(ofd.FileName,
Missing.Value,
                Missing.Value,
Missing.Value, Missing.Value, Missing.Value,
                Missing.Value,
Missing.Value, Missing.Value, Missing.Value,
                Missing.Value,
Missing.Value, Missing.Value, Missing.Value,
                Missing.Value);
 //Устанавливаем номер листа из котрого
будут извлекаться данные
NwSheet =
(ExcelObj.Worksheet)workbook.Sheets.get_Item
(1);
 ShtRange = NwSheet.UsedRange;
                for (int Cnum = 1; Cnum <=
ShtRange.Columns.Count; Cnum++)
                {
                    dt.Columns.Add(
                       new
DataColumn((ShtRange.Cells[1, Cnum] as
ExcelObj.Range).Value2.ToString()));
                }
                dt.AcceptChanges();

                string[] columnNames = new
String[dt.Columns.Count];
                for (int i = 0; i <
dt.Columns.Count; i++)
                {
                    columnNames[0] =
dt.Columns[i].ColumnName;
                }

                for (int Rnum = 2; Rnum <=
ShtRange.Rows.Count; Rnum++)
                {
                    DataRow dr =
dt.NewRow();
                    for (int Cnum = 1; Cnum
<= ShtRange.Columns.Count; Cnum++)
                    {
                        if
((ShtRange.Cells[Rnum, Cnum] as
ExcelObj.Range).Value2 != null)
                        {
                            dr[Cnum - 1] =
                (ShtRange.Cells[Rnum, Cnum]
as ExcelObj.Range).Value2.ToString();
                        }
                    }
                    dt.Rows.Add(dr);
                    dt.AcceptChanges();
                }

                dataGridView1.DataSource =
dt;
                app.Quit();
            }
            else
                Application.Exit();
        }
```

Листинг программы на C# при нажатии кнопки «Весь метод»

```
private void
allMethod1ToolStripMenuItem_Click(object
sender, EventArgs e)
        {
Microsoft.Office.Interop.Excel._Application
app = new
Microsoft.Office.Interop.Excel.Application()
;
Microsoft.Office.Interop.Excel._Workbook
//открытие макроса Excel с расчетами
workbook =
app.Workbooks.Open(@"D:\Метод1_New!!!.xlsm")
;
Microsoft.Office.Interop.Excel._Worksheet
worksheet = null;
app.Visible = false;
worksheet = workbook.Sheets["100"];
worksheet = workbook.ActiveSheet;
System.Diagnostics.Process excelProc =
System.Diagnostics.Process.GetProcessesByNam
e("EXCEL").Last();
            for (int i = 0; i <
dataGridView1.Rows.Count; i++)
   { for (int j = 0; j <
dataGridView1.Columns.Count; j++)
    { if
(dataGridView1.Rows[i].Cells[j].Value !=
null)
     {
     worksheet.Cells[i + 3, j + 1] =
dataGridView1.Rows[i].Cells[j].Value;
                    }
                    else
                    {
                        worksheet.Cells[i +
3, j + 1] = "";
                    }
                }
            }
            // Запускаем макрос
            app.Run("macros1");
            app.ScreenUpdating = true;
            tabControl1.Visible = true;
            ExcelObj.Application app1 = new
ExcelObj.Application();
Microsoft.Office.Interop.Excel._Workbook
workbook1 =
app1.Workbooks.Open(@"D:\Temp.xls");

Microsoft.Office.Interop.Excel._Worksheet
worksheet1 = null;
            app1.Visible = false;
            worksheet1 =
workbook1.Sheets["Itog"];
            ExcelObj.Range ShtRange;
            ExcelObj.Range ShtRange1;
            DataTable dt = new DataTable();
            ShtRange = worksheet1.UsedRange;
            ShtRange1 =
worksheet1.UsedRange;
dataGridView2.ColumnHeadersDefaultCellStyle.
BackColor = Color.Gray;
            for (int j = 1; j <= 10; j++)
            {
                dataGridView2.Rows.Add();
            }
            //добавление в таблицу данных с
1-й по 7 колонку
            for (int x = 1; x <= 7; x++)
            {

                //добавление данных, начиная
со 2-й строки (Excel) по 12
                for (int y = 2; y <= 10;
y++)
                {

                    if ((ShtRange.Cells[y,
x] as ExcelObj.Range).Value2 != null)
                    {
                        dataGridView2.Rows[y
- 2].Cells[x - 1].Value =
worksheet1.Cells[y, x].Value2.ToString();
                    }
                }
            }
            //
customersDataGridView3.Columns[0].Visible =
false;
            dataGridView4.Visible = false;
            //создание в гриде таблицы с 1-й
строки до последней заполненной в Excel

            for (int j = 15; j <=
ShtRange1.Rows.Count; j++)
            {
                dataGridView3.Rows.Add();
            }
            //добавление в таблицу данных с
1-й по 16 колонку
            for (int x = 1; x <= 16; x++)
            {

                //добавление данных, начиная
со 2-й строки (Excel) по 12
                for (int y = 15; y <=
ShtRange1.Rows.Count; y++)
                {

                    if ((ShtRange1.Cells[y,
x] as ExcelObj.Range).Value2 != null)
                    {
                        dataGridView3.Rows[y
- 15].Cells[x - 1].Value =
worksheet1.Cells[y, x].Value2.ToString();
                    }
                }
            }

            workbook1.Close();
            app1.Quit();
            app.Application.Quit();
            //excelProc.Kill();

                        }
```

Листинг фрагмента программы на VBA для заполнения листов расчетов

```
Public koefa As Single
Dim koefb As Single
Public kol As Integer
Public kolvo As Integer
Private Sub CommandButton1_Click()
Dim num As Single
Dim rwIndex1 As Integer
Dim rwIndex2 As Integer
Dim rwIndex3 As Integer
Dim rwIndex55 As Integer
Dim num1 As Single
Dim koef As Single
Dim koefap As Single
Dim koefbp As Single
Dim koef1 As Single
Dim eq As Single
Dim eq95 As Single
Dim R1 As Single
Dim R2 As Single
Dim R As Single
Dim num95 As Single
Dim num951 As Single
Dim mox95 As Single
Dim moy95 As Single
Dim koef95 As Single
Dim koef951 As Single
Dim koefa95 As Single
Dim koefb95 As Single
Dim A5 As String
kolv = 0
i = 1
   While Cells(i, 1).Value <> ""
      kolv = kolv + 1
      i = i + 1
   Wend
kol = kolv - 2
Cells(1, 2) = kol

For rwIndex55 = 3 To kol + 2
   For j = 1 To 2
   A5 = Cells(rwIndex55, j)
   If Not IsNumeric(A5) Then If MsgBox("Ошибка!
Повторить ввод?", vbOK) = vbOK Or vbCancel Then
Exit Sub
   Next j
Next rwIndex55

Range(Cells(3, 1), Cells(2 + kol, 2)).Select
   Range(Cells(3, 1), Cells(2 + kol, 2)).Sort
Key1:=Range("A3"), Order1:=xlAscending,
Header:=xlGuess, OrderCustom:=1, MatchCase:=False,
Orientation:=xlTopToBottom, _
      DataOption1:=xlSortNormal
num = 0
   For rwIndex = 3 To 3 + kol - 1
      num = num + Cells(rwIndex, 1)
   Next rwIndex
Cells(2, 1) = "X"
Cells(2, 2) = "Y"
Cells(2, 4) = "Мат. ожидание"
mox = num / kol
Cells(2, 7) = num / kol
Cells(4 + kol, 1) = "Коэф-ты ур-я ax+b"
Cells(4 + kol + 1, 1) = "a"
Cells(4 + kol + 1, 2) = "b"
num1 = 0
   For rwIndex = 3 To 3 + kol - 1
      num1 = num1 + Cells(rwIndex, 2)
   Next rwIndex
moy = num1 / kol
koef = 0
koef1 = 0
   For rwIndex = 3 To 3 + kol - 1
      koef = koef + ((Cells(rwIndex, 1) - mox) *
(Cells(rwIndex, 2) - moy))
      koef1 = koef1 + ((Cells(rwIndex, 1) - mox) *
(Cells(rwIndex, 1) - mox))
   Next rwIndex
koefa = koef / koef1
Cells(4 + kol + 2, 1) = koefa
koefb = moy - koefa * mox
Cells(4 + kol + 2, 2) = koefb
Cells(4 + kol, 7) = "Коэф-ты перпендикуляра"
Cells(4 + kol + 1, 7) = "a"
Cells(4 + kol + 1, 8) = "b"
koefap = -1 / koefa
Cells(4 + kol + 2, 7) = koefap
koefbp = moy - koefap * mox
Cells(4 + kol + 2, 8) = koefbp
Cells(4 + kol + 4, 1) = "Yпрогн"
If Cells(2, 9).Value <> "" Then
xpr = Cells(2, 9).Value
Else: xpr = Cells(kol + 1, 1)
Cells(2, 9) = xpr
End If
ypr = koefa * xpr + koefb
Cells(4 + kol + 4, 2) = ypr
Cells(1, 11) = kol - 2
Cells(4 + kol + 6, 1) = "Ур-е"
Cells(4 + kol + 6, 2) = "e"
Cells(4 + kol + 6, 3) = "e^2"
   For rwIndex = 3 To 3 + kol - 1
      For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
         eq = koefa * Cells(rwIndex, 1) + koefb
         Cells(rwIndex1, 1) = eq
         rwIndex = rwIndex + 1
      Next rwIndex1
   Next rwIndex
   For rwIndex = 3 To 3 + kol - 1
      For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
         e = Cells(rwIndex, 2) - (koefa * Cells(rwIndex, 1)
+ koefb)
         Cells(rwIndex1, 2) = e
         rwIndex = rwIndex + 1
      Next rwIndex1
   Next rwIndex
   For rwIndex = 3 To 3 + kol - 1
      For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
         e = Cells(rwIndex, 2) - (koefa * Cells(rwIndex, 1)
+ koefb)
```

```
        e2 = e * e
        Cells(rwIndex1, 3) = e2
        rwIndex = rwIndex + 1
      Next rwIndex1
   Next rwIndex
Cells(4 + kol + 7, 5) = "s^2"
   For rwIndex = 4 + kol + 7 To 4 + kol + 7 + kol - 1
      S = S + Cells(rwIndex, 2) * Cells(rwIndex, 2)
   Next rwIndex
s2 = S / (kol - 2)
Cells(4 + kol + 8, 5) = s2
Cells(4 + kol + 7, 6) = "сигма"
sig = (s2 * (1 + 1 / kol)) ^ (1 / 2)
Cells(4 + kol + 8, 6) = sig

Cells(4 + kol + 6, 8) = "Ур-е перп."
Cells(4 + kol + 6, 9) = "e"
Cells(4 + kol + 6, 10) = "e^2"
   For rwIndex = 3 To 3 + kol - 1
      For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
         eqp = koefap * Cells(rwIndex, 1) + koefbp
         Cells(rwIndex1, 8) = eqp
         rwIndex = rwIndex + 1
      Next rwIndex1
   Next rwIndex
   For rwIndex = 3 To 3 + kol - 1
      For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
         ep = Cells(rwIndex, 2) - (koefap * Cells(rwIndex,
1) + koefbp)
         Cells(rwIndex1, 9) = ep
         rwIndex = rwIndex + 1
      Next rwIndex1
   Next rwIndex
   For rwIndex = 3 To 3 + kol - 1
      For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
         ep = Cells(rwIndex, 2) - (koefap * Cells(rwIndex,
1) + koefbp)
         ep2 = ep * ep
         Cells(rwIndex1, 10) = ep2
         rwIndex = rwIndex + 1
      Next rwIndex1
   Next rwIndex
Cells(4 + kol + 7, 12) = "s^2"
   For rwIndex = 4 + kol + 7 To 4 + kol + 7 + kol - 1
      sp = sp + Cells(rwIndex, 9) * Cells(rwIndex, 9)
   Next rwIndex
sp2 = sp / (kol - 2)
Cells(4 + kol + 8, 12) = sp2
Cells(4 + kol + 7, 13) = "сигма"
sigp = (sp2 * (1 + 1 / kol)) ^ (1 / 2)
Cells(4 + kol + 8, 13) = sigp

Cells(4 + kol + 1, 4) = "R^2"
R1 = 0
R2 = 0
R = 0
   For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol - 1
      R1 = R1 + ((Cells(rwIndex1, 1) - moy) ^ 2)
   Next rwIndex1
   For rwIndex = 3 To 3 + kol - 1
      R2 = R2 + ((Cells(rwIndex, 2) - moy) ^ 2)
   Next rwIndex
R = R1 / R2
max1 = Cells(4 + kol + 7, 2)
   For rwIndex1 = 4 + kol + 7 + 1 To 4 + kol + 7 + kol -
1
      If Cells(rwIndex1, 2) > max1 Then
      max1 = Cells(rwIndex1, 2)
      End If
   Next rwIndex1
min1 = Cells(4 + kol + 7, 2)
   For rwIndex1 = 4 + kol + 7 + 1 To 4 + kol + 7 + kol -
1
      If Cells(rwIndex1, 2) < min1 Then
      min1 = Cells(rwIndex1, 2)
      End If
   Next rwIndex1
di = (Abs(min1) + max1) / 2 * 100 / ypr
Cells(4 + kol + 2, 4) = R
Cells(4 + kol + 1, 5) = "DI,%"
Cells(4 + kol + 2, 5) = di
Worksheets("Itog").Cells(2, 2) = R
Worksheets("Itog").Cells(2, 3) = di

Cells(6, 5) = "Вероятность попадания в интервал"
Cells(6, 8) = "0,9"
Cells(6, 9) = "0,85"
Cells(6, 10) = "0,8"
Cells(6, 11) = "0,75"
Cells(6, 12) = "0,7"
Cells(6, 13) = "0,65"
Cells(6, 14) = "0,6"
Cells(6, 15) = "0,5"
Cells(7, 5) = "Величина сдвига"
Cells(7, 8) = 1.95 * sig
Cells(7, 9) = 1.75 * sig
Cells(7, 10) = 1.6 * sig
Cells(7, 11) = 1.5 * sig
Cells(7, 12) = 1.4 * sig
Cells(7, 13) = 1.3 * sig
Cells(7, 14) = 1.2 * sig
Cells(7, 15) = 1.05 * sig
Cells(8, 8) = 1.95 * sigp
Cells(8, 9) = 1.75 * sigp
Cells(8, 10) = 1.6 * sigp
Cells(8, 11) = 1.5 * sigp
Cells(8, 12) = 1.4 * sigp
Cells(8, 13) = 1.3 * sigp
Cells(8, 14) = 1.2 * sigp
Cells(8, 15) = 1.05 * sigp
kolvo = 0
   For rwIndex2 = 1 To kol
      For rwIndex = 3 To 3 + kol - 2
         For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol -
1
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 1) +
Cells(7, 8).Value) Then
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 8) +
Cells(8, 8).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 1) -
Cells(7, 8).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 8) -
Cells(8, 8).Value) Then
               Worksheets("0,90").Cells(rwIndex2, 1) =
Worksheets("100").Cells(rwIndex, 1)
```

```
            Worksheets("0,90").Cells(rwIndex2, 2) =
Worksheets("100").Cells(rwIndex, 2)
            kolvo = kolvo + 1
        Else
        Worksheets("0,90").Cells(rwIndex, 4) =
Worksheets("100").Cells(rwIndex, 1)
        Worksheets("0,90").Cells(rwIndex, 5) =
Worksheets("100").Cells(rwIndex, 2)
          End If
          End If
          End If
          End If
        rwIndex = rwIndex + 1
        rwIndex2 = rwIndex2 + 1
        Next rwIndex1
      Next rwIndex
   Next rwIndex2
   For rwIndex2 = 1 To kolvo
      If Worksheets("0,90").Cells(rwIndex2, 1).Value =
"" Then
        Worksheets("0,90").Cells(rwIndex2,
1).EntireRow.Delete
        rwIndex2 = rwIndex2 - 1
      End If
   Next rwIndex2
num95 = 0
   For rwIndex2 = 1 To kolvo
      num95 = num95 +
Worksheets("0,90").Cells(rwIndex2, 1)
   Next rwIndex2
mox95 = num95 / kolvo
Worksheets("0,90").Cells(2 + kolvo, 1) = "Коэф-ты ур-я
ax+b"
Worksheets("0,90").Cells(2 + kolvo + 1, 1) = "a"
Worksheets("0,90").Cells(2 + kolvo + 1, 2) = "b"
num951 = 0
   For rwIndex2 = 1 To kolvo
      num951 = num951 +
Worksheets("0,90").Cells(rwIndex2, 2)
   Next rwIndex2
moy95 = num951 / kolvo
koef95 = 0
koef951 = 0
   For rwIndex2 = 1 To kolvo
      koef95 = koef95 +
((Worksheets("0,90").Cells(rwIndex2, 1) - mox95) *
(Worksheets("0,90").Cells(rwIndex2, 2) - moy95))
      koef951 = koef951 +
((Worksheets("0,90").Cells(rwIndex2, 1) - mox95) ^ 2)
   Next rwIndex2
koefa95 = koef95 / koef951
Worksheets("0,90").Cells(2 + kolvo + 2, 1) = koefa95
koefb95 = moy95 - koefa95 * mox95
Worksheets("0,90").Cells(2 + kolvo + 2, 2) = koefb95
Worksheets("0,90").Cells(2 + kolvo + 6, 1) = "Ур-е"
Worksheets("0,90").Cells(2 + kolvo + 6, 2) = "e"
   For rwIndex2 = 1 To kolvo
      For rwIndex1 = 2 + kolvo + 7 To 2 + kolvo + 7 +
kolvo - 1
        eq95 = koefa95 *
Worksheets("0,90").Cells(rwIndex2, 1) + koefb95
        Worksheets("0,90").Cells(rwIndex1, 1) = eq95
        rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
   For rwIndex2 = 1 To kolvo
      For rwIndex1 = 2 + kolvo + 7 To 2 + kolvo + 7 +
kolvo - 1
        e95 = Worksheets("0,90").Cells(rwIndex2, 2) -
(koefa95 * Worksheets("0,90").Cells(rwIndex2, 1) +
koefb95)
        Worksheets("0,90").Cells(rwIndex1, 2) = e95
        rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
Worksheets("0,90").Cells(2 + kolvo + 1, 4) = "R^2"
R195 = 0
R295 = 0
R95 = 0
   For rwIndex1 = 2 + kolvo + 7 To 2 + kolvo + 7 +
kolvo - 1
      R195 = R195 +
((Worksheets("0,90").Cells(rwIndex1, 1) - moy95) ^ 2)
   Next rwIndex1
   For rwIndex = 1 To 1 + kolvo - 1
      R295 = R295 +
((Worksheets("0,90").Cells(rwIndex, 2) - moy95) ^ 2)
   Next rwIndex
R95 = R195 / R295
max951 = Worksheets("0,90").Cells(2 + kolvo + 7, 2)
   For rwIndex1 = 2 + kolvo + 7 + 1 To 2 + kolvo + 7 +
kolvo - 1
      If Worksheets("0,90").Cells(rwIndex1, 2) > max951
Then
      max951 = Worksheets("0,90").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
min951 = Worksheets("0,90").Cells(2 + kolvo + 7, 2)
   For rwIndex1 = 2 + kol + 7 + 1 To 2 + kol + 7 + kol -
1
      If Worksheets("0,90").Cells(rwIndex1, 2) < min951
Then
      min951 = Worksheets("0,90").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
ypr95 = xpr * koefa95 + koefb95
di95 = (Abs(min951) + max951) / 2 * 100 / ypr95
delta95 = Abs(ypr95 - ypr) / ypr * 100
Worksheets("0,90").Cells(2 + kolvo + 2, 4) = R95
Worksheets("0,90").Cells(2 + kolvo + 1, 5) = "DI,%"
Worksheets("0,90").Cells(2 + kolvo + 2, 5) = di95
Worksheets("0,90").Cells(2 + kolvo + 1, 6) = "Delta,%"
Worksheets("0,90").Cells(2 + kolvo + 2, 6) = delta95
Worksheets("0,90").Cells(2 + kolvo + 4, 2) = ypr95

Worksheets("Itog").Cells(2, 1) = "100"
Worksheets("Itog").Cells(3, 1) = "90"
Worksheets("Itog").Cells(4, 1) = "85"
Worksheets("Itog").Cells(5, 1) = "80"
Worksheets("Itog").Cells(6, 1) = "75"
Worksheets("Itog").Cells(7, 1) = "70"
Worksheets("Itog").Cells(8, 1) = "65"
Worksheets("Itog").Cells(9, 1) = "60"
Worksheets("Itog").Cells(10, 1) = "50"
Worksheets("Itog").Cells(1, 2) = "R^2"
Worksheets("Itog").Cells(1, 3) = "DI,%"
```

```
Worksheets("Itog").Cells(1, 4) = "Delta,%"
Worksheets("Itog").Cells(1, 5) = "Количество точек"
Worksheets("Itog").Cells(1, 6) = "Точность"
Worksheets("Itog").Cells(3, 2) = R95
Worksheets("Itog").Cells(3, 3) = di95
Worksheets("Itog").Cells(3, 4) = delta95
Worksheets("Itog").Cells(2, 5) = kol
Worksheets("Itog").Cells(3, 5) = kolvo
Worksheets("Itog").Cells(3, 6) = R95 * kolvo / kol
kolvo85 = 0
   For rwIndex2 = 1 To kol
      For rwIndex = 3 To 3 + kol - 2
         For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol -
1
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 1) +
Cells(7, 9).Value) Then
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 8) +
Cells(8, 9).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 1) -
Cells(7, 9).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 8) -
Cells(8, 9).Value) Then
               Worksheets("0,85").Cells(rwIndex2, 1) =
Worksheets("100").Cells(rwIndex, 1)
               Worksheets("0,85").Cells(rwIndex2, 2) =
Worksheets("100").Cells(rwIndex, 2)
               kolvo85 = kolvo85 + 1
         Else
         Worksheets("0,85").Cells(rwIndex, 4) =
Worksheets("100").Cells(rwIndex, 1)
         Worksheets("0,85").Cells(rwIndex, 5) =
Worksheets("100").Cells(rwIndex, 2)
            End If
            End If
            End If
            End If
         rwIndex = rwIndex + 1
         rwIndex2 = rwIndex2 + 1
         Next rwIndex1
      Next rwIndex
   Next rwIndex2
   For rwIndex2 = 1 To kolvo85
      If Worksheets("0,85").Cells(rwIndex2, 1).Value =
"" Then
         Worksheets("0,85").Cells(rwIndex2,
1).EntireRow.Delete
          rwIndex2 = rwIndex2 - 1
      End If
   Next rwIndex2
num85 = 0
   For rwIndex2 = 1 To kolvo85
      num85 = num85 +
Worksheets("0,85").Cells(rwIndex2, 1)
   Next rwIndex2
mox85 = num85 / kolvo85
Worksheets("0,85").Cells(2 + kolvo85, 1) = "Коэф-ты
ур-я ax+b"
Worksheets("0,85").Cells(2 + kolvo85 + 1, 1) = "a"
Worksheets("0,85").Cells(2 + kolvo85 + 1, 2) = "b"
num851 = 0
   For rwIndex2 = 1 To kolvo85
      num851 = num851 +
Worksheets("0,85").Cells(rwIndex2, 2)
   Next rwIndex2
moy85 = num851 / kolvo85
koef85 = 0
koef851 = 0
   For rwIndex2 = 1 To kolvo85
       koef85 = koef85 +
((Worksheets("0,85").Cells(rwIndex2, 1) - mox85) *
(Worksheets("0,85").Cells(rwIndex2, 2) - moy85))
       koef851 = koef851 +
((Worksheets("0,85").Cells(rwIndex2, 1) - mox85) ^ 2)
   Next rwIndex2
koefa85 = koef85 / koef851
Worksheets("0,85").Cells(2 + kolvo85 + 2, 1) = koefa85
koefb85 = moy85 - koefa85 * mox85
Worksheets("0,85").Cells(2 + kolvo85 + 2, 2) = koefb85
Worksheets("0,85").Cells(2 + kolvo85 + 6, 1) = "Ур-е"
Worksheets("0,85").Cells(2 + kolvo85 + 6, 2) = "e"
   For rwIndex2 = 1 To kolvo85
      For rwIndex1 = 2 + kolvo85 + 7 To 2 + kolvo85 + 7
+ kolvo85 - 1
         eq85 = koefa85 *
Worksheets("0,85").Cells(rwIndex2, 1) + koefb85
         Worksheets("0,85").Cells(rwIndex1, 1) = eq85
         rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
   For rwIndex2 = 1 To kolvo85
      For rwIndex1 = 2 + kolvo85 + 7 To 2 + kolvo85 + 7
+ kolvo85 - 1
         e85 = Worksheets("0,85").Cells(rwIndex2, 2) -
(koefa85 * Worksheets("0,85").Cells(rwIndex2, 1) +
koefb85)
         Worksheets("0,85").Cells(rwIndex1, 2) = e85
         rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
Worksheets("0,85").Cells(2 + kolvo85 + 1, 4) = "R^2"
R185 = 0
R285 = 0
R85 = 0
   For rwIndex1 = 2 + kolvo85 + 7 To 2 + kolvo85 + 7 +
kolvo85 - 1
      R185 = R185 +
((Worksheets("0,85").Cells(rwIndex1, 1) - moy85) ^ 2)
   Next rwIndex1
   For rwIndex = 1 To 1 + kolvo85 - 1
      R285 = R285 +
((Worksheets("0,85").Cells(rwIndex, 2) - moy85) ^ 2)
   Next rwIndex
R85 = R185 / R285
max851 = Worksheets("0,85").Cells(2 + kolvo85 + 7, 2)
   For rwIndex1 = 2 + kolvo85 + 7 + 1 To 2 + kolvo85 +
7 + kolvo85 - 1
      If Worksheets("0,85").Cells(rwIndex1, 2) > max851
Then
      max851 = Worksheets("0,85").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
min851 = Worksheets("0,85").Cells(2 + kolvo85 + 7, 2)
   For rwIndex1 = 2 + kolvo85 + 7 + 1 To 2 + kolvo85 +
7 + kolvo85 - 1
      If Worksheets("0,85").Cells(rwIndex1, 2) < min851
Then
```

```
      min851 = Worksheets("0,85").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
ypr85 = xpr * koefa85 + koefb85
di85 = (Abs(min851) + max851) / 2 * 100 / ypr85
delta85 = Abs(ypr85 - ypr) / ypr * 100
Worksheets("0,85").Cells(2 + kolvo85 + 2, 4) = R85
Worksheets("0,85").Cells(2 + kolvo85 + 1, 5) = "DI,%"
Worksheets("0,85").Cells(2 + kolvo85 + 2, 5) = di85
Worksheets("0,85").Cells(2 + kolvo85 + 1, 6) =
"Delta,%"
Worksheets("0,85").Cells(2 + kolvo85 + 2, 6) = delta85
Worksheets("0,85").Cells(2 + kolvo85 + 4, 2) = ypr85
Worksheets("Itog").Cells(4, 2) = R85
Worksheets("Itog").Cells(4, 3) = di85
Worksheets("Itog").Cells(4, 4) = delta85
Worksheets("Itog").Cells(4, 5) = kolvo85
Worksheets("Itog").Cells(4, 6) = R85 * kolvo85 / kol
kolvo80 = 0
   For rwIndex2 = 1 To kol
      For rwIndex = 3 To 3 + kol - 2
         For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol -
1
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 1) +
Cells(7, 10).Value) Then
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 8) +
Cells(8, 10).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 1) -
Cells(7, 10).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 8) -
Cells(8, 10).Value) Then
               Worksheets("0,80").Cells(rwIndex2, 1) =
Worksheets("100").Cells(rwIndex, 1)
               Worksheets("0,80").Cells(rwIndex2, 2) =
Worksheets("100").Cells(rwIndex, 2)
               kolvo80 = kolvo80 + 1
         Else
         Worksheets("0,80").Cells(rwIndex, 4) =
Worksheets("100").Cells(rwIndex, 1)
         Worksheets("0,80").Cells(rwIndex, 5) =
Worksheets("100").Cells(rwIndex, 2)
            End If
            End If
            End If
            End If
         rwIndex = rwIndex + 1
         rwIndex2 = rwIndex2 + 1
         Next rwIndex1
      Next rwIndex
   Next rwIndex2
   For rwIndex2 = 1 To kolvo80
      If Worksheets("0,80").Cells(rwIndex2, 1).Value =
"" Then
         Worksheets("0,80").Cells(rwIndex2,
1).EntireRow.Delete
         rwIndex2 = rwIndex2 - 1
      End If
   Next rwIndex2
num80 = 0
   For rwIndex2 = 1 To kolvo80
      num80 = num80 +
Worksheets("0,80").Cells(rwIndex2, 1)
   Next rwIndex2
mox80 = num80 / kolvo80
Worksheets("0,80").Cells(2 + kolvo80, 1) = "Коэф-ты
ур-я ax+b"
Worksheets("0,80").Cells(2 + kolvo80 + 1, 1) = "a"
Worksheets("0,80").Cells(2 + kolvo80 + 1, 2) = "b"
num801 = 0
   For rwIndex2 = 1 To kolvo80
      num801 = num801 +
Worksheets("0,80").Cells(rwIndex2, 2)
   Next rwIndex2
moy80 = num801 / kolvo80
koef80 = 0
koef801 = 0
   For rwIndex2 = 1 To kolvo80
      koef80 = koef80 +
((Worksheets("0,80").Cells(rwIndex2, 1) - mox80) *
(Worksheets("0,80").Cells(rwIndex2, 2) - moy80))
      koef801 = koef801 +
((Worksheets("0,80").Cells(rwIndex2, 1) - mox80) ^ 2)
   Next rwIndex2
koefa80 = koef80 / koef801
Worksheets("0,80").Cells(2 + kolvo80 + 2, 1) = koefa80
koefb80 = moy80 - koefa80 * mox80
Worksheets("0,80").Cells(2 + kolvo80 + 2, 2) = koefb80
Worksheets("0,80").Cells(2 + kolvo80 + 6, 1) = "Ур-е"
Worksheets("0,80").Cells(2 + kolvo80 + 6, 2) = "e"
   For rwIndex2 = 1 To kolvo80
      For rwIndex1 = 2 + kolvo80 + 7 To 2 + kolvo80 + 7
+ kolvo80 - 1
         eq80 = koefa80 *
Worksheets("0,80").Cells(rwIndex2, 1) + koefb80
         Worksheets("0,80").Cells(rwIndex1, 1) = eq80
         rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
   For rwIndex2 = 1 To kolvo80
      For rwIndex1 = 2 + kolvo80 + 7 To 2 + kolvo80 + 7
+ kolvo80 - 1
         e80 = Worksheets("0,80").Cells(rwIndex2, 2) -
(koefa80 * Worksheets("0,80").Cells(rwIndex2, 1) +
koefb80)
         Worksheets("0,80").Cells(rwIndex1, 2) = e80
         rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
Worksheets("0,80").Cells(2 + kolvo80 + 1, 4) = "R^2"
R180 = 0
R280 = 0
R80 = 0
   For rwIndex1 = 2 + kolvo80 + 7 To 2 + kolvo80 + 7 +
kolvo80 - 1
      R180 = R180 +
((Worksheets("0,80").Cells(rwIndex1, 1) - moy80) ^ 2)
   Next rwIndex1
   For rwIndex = 1 To 1 + kolvo80 - 1
      R280 = R280 +
((Worksheets("0,80").Cells(rwIndex, 2) - moy80) ^ 2)
   Next rwIndex
R80 = R180 / R280
max801 = Worksheets("0,80").Cells(2 + kolvo80 + 7, 2)
   For rwIndex1 = 2 + kolvo80 + 7 + 1 To 2 + kolvo80 +
7 + kolvo80 - 1
```

```
      If Worksheets("0,80").Cells(rwIndex1, 2) > max801
Then
      max801 = Worksheets("0,80").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
min801 = Worksheets("0,80").Cells(2 + kolvo80 + 7, 2)
   For rwIndex1 = 2 + kolvo80 + 7 + 1 To 2 + kolvo80 +
7 + kolvo80 - 1
      If Worksheets("0,80").Cells(rwIndex1, 2) < min801
Then
      min801 = Worksheets("0,80").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
ypr80 = xpr * koefa80 + koefb80
di80 = (Abs(min801) + max801) / 2 * 100 / ypr80
delta80 = Abs(ypr80 - ypr) / ypr * 100
Worksheets("0,80").Cells(2 + kolvo80 + 2, 4) = R80
Worksheets("0,80").Cells(2 + kolvo80 + 1, 5) = "DI,%"
Worksheets("0,80").Cells(2 + kolvo80 + 2, 5) = di80
Worksheets("0,80").Cells(2 + kolvo80 + 1, 6) =
"Delta,%"
Worksheets("0,80").Cells(2 + kolvo80 + 2, 6) = delta80
Worksheets("0,80").Cells(2 + kolvo80 + 4, 2) = ypr80
Worksheets("Itog").Cells(5, 2) = R80
Worksheets("Itog").Cells(5, 3) = di80
Worksheets("Itog").Cells(5, 4) = delta80
Worksheets("Itog").Cells(5, 5) = kolvo80
Worksheets("Itog").Cells(5, 6) = R80 * kolvo80 / kol
kolvo75 = 0
   For rwIndex2 = 1 To kol
      For rwIndex = 3 To 3 + kol - 2
         For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol -
1
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 1) +
Cells(7, 11).Value) Then
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 8) +
Cells(8, 11).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 1) -
Cells(7, 11).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 8) -
Cells(8, 11).Value) Then
               Worksheets("0,75").Cells(rwIndex2, 1) =
Worksheets("100").Cells(rwIndex, 1)
               Worksheets("0,75").Cells(rwIndex2, 2) =
Worksheets("100").Cells(rwIndex, 2)
               kolvo75 = kolvo75 + 1
         Else
         Worksheets("0,75").Cells(rwIndex, 4) =
Worksheets("100").Cells(rwIndex, 1)
         Worksheets("0,75").Cells(rwIndex, 5) =
Worksheets("100").Cells(rwIndex, 2)
            End If
            End If
            End If
            End If
         rwIndex = rwIndex + 1
         rwIndex2 = rwIndex2 + 1
         Next rwIndex1
      Next rwIndex
   Next rwIndex2
   For rwIndex2 = 1 To kolvo75
      If Worksheets("0,75").Cells(rwIndex2, 1).Value =
"" Then
         Worksheets("0,75").Cells(rwIndex2,
1).EntireRow.Delete
         rwIndex2 = rwIndex2 - 1
      End If
   Next rwIndex2
num75 = 0
   For rwIndex2 = 1 To kolvo75
      num75 = num75 +
Worksheets("0,75").Cells(rwIndex2, 1)
   Next rwIndex2
mox75 = num75 / kolvo75
Worksheets("0,75").Cells(2 + kolvo75, 1) = "Коэф-ты
ур-я ax+b"
Worksheets("0,75").Cells(2 + kolvo75 + 1, 1) = "a"
Worksheets("0,75").Cells(2 + kolvo75 + 1, 2) = "b"
num751 = 0
   For rwIndex2 = 1 To kolvo75
      num751 = num751 +
Worksheets("0,75").Cells(rwIndex2, 2)
   Next rwIndex2
moy75 = num751 / kolvo75
koef75 = 0
koef751 = 0
   For rwIndex2 = 1 To kolvo75
      koef75 = koef75 +
((Worksheets("0,75").Cells(rwIndex2, 1) - mox75) *
(Worksheets("0,75").Cells(rwIndex2, 2) - moy75))
      koef751 = koef751 +
((Worksheets("0,75").Cells(rwIndex2, 1) - mox75) ^ 2)
   Next rwIndex2
koefa75 = koef75 / koef751
Worksheets("0,75").Cells(2 + kolvo75 + 2, 1) = koefa75
koefb75 = moy75 - koefa75 * mox75
Worksheets("0,75").Cells(2 + kolvo75 + 2, 2) = koefb75
Worksheets("0,75").Cells(2 + kolvo75 + 6, 1) = "Ур-е"
Worksheets("0,75").Cells(2 + kolvo75 + 6, 2) = "e"
   For rwIndex2 = 1 To kolvo75
      For rwIndex1 = 2 + kolvo75 + 7 To 2 + kolvo75 + 7
+ kolvo75 - 1
         eq75 = koefa75 *
Worksheets("0,75").Cells(rwIndex2, 1) + koefb75
         Worksheets("0,75").Cells(rwIndex1, 1) = eq75
         rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
   For rwIndex2 = 1 To kolvo75
      For rwIndex1 = 2 + kolvo75 + 7 To 2 + kolvo75 + 7
+ kolvo75 - 1
         e75 = Worksheets("0,75").Cells(rwIndex2, 2) -
(koefa75 * Worksheets("0,75").Cells(rwIndex2, 1) +
koefb75)
         Worksheets("0,75").Cells(rwIndex1, 2) = e75
         rwIndex2 = rwIndex2 + 1
      Next rwIndex1
   Next rwIndex2
Worksheets("0,75").Cells(2 + kolvo75 + 1, 4) = "R^2"
R175 = 0
R275 = 0
R75 = 0
   For rwIndex1 = 2 + kolvo75 + 7 To 2 + kolvo75 + 7 +
kolvo75 - 1
      R175 = R175 +
((Worksheets("0,75").Cells(rwIndex1, 1) - moy75) ^ 2)
```

```
   Next rwIndex1
   For rwIndex = 1 To 1 + kolvo75 - 1
      R275 = R275 +
((Worksheets("0,75").Cells(rwIndex, 2) - moy75) ^ 2)
   Next rwIndex
R75 = R175 / R275
max751 = Worksheets("0,75").Cells(2 + kolvo75 + 7, 2)
   For rwIndex1 = 2 + kolvo75 + 7 + 1 To 2 + kolvo75 +
7 + kolvo75 - 1
      If Worksheets("0,75").Cells(rwIndex1, 2) > max751
Then
      max751 = Worksheets("0,75").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
min751 = Worksheets("0,75").Cells(2 + kolvo75 + 7, 2)
   For rwIndex1 = 2 + kolvo75 + 7 + 1 To 2 + kolvo75 +
7 + kolvo75 - 1
      If Worksheets("0,75").Cells(rwIndex1, 2) < min751
Then
      min751 = Worksheets("0,75").Cells(rwIndex1, 2)
      End If
   Next rwIndex1
ypr75 = xpr * koefa75 + koefb75
di75 = (Abs(min751) + max751) / 2 * 100 / ypr75
delta75 = Abs(ypr75 - ypr) / ypr * 100
Worksheets("0,75").Cells(2 + kolvo75 + 2, 4) = R75
Worksheets("0,75").Cells(2 + kolvo75 + 1, 5) = "DI,%"
Worksheets("0,75").Cells(2 + kolvo75 + 2, 5) = di75
Worksheets("0,75").Cells(2 + kolvo75 + 1, 6) =
"Delta,%"
Worksheets("0,75").Cells(2 + kolvo75 + 2, 6) = delta75
Worksheets("0,75").Cells(2 + kolvo75 + 4, 2) = ypr75
Worksheets("Itog").Cells(6, 2) = R75
Worksheets("Itog").Cells(6, 3) = di75
Worksheets("Itog").Cells(6, 4) = delta75
Worksheets("Itog").Cells(6, 5) = kolvo75
Worksheets("Itog").Cells(6, 6) = R75 * kolvo75 / kol
kolvo70 = 0
   For rwIndex2 = 1 To kol
      For rwIndex = 3 To 3 + kol - 2
         For rwIndex1 = 4 + kol + 7 To 4 + kol + 7 + kol -
1
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 1) +
Cells(7, 12).Value) Then
            If Cells(rwIndex, 2) < (Cells(rwIndex1, 8) +
Cells(8, 12).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 1) -
Cells(7, 12).Value) Then
            If Cells(rwIndex, 2) > (Cells(rwIndex1, 8) -
Cells(8, 12).Value) Then
    …
End Sub
```

ПРИЛОЖЕНИЕ В

Экспериментальные данные

Таблица В.1 – Исходные данные эксперимента 3

X	Y
301	52,46
328	72,3
353	54,08
372	62,98
386	52,95
389	53,71
401	63,69
408	58,99
415	66,8
444	59,74
446	71,66
457	72,81
458	68,44
463	69,33
484	70,77
491	79,38
503	74,39
512	85,58
517	82,03
527	94,44
535	70,84
547	89,18
596	93,24
623	90,5

Таблица В.2 – Результаты применения метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество данных	Точность
100	0,71	16,19		24	
90	0,80	13,39	3,30	22	0,73
85	0,81	10,04	2,59	20	0,68
80	0,76	10,02	2,14	19	0,60
75	0,76	10,02	2,14	19	0,60
70	0,84	8,44	5,98	18	0,63
65	0,87	6,54	6,44	17	0,62
60	0,84	6,41	5,84	15	0,53
50	0,78	6,10	2,34	13	0,42

Таблица В.3 – Результаты применения первой модификации метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,71	16,19		24	
95	0,81	13,56	1,80	23	0,78
90	0,84	10,04	0,07	22	0,77
85	0,87	8,52	1,63	21	0,76
80	0,89	7,47	1,99	20	0,74
75	0,91	6,54	1,48	18	0,68
70	0,91	6,54	1,48	18	0,68
65	0,92	6,07	0,65	17	0,65
60	0,92	6,07	0,65	17	0,65
55	0,94	4,38	0,49	15	0,59
50	0,94	4,38	0,49	15	0,59

Таблица В.4 – Результаты применения второй модификации метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,71	16,19		24	
95	0,68	16,53	0,83	23	0,65
90	0,65	16,89	1,29	22	0,59
85	0,76	13,21	5,24	20	0,63
80	0,76	13,21	5,24	20	0,63
75	0,73	13,17	5,69	19	0,58
70	0,69	13,28	4,62	18	0,52
65	0,72	13,09	6,74	17	0,51
60	0,84	8,02	12,72	16	0,56
55	0,74	7,25	5,62	13	0,40
50	0,69	7,47	4,81	12	0,34

Таблица В.5 – Результаты применения метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,71	16,19	
90	0,75	14,36	0,48
85	0,76	13,93	1,02
80	0,78	13,80	1,72
75	0,79	13,78	2,34
70	0,79	13,39	3,05
65	0,80	12,98	3,75
60	0,81	12,57	4,46
50	0,82	11,97	5,55

Таблица В.6 – Результаты применения первой модификации метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,71	16,19	
95	0,74	14,30	0,34
90	0,77	13,21	0,31
85	0,79	12,29	0,31
80	0,81	10,95	0,45
75	0,82	10,02	0,54
70	0,84	9,20	0,60
65	0,86	8,22	0,65
60	0,87	7,54	0,64
55	0,89	6,76	0,58
50	0,90	6,21	0,51

Таблица В.7 – Результаты применения второй модификации метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,709	16,19	
95	0,710	16,24	0,11
90	0,714	16,83	1,24
85	0,716	17,11	2,61
80	0,730	16,68	4,28
75	0,735	16,40	5,51
70	0,736	16,14	6,67
65	0,736	15,91	8,42
60	0,741	15,23	10,17
55	0,734	14,41	12,55
50	0,725	13,82	14,41

Таблица В.8 – Исходные данные эксперимента 4

X	Y	X	Y	X	Y
500	1070	1020	1200	1160	1330
850	1060	1040	1230	1170	900
880	1130	1040	1250	1170	1370
900	1110	1040	1250	1200	1370
900	1100	1060	1270	1200	1370
920	1140	1060	1270	1210	1370
920	1150	1060	1270	1220	1390
920	1150	1070	1280	1230	1420
930	1125	1070	1290	1230	1400
935	1152	1070	1260	1230	1380
940	1300	1080	1250	1240	1440
940	1130	1080	1290	1240	1700
950	1150	1090	1280	1240	1410
960	1150	1090	1330	1250	1420
960	1160	1090	1260	1250	1440
970	1170	1090	1290	1260	1450
980	1180	1100	1320	1280	1440
980	1190	1110	1300	1280	1470
990	1230	1110	1300	1290	1450
1000	1230	1120	1310	1300	1460
1000	1180	1140	1320	1300	1480
1000	1210	1150	1350	1300	1490
1010	1230	1150	1370	1310	1490
1010	1210	1150	1330	1320	1510
1010	1190	1150	1310	1320	1520
1020	1250	1150	1320	1360	1550
				1520	1540

Таблица В.9 – Результаты применения метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,72	25,00		79	
90	0,95	5,66	1,77	75	0,90
85	0,95	5,64	1,59	74	0,89
80	0,97	2,39	1,77	72	0,89
75	0,97	2,38	1,79	70	0,86
70	0,97	2,38	1,61	69	0,85
65	0,96	2,39	1,52	63	0,76
60	0,96	2,39	1,74	61	0,74
50	0,91	2,41	2,37	53	0,61

Таблица В.10 – Результаты применения первой модификации метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,72	25,00		79	
95	0,95	8,26	1,44	76	0,91
90	0,96	5,29	1,60	75	0,92
85	0,96	5,29	1,60	75	0,92
80	0,96	5,29	1,60	75	0,92
75	0,96	5,29	1,60	75	0,92
70	0,96	5,29	1,60	75	0,92
65	0,98	2,49	2,38	74	0,92
60	0,98	2,49	2,38	74	0,92
55	0,98	2,45	2,16	72	0,89
50	0,98	2,45	2,16	72	0,89

Таблица В.11 – Результаты применения второй модификации метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,72	25,00		79	
95	0,75	24,18	2,13	77	0,73
90	0,76	19,67	1,11	74	0,72
85	0,74	19,62	0,74	72	0,68
80	0,66	19,51	0,13	65	0,54
75	0,56	19,40	0,41	58	0,41
70	0,51	19,27	0,93	55	0,35
65	0,37	18,88	2,60	49	0,23
60	0,81	5,51	0,60	45	0,46
55	0,74	5,48	1,16	40	0,37
50	0,73	5,34	1,53	39	0,36

Таблица В.12 – Результаты применения метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,717	25,00	
90	0,903	10,65	1,05
85	0,912	9,89	1,26
80	0,913	9,41	1,50
75	0,910	9,12	1,69
70	0,903	8,85	1,92
65	0,889	8,93	2,30
60	0,867	9,70	2,77
50	0,813	11,17	3,65

Таблица В.13 – Результаты применения первой модификации метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,717	25,00	
95	0,896	10,17	0,54
90	0,912	8,84	0,64
85	0,923	7,98	0,71
80	0,932	7,21	0,77
75	0,938	6,93	0,82
70	0,942	6,77	0,85
65	0,948	6,43	0,93
60	0,952	6,15	0,98
55	0,956	5,84	1,05
50	0,959	5,60	1,06

Таблица В.14 – Результаты применения второй модификации метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,72	25,00	
95	0,77	24,28	1,89
90	0,75	25,17	2,42
85	0,72	26,55	2,78
80	0,67	27,76	3,45
75	0,62	28,58	4,10
70	0,56	29,31	4,79
65	0,47	30,24	5,49
60	0,38	31,49	5,21
55	0,25	32,80	3,69
50	0,16	33,37	1,42

Таблица В.15 – Исходные данные эксперимента 5

X	Y	X	Y	X	Y
1008	217,7	1846	271,7	2410	364,3
1120	299	1848	455	2448	466
1129	276,6	1884	441,7	2473	679,9
1184	221,9	1942	364	2479	613,5
1331	304,3	1968	615,8	2493	768,7
1344	288,4	1970	457,9	2496	510,7
1392	355,4	2008	598,7	2707	548,4
1416	379,3	2042	405,9	2710	511,1
1423	374,8	2050	353,1	2826	726,6
1433	315	2063	612,4	2936	554
1439	365,2	2070	416,7	2978	463,3
1551	332,84	2070	651,3	2991	749,74
1605	314,1	2073	429	3032	606
1606	393,1	2080	363,5	3130	854,7
1615	517,7	2089	374,1	3136	766,4
1674	397	2100	356,2	3202	635,7
1753	305,1	2132	320	3244	786
1799	543	2191	711,8	3356	912,7
1817	471	2194	681,4	4065	811,8
1822	396,7	2230	350,7		

Таблица В.16 – Результаты применения метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,630	27,84		59	
90	0,626	27,34	0,91	57	0,61
85	0,630	24,46	0,53	54	0,58
80	0,631	22,36	0,91	48	0,51
75	0,570	21,19	5,60	42	0,41
70	0,581	18,61	5,31	39	0,38
65	0,569	18,83	7,77	35	0,34
60	0,587	18,80	8,73	33	0,33
50	0,583	15,59	15,24	23	0,23

Таблица В.17 – Результаты применения первой модификации метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,63	27,84		59	
95	0,65	27,50	0,59	58	0,64
90	0,73	22,23	0,13	53	0,65
85	0,75	20,61	3,49	48	0,61
80	0,82	17,31	4,05	44	0,61
75	0,83	17,27	4,17	43	0,61
70	0,86	14,86	4,50	40	0,59
65	0,89	13,21	2,82	36	0,54
60	0,90	11,99	3,25	34	0,52
55	0,91	10,69	0,17	31	0,48
50	0,90	9,24	0,02	29	0,44

Таблица В.18 – Результаты применения второй модификации метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество точек	Точность
100	0,63	27,84		59	
95	0,61	27,83	1,58	58	0,60
90	0,53	27,88	1,83	55	0,49
85	0,41	27,98	5,12	50	0,35
80	0,32	28,34	7,81	47	0,25
75	0,32	26,75	2,08	43	0,23
70	0,22	27,58	5,48	39	0,14
65	0,20	27,30	4,62	37	0,12
60	0,18	26,20	0,86	34	0,10
55	0,14	26,50	0,91	32	0,08
50	0,17	25,55	4,96	31	0,09

Таблица В.19 – Результаты применения метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,630	27,84	
90	0,634	27,63	2,16
85	0,637	25,27	2,49
80	0,639	24,77	3,05
75	0,641	24,89	3,81
70	0,643	24,96	4,89
65	0,644	25,05	6,13
60	0,645	25,16	7,54
50	0,646	25,33	10,44

Таблица В.20 – Результаты применения первой модификации метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,630	27,84	
95	0,631	27,65	0,01
90	0,645	23,82	0,02
85	0,67	20,96	0,11
80	0,693	18,44	0,40
75	0,715	16,78	0,62
70	0,737	15,39	0,82
65	0,765	13,73	1,00
60	0,787	12,58	1,02
55	0,81	11,26	1,01
50	0,83	10,35	0,93

Таблица В.21 – Результаты применения второй модификации метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,630	27,84	
95	0,634	27,82	2,10
90	0,622	27,79	2,91
85	0,607	28,04	4,42
80	0,585	29,36	6,51
75	0,571	29,22	8,63
70	0,560	28,67	10,99
65	0,546	27,84	14,84
60	0,533	27,54	17,70
55	0,52	27,40	22,05
50	0,51	27,27	25,50

ПРИЛОЖЕНИЕ Г

Копии документов о внедрении результатов исследований

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ
НАУЧНО-ПРОИЗВОДСТВЕННОЕ ОБЪЕДИНЕНИЕ
«ИНТЕРМЕТ»

Донецкая Народная Республика, 83117, город Донецк, Кировский район, улица Лишенко, дом 8,
тел. (062) 386 52 52, (071) 333 31 12, идентификационный код юридического лица 20341814
mail: info@vnet.dn.ua

Исх. № _____
от «23» июня 2021 г.

Диссертационный совет Д 01.024.04
при ГОУВПО "ДОНЕЦКИЙ
НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ"

СПРАВКА

о внедрении результатов исследований диссертационной работы Рычки Ольги Валентиновны на тему "Совершенствование методов выявления и корректировки аномальных измерений для повышения качества линейных регрессионных моделей", представленную на соискание ученой степени кандидата технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки)

Результаты диссертационных исследований Рычки О.В., а именно: рекомендации по выявлению аномальных данных при работе с большим объёмом данных, комплекс программ, реализующий предложенные методы обнаружения и корректировки аномальных данных были рассмотрены и приняты к использованию в ООО НПО «Интермет».

С уважением,
Генеральный директор Т.А. Белик

Соответствует оригиналу
Ученый секретарь Д 01.024.04
Т.В. Завадская

ДОНЕЦКАЯ НАРОДНАЯ РЕСПУБЛИКА
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
"ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ"
83001, г. Донецк, ул. Артема, 58 тел.: (062) 337-17-33, 335-75-62, факс: (062) 304-12-78
эл. почта: donntu.info@mail.ru

05.07.21 № 29-13/15
На №

Диссертационный совет Д 01.024.04
при ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

СПРАВКА

о внедрении результатов исследований диссертационной работы Рычки Ольги Валентиновны на тему "Совершенствование методов выявления и корректировки аномальных измерений для повышения качества линейных регрессионных моделей", представленную на соискание ученой степени кандидата технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки) в научно-исследовательскую деятельность ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Результаты диссертационных исследований Рычки О.В., а именно: алгоритм поиска и последующей обработки аномальных данных, направленный на повышение качества линейных регрессионных моделей, обоснования и рекомендации по применению методов корректировки данных, программное обеспечение, позволяющее исследовать исходные статистические данные, выполнять оперативный поиск аномальных данных были использованы при выполнении научно-исследовательских работ Н-1-16 "Анализ современных методов инженерии программного обеспечения для информационно-вычислительных и интеллектуальных систем", Н-16-18 "Исследование методов, технологий и средств инженерии программного обеспечения на различных классах приложений", Н-2020-14 "Усовершенствование средств инженерии программного обеспечения для актуальных классов IT-приложений" в 2016-2020 гг.

Проректор по научной работе
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,
д-р техн. наук, профессор С.В. Борщевский

Начальник научно-исследовательской части
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,
д-р техн. наук, профессор К.Н. Лабинский

Заведующий кафедрой
программной инженерии им. Л.П. Фельдмана
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,
д-р техн. наук, доцент С.А. Зори

Соответствует оригиналу
Ученый секретарь Д 01.024.04
Т.Б. Завадская

ДОНЕЦКАЯ НАРОДНАЯ РЕСПУБЛИКА
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
"ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ"
[illegible]

05.07.21 № 29-12/63
На №

Диссертационный совет Д 01.024.04
при ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

СПРАВКА

о внедрении результатов исследований диссертационной работы Рычки Ольги Валентиновны на тему "Совершенствование методов выявления и корректировки аномальных измерений для повышения качества линейных регрессионных моделей", представленную на соискание ученой степени кандидата технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ (технические науки) в учебный процесс ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Результаты диссертационных исследований Рычки О.В., а именно: усовершенствованные методы поиска и корректировки аномальных данных, направленные на повышение качества линейных регрессионных моделей, осуществление прогноза на базе построенной модели, комплекс программ, разработанный в соответствии с алгоритмами предложенных в работе методов, рекомендации по применению методов, внедрены в учебный процесс при чтении курсов лекций по дисциплинам «Эмпирические методы программной инженерии», «Численные методы в информатике» для студентов направления подготовки 09.03.04 «Программная инженерия», что отражено в учебных программах вышеуказанных дисциплин.

Проректор по научно-педагогической работе
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,
д-р техн. наук, профессор А.Б. Бирюков

Начальник учебного отдела
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,
канд. техн. наук, доцент Б.В. Гавриленко

Заведующий кафедрой
программной инженерии им. Л.П. [illegible]
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»,
д-р. техн. наук, доцент [illegible]

Соответствует оригиналу
Учен[illegible] 05.13.18, 01.024.04
[illegible] Завадская