2.4 ПРОГРАММА ДЛЯ РАСЧЁТА УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ ПРИ ОБРАБОТКЕ БОЛЬШИХ ДАННЫХ

Мельников Ф. В.

Технология анализа больших данных (Big Data) рассматривается как сквозная технология Национальной технологической инициативы (Постановление Правительства Российской Федерации от 18.04.2016 № 317 (ред. от 20.04.2019) «О реализации Национальной технологической инициативы»). Она должна трансформировать научно-технологический уклад многих отраслей человеческой деятельности, включая и сферу образования. Аналитика данных рассматривается как новый инструмент для реформирования образования на базе принципов персонализации, повышения качества образовательных результатов и управления системами образования на основании данных. Образовательная политика начинает строиться на образовательной аналитике.

По мнению Фиофановой О. А. [1, с. 8.] можно говорить о появлении «новой области междисциплинарного знания — «Педагогике, основанной на данных» (Data Driven Pedagogy), раскрывающей методологию и технологии анализа данных об образовании и детском развитии для использования в общеобразовательной практике и практике управления образованием». Одной из составляющей этого междисциплинарного знания является математическая и инженерно-технологическая составляющие. Big data — это большой массив разнородных данных. Чтобы они принесли пользу, в них нужно найти какие-то полезные закономерности: сходства, различия, общие категории и так далее. Процесс поиска таких закономерностей называют data mining — добыча данных, или глубинный анализ данных. К методам data mining относят и статистические методы, в том числе и регрессионный анализ, который позволяет найти важные факторы, влияющие на какой-либо заданный параметр. Регрессионный анализ используется при проведении педагогических, психологических, социологических исследований; решении эконометрических задач; обработке данных, полученных в ходе эксперимента. Регрессионный анализ используется при решении большого количества практических задач, которые предполагают определение связи между некоторыми факторами и результативными значениями и определение прогнозных значений результативного признака при влиянии определённых факторов. Анализ производится, как правило, с большим объёмом статистических данных, в связи с чем целесообразно производить вычисления с использованием ЭВМ.

Одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным является метод наименьших квадратов [2, 3, 4, 5]. Данный метод основан на минимизации суммы квадратов отклонений некоторых функций от исходных переменных.

Линия регрессии может быть найдена в виде линейной (линейная парная регрессия) или полиномиальной функции (полиномиальная парная регрессия или множественная линейная регрессия), наилучшим образом приближающей искомую кривую. Выбор типа регрессии зависит от поставленной задачи.

Общая постановка задачи заключается в следующем. Объект исследования представлен наблюдаемыми величинами x и y. Предполагается, что между зависимой переменной y и независимой переменной x существует объективная причинная связь. Также предполагается, что имеются значения наблюдаемых величин, т.е. по каждой из них имеется ряд данных. Необходимо оценить параметры уравнения регрессии и спрогнозировать возможные значения зависимой переменной с помощью независимых.

В соответствии с предпосылками классической многомерной линейной регрессионной модели количество наблюдений должно быть больше количества оцениваемых параметров. В случае полиномиальной парной регрессии количество наблюдений должно быть больше, чем максимальное значение степени полинома.

Проведение регрессионного анализа без использования ЭВМ не является целесообразным, т.к. анализ предполагает большой объём вычислений.

Регрессионный анализ может быть произведён с помощью:

- электронных таблиц (Microsoft Office Excel, LibreOffice Calc и т.д.)
- систем компьютерной алгебры (Maxima, Scilab и т.д.),
- а также различных специализированных программных пакетов, таких как Statistica, Stadia.

При использовании электронных таблиц необходимо подключать дополнительные пакеты (например, «Пакет анализа» в Excel), что может быть сопряжено с трудностями при настройке.

Системы компьютерной алгебры позволяют производить анализ без установки пакетов, однако в данном случае требуется понимание математического аппарата регрессионного анализа, изучение которого требует временных затрат. При этом процесс ввода матриц и формул является достаточно длительным, возможны ошибки, которые влияют на результат вычислений. Операции над матрицами, их свойства не относятся к сфере профессиональной компетенции педагогов и психологов.

Специализированные программные пакеты позволяют решать широкий спектр задач. В связи с этим они могут быть достаточно сложны, и, следовательно, могут требовать предварительного изучения.

Другим недостатком систем компьютерной алгебры и специализированных пакетов может являться их стоимость, что препятствует их использованию для непериодического решения задач только одного типа.

Таким образом, использование систем компьютерной алгебры и пакетов статистического анализа для обработки результатов педагогических, психологических и социологических исследований не во всех случаях является целесообразным.

Использование программного продукта, который позволяет решать только определённый вид задач, позволяет оптимизировать и облегчить процесс решения. Не требуется время на изучение функций программы, её настройку, т.к. она выполняет только одну основную функцию. Автором данного параграфа разработан программный продукт «Регрессионный анализ» для использования педагогами, психологами и другими специалистами, деятельность которых предполагает решение задач регрессионного анализа.

Программа предназначена для обработки результатов исследования с использованием математического аппарата метода наименьших квадратов и позволяет рассчитать уравнение парной линейной или полиномиальной регрессии.

Программный продукт может быть использован при решении задач в следующих сферах:

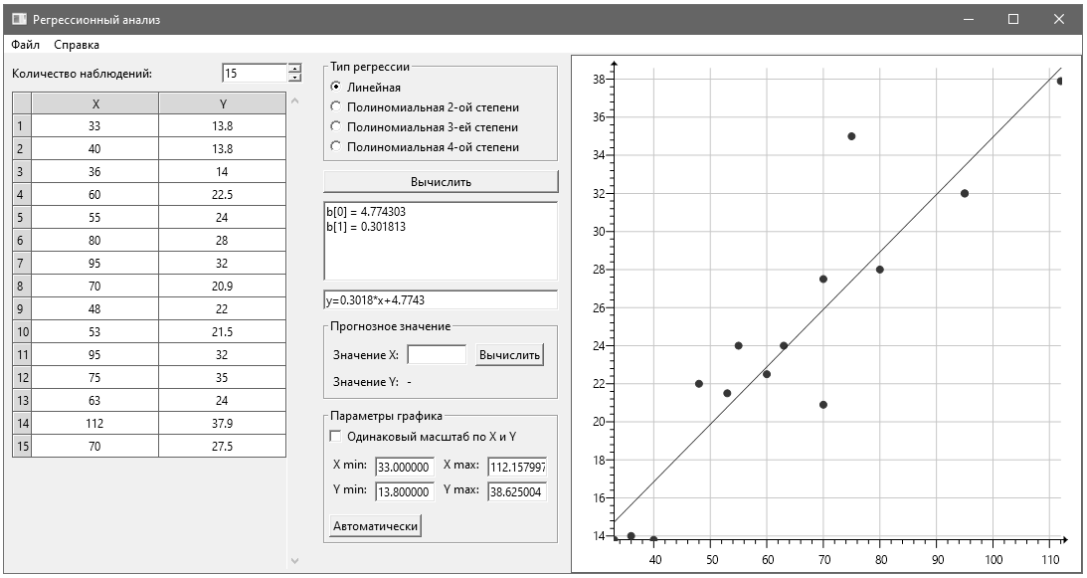
- педагогические исследования;
- психологические исследования;
- социологические исследования;
- эконометрика
- обработка экспериментальных данных (физика, биология)
- аппроксимация показаний измерительных приборов для их калибровки.

Программа реализует следующие функции:

- вычисление коэффициентов уравнения парной линейной или полиномиальной регрессии;
- вычисление прогнозного значения зависимой переменной с помощью независимых;
- ввод зависимых и независимых величин в виде таблицы;
- загрузка таблицы из файла;

- сохранение таблицы в файл;
- выбор типа уравнения регрессии;
- вывод коэффициентов уравнения;
- вывод уравнения регрессии;
- построение линии регрессии;
- настройка параметров осей графика.

Программа написана на языке C++. Для построения графического интерфейса пользователя используются библиотеки IUP, CD, которые разработаны Tecgraf/PUC-Rio в сотрудничестве с PETROBRAS/CENPES и имеют свободную лицензию MIT. Общий вид программы представлен на рис. 2.4.1.



	X	Y
1	33	13.8
2	40	13.8
3	36	14
4	60	22.5
5	55	24
6	80	28
7	95	32
8	70	20.9
9	48	22
10	53	21.5
11	95	32
12	75	35
13	63	24
14	112	37.9
15	70	27.5

Рисунок 2.4.1 – общий вид программы

Пользователь программного продукта «Регрессионный анализ» должен следовать ниже описанному алгоритму.

В поле «Количество наблюдений» вводится число, соответствующее количеству наблюдений, и вводятся данные в таблицу. Возможно сохранить данные наблюдений в текстовый файл или загрузить их из файла. Файл содержит данные наблюдений, значения X и Y разделены пробелом, каждое наблюдение записано в отдельной строке.

Затем необходимо выбрать тип регрессии. Количество наблюдений должно превышать значение степени полинома. Например, невозможно вычисление параметров полиномиальной регрессии, если количество наблюдений равно двум. Если установить количество наблюдений меньшее, чем значение максимальной степени полинома, программа выведет сообщение о невозможности вычисления обратной матрицы, или результаты вычислений будут неверными. Таким образом, минимальное количество наблюдений составляет:

- 2 при линейной регрессии,
- 3 при полиномиальной регрессии 2-ой степени,
- 4 при полиномиальной регрессии 3-ей степени,
- 5 при полиномиальной регрессии 4-ой степени.

После нажатия кнопки «Вычислить» производится вычисление, и затем в текстовых полях выводятся уравнение регрессии и его коэффициенты. Программа строит график зависимости $Y(X)$ и выводит линию регрессии в области построения.

Возможно настроить параметры графика – установить одинаковый масштаб по осям X и Y, задать границы осей или установить их автоматически (рис. 2.4.2).

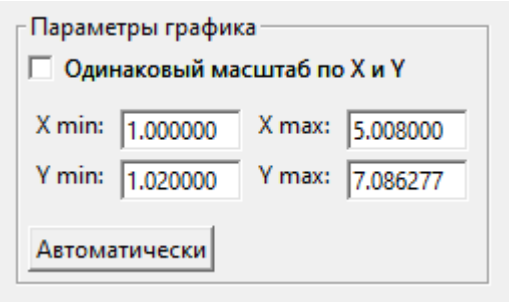


Рисунок 2.4.2. Настройка параметров области построения

Для масштабирования графика необходимо зажать клавишу Ctrl и использовать колёсико мыши, либо использовать контекстное меню (команды Zoom In, Zoom Out, Reset Zoom), которое открывается при нажатии правой кнопки мыши в области построения.

В контекстном меню доступны следующие функции:

- Show/Hide Legend – включение или выключение легенды графика (plot 0 – значения наблюдений, plot 1 – линия регрессии),
- Show/Hide Grid – включение или выключение отображения сетки,
- Copy – копирование изображения,
- Export – сохранение графика как изображения,
- Print – печать графика.

Для определения прогнозного значения необходимо ввести значение зависимой переменной X. Вычисленное значение независимой переменной Y будет отображено в метке «Значение Y» (рис .2.4.3).

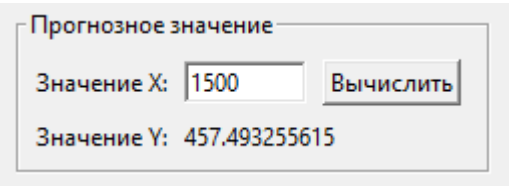


Рисунок 2.4.3. Расчёт прогнозного значения

Использование сложных, многофункциональных программных продуктов не во всех случаях является целесообразным для специалистов, для которых статистическая обработка данных не является основным видом деятельности.

Для решения задач регрессионного анализа педагогами, психологами и другими специалистами была разработана программа, которая не требует предварительного изучения интерфейса и детального знания математического аппарата методов регрессионного анализа.

Программа позволяет определить зависимость между зависимой и независимой переменными. Реализована возможность вычисления параметров линейной и полиномиальной парной регрессии, вычисления прогнозного значения зависимой переменной при заданном значении независимой переменной. Результаты отображаются в графическом виде.

Программа может применяться при проведении педагогических, психологических, социологических и других исследований.

Список источников

1. Фиофанова О. А. Анализ больших данных в сфере образования: методология и технологии: монография / О. А. Фиофанова. — М.: Издательский технологии: монография / О. А. Фиофанова. — М.: Издательский дом «Дело» РАНХиГС, 2020. — 200 с.
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 576 с.
3. Домбровский, В. В. Эконометрика / В. В. Домбровский. — Текст : электронный // Научная библиотека Томского государственного университета : [сайт]. — URL: http://sun.tsu.ru/mminfo/2016/Dombrovski/start.htm (дата обращения: 28.04.2021).
4. Селютин В.Д., Лебедева Е.В., Яремко Н.Н. Применение линейных регрессионных моделей в педагогических исследованиях // Ученые записки ОГУ. Серия: Гуманитарные и социальные науки. 2018. №3 (80). URL: https://cyberleninka.ru/article/n/primenenie-lineynyh-regressionnyh-modeley-v-pedagogicheskih-issledovaniyah (дата обращения: 19.11.2021).
5. Свидетельство RU 2021612179. Обучающая программа-тренажер по численным методам решения задач линейной алгебры: программа для ЭВМ / Гончарова С.В., Власова Е.З., Свистунова М.П., Войтин Е.В., Сухачева В.А., Стрижов Е.Д. (RU); правообладатели Гончарова С.В., Власова Е.З., Свистунова М.П., Войтин Е.В., Сухачева В.А., Стрижов Е.Д. № 2020666821; заявл. 11.12.2020; опубл. 12.02.2021