Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное

образовательное учреждение высшего образования

«Воронежский государственный лесотехнический

университет имени Г.Ф. Морозова»

Кафедра автоматизации производственных процессов

РАСЧЕТНО-ГРАФИЧЕСКАЯ РАБОТА

По дисциплине: Технологии обработки информации

Обработка экспериментальных данных

Вариант 20

Студент группы ИС2–191–ОБ \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Шпинев А.В.

(номер группы) (подпись) (инициалы и фамилия)

Руководитель к.т.н. доцент \_\_\_\_\_\_\_\_\_\_\_\_\_ Мещерякова А.А.

(ученая степень, ученое звание) (подпись) (инициалы и фамилия)

Воронеж 2021

СОДЕРЖАНИЕ

[СОДЕРЖАНИЕ 2](#_Toc70537360)

[ВВЕДЕНИЕ 3](#_Toc70537361)

[1 КОРРЕЛЯЦИЯ И РЕГРЕССИЯ 4](#_Toc70537362)

[2 ВЫЧИСЛЕНИЕ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ 10](#_Toc70537363)

[3 УРАВНЕНИЕ ЛИНИИ РЕГРЕССИИ 12](#_Toc70537364)

[4 ТАБЛИЦЫ РЕЗУЛЬТАТОВ 13](#_Toc70537365)

[5 СТРУКТУРА ПРОГРАММЫ И РЕЗУЛЬТАТЫ ЕЁ РАБОТЫ 15](#_Toc70537366)

[ЗАКЛЮЧЕНИЕ 19](#_Toc70537367)

[СПИСОК ЛИТЕРАТУРЫ 20](#_Toc70537368)

ВВЕДЕНИЕ

В данной работе будет произведён анализ экспериментальных данных, полученных в виде набора значений двух зависимых величин. Будет сделан вывод о связи между ними на основании вычисления коэффициента корреляции и построено уравнение линейной регрессии. Полученная зависимость будет использована для прогнозирования зависимой величины.

Коэффициент корреляции используется для обозначения силы линейных взаимоотношений между двумя переменными. Регрессионный анализ используют для оценки уравнения, которое в наибольшей степени соответствует совокупности наблюдений зависимых и независимых переменных. С помощью оцененного таким образом уравнения можно предсказать, каково будет значение зависимой переменной для данного значения независимой переменной.

1 КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Одним из важных методов анализа экспериментальных данных является корреляционный анализ. Он позволяет установить наличие и степень связи между случайными величинами [1].

Простая связь означает наличие двух случайных переменных. Множественная связь предполагает действие нескольких переменных.

Корреляционный анализ отвечает на следующие вопросы:

1. Существует ли связь между переменными?
2. Какой тип имеет эта связь?
3. Насколько сильна эта связь?
4. Какой прогноз можно сделать с учётом этой связи?

Примером простой связи является измерение роста и веса пациентов при врачебном осмотре, или зависимость годового объема продаж от средств, потраченных на рекламу. В качестве множественной связи можно представить зависимость ощущения температуры человеком в зависимости от температуры воздуха и влажности. Кроме того, важно установить и направление связи. Какая переменная оказывает влияние, а какая является зависимой? Например, при установлении связи между затратами на рекламу и прибылью, очевидно, независимой переменной являются затраты на рекламу.

Независимая переменная – это та, значение которой можно изменять. Зависимая переменная – это переменная, которую нельзя менять по желанию исследователя. Её значение является следствием определённого числа скрытых причин. Для выявления зависимости переменных можно строить графическое представление данных и визуально определять, имеет место зависимость и каково её направление.

Предположим, что в результате эксперимента измеряются две случайные величины  и . Их выборки представляют собой пары чисел (точки):

 (1.1),

где  – число испытаний. Вместе с анализом величин  и  по отдельности, нужно исследовать их возможную зависимость. Является ли эти величины независимыми? Если же между ними имеется зависимость, то какого рода?

Если между переменными имеется связь, то говорят, что  и  коррелированы. Для определения такого рода зависимости вычисляют величину, называемую коэффициентом корреляции . В случаях, когда других переменных нет, его обозначают просто . Эта величина помогает установить характер связи между исследуемыми переменными:

 (1.2)

Чем ближе значение  к нулю, тем слабее корреляция. Если же  близок к  или , тем корреляция сильнее, то есть зависимость между  и  близка к линейной. В случае, если  или , все точки выборки лежат на одной прямой.

Таким образом, коэффициент корреляции отражает степень именно линейной зависимости между исследуемыми величинами. При наличии зависимости другого вида (например, кубической) он может быть близок к нулю.

Приведём формулы для вычисления .

 (1.3),

 (1.4),

 (1.5),

 (1.6),

 (1.7),

(1.8).

Существует общепринятая шкала для интерпретации значений коэффициента корреляции.

**Таблица 1.1 – Шкала интерпретации **

|  |  |
| --- | --- |
| Значение | Уровень связи между переменными |
| 0,75 – 1,00 | Очень высокая положительная |
| 0,50 – 0,74 | Высокая положительная |
| 0,25 – 0,49 | Средняя положительная |
| 0,00 – 0,24 | Слабая положительная |
| 0,00 – -0,24 | Слабая отрицательная |
| -0,25 – -0,49 | Средняя отрицательная |
| -0,50 – -0,74 | Высокая отрицательная |
| -0,75 – -1,00 | Очень высокая отрицательная |

Если коэффициент корреляции близок к единице, то линейная зависимость существует, и этим можно воспользоваться для прогнозирования числа зрителей. Для этого применяется регрессионный анализ, тесно связанный с корреляционным.

Когда установлена линейная связь между переменными, исследователи должны рассмотреть возможные **виды связи** и выбрать ту, которая диктуется логикой данного исследования. Существует несколько видов связи.

Прямая причинно-следственная связь между исследуемыми переменными. В этом случае переменная  влияет на переменную . Например, наличие воды ускоряет рост растений, а яд вызывает смерть.

Обратная причинно-следственная связь. В этом случае переменная  влияет на значение . Можно предположить, что употребление большого количества чая вызывает нервозность. Но также может быть, что нервный человек пьет чай, чтобы успокоиться.

Связь между исследуемыми переменными может быть вызвана третьей переменной. Например, исследователем установлено, что существует определённая корреляция между числом посещений магазинов вечером в холодную погоду и уменьшением продаж прохладительных напитков. Очевидно, несмотря на зависимость, причиной обоих этих явлений является третье явление – холодная погода.

Взаимосвязь между несколькими переменными. Можно обнаружить связь между оценками студентов в академии и их оценками в школе. Но в этом случае могут действовать и другие переменные: уровень мотивации, жизненные обстоятельства, значимость предмета для студента.

Кроме всего вышеперечисленного, зависимость между событиями может быть случайна. Исследователь может найти значимую зависимость между уменьшением числа мышей весной и ростом солнечной активности. Но здравый смысл говорит о том, что связь между этими переменными случайна.

Таким образом, коэффициент корреляции показывает исследователю не причинно-следственную связь между событиями, а наличие линейной связи между ними и степень этой связи.

На рисунке 1 (с линейно расположенными точками) видно, что зависимость имеет приближенно линейный характер. Значения переменных расположены вокруг некоей прямой линии. Она называется линией регрессии. Для её построения несколько способов. Один из них – непосредственный. Если представить натянутую нить между двумя точками на рисунке, то можно выбрать визуально наиболее подходящее положение. Если после этого нарисовать эту нить, то при помощи измерений можно определить уравнение этой прямой. Эта грубая оценка пригодна в некоторых случаях. Также для этой цели существует несколько методов. Рассмотрим один из них, называемый методом наименьших квадратов.

Если установлена линейная связь между переменными  и , то можно отыскать функцию вида , выражающую зависимость  от .

Пусть даны пары чисел (иначе говоря, точек)

 (1.9).

Требуется найти такую прямую, чтобы сумма квадратов отклонений координат этих точек от прямой была как можно меньше.

Это означает, что выражение

 (1.10)

должно быть минимальным.

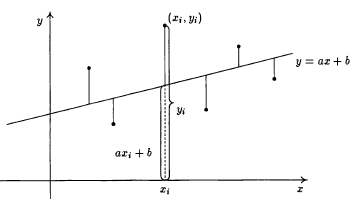


Рисунок 1.1 **–** Иллюстрация метода наименьших квадратов

Это выражение представляет собой функцию двух переменных  и , поскольку результаты наблюдений  и  заданы. Это выражение принимает минимальное значение, если величины  и  связаны соотношениями

 (1.11).

Эта система имеет единственное решение

,  (1.12).

Отыскав значения  и , мы сможем записать уравнение прямой, наилучшим образом выражающую статистическую связь между переменными  и . Эта прямая называется прямой регрессии  на .

После отыскания коэффициентов линии регрессии, можно оценить качество приближения результатов наблюдений. Подставив в выражение

 (1.13)

Найденные значения  и , вычислим среднюю квадратичную погрешность, иначе называемую ошибкой уравнения регрессии.

 (1.14).

Эта величина отражает среднюю длину вертикальных отклонений исследуемых точек от прямой регрессии. Чем меньше , тем ближе результаты наблюдений к прямой регрессии.

2 ВЫЧИСЛЕНИЕ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Рассмотрим задачу, решаемую туристическими фирмами, которые продают туристические услуги. Чтобы продать туристические услуги, фирмы покупают рекламу за x для привлечения туристов y. Например, можно предположить, что количество туристов, воспользовавшихся услугами каждой фирмы, будет зависеть от количества затрат на рекламу каждой фирмы. Пусть данные шести туристических фирм следующие:

**Таблица 2.1 – Исходные данные**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | 1.5 | 2.2 | 2.5 | 6.1 | 6.5 | 7.1 |
| y | 2.2 | 2.36 | 0.8 | 2.68 | 7.4 | 5.48 |

Требуется определить коэффициент корреляции между числом затрат на рекламу и числом туристов, воспользовавшихся услугами каждой фирмы, прямую регрессии, а также сделать прогноз числа туристов при затратах на рекламу x = 7.

В таблице приведены шесть реализаций пары *x* и *y* случайных величин. Рассчитаем коэффициент корреляции.

Согласно таблице 1.1, наблюдается очень высокая положительная корреляция.

Результаты вычислений занесём в таблицу 4.1.

Таким образом, в нашем случае коэффициент корреляции довольно близок к единице. Следовательно, линейная зависимость существует, и этим обстоятельством можно воспользоваться для прогнозирования числа туристов.

3 УРАВНЕНИЕ ЛИНИИ РЕГРЕССИИ

Для прогнозирования числа туристов, рассмотренного в предыдущем примере, нужно найти прямую регрессии  на . Для этого подставим найденные ранее значения , ,  и  в формулы для коэффициентов регрессии. Получим:

Запишем уравнение прямой регрессии:

0,25094.

Если затраты на рекламу составляют 7 усл. ден. ед., то можно предположить, что число туристов составит:

тыс. туристов.

Найдём ошибку уравнения регрессии.

Итак, среднеквадратичная ошибка составляет 1,439532 тыс. туристов. Таким образом тыс. туристов. Результаты вычислений занесём в таблицу 4.2.

4 ТАБЛИЦЫ РЕЗУЛЬТАТОВ

Таблица 4.1 – Результаты вычислений коэффициента корреляции

|  |  |
| --- | --- |
| Параметр | Значение |
|  | 25,9 |
|  | 20,92 |
|  | 143,21 |
|  | 103,0224 |
|  | 113,848 |
|  | 4,316667 |
|  | 3,486667 |
| , | 5,2347 |
| , | 5,0135 |
| , | 3,9239 |
|  | 0,7659  Очень высокая положительная |

Таблица 4.2 – Результаты вычислений коэффициентов линии регрессии

|  |  |
| --- | --- |
| Параметр | Значение |
|  | 0,74959 |
|  | 0,25094 |
| Уравнение линии регрессии | 0,25094 |
|  | 1,439532 |
| 7 усл. ден. ед. | тыс. туристов |

5 СТРУКТУРА ПРОГРАММЫ И РЕЗУЛЬТАТЫ ЕЁ РАБОТЫ

Программа предназначена для вычисления коэффициента корреляции заданной пары последовательностей экспериментальных данных и выполнена в пакете прикладных математических программ для инженерных (технических) и научных расчётов Scilab 5.5.1 [2]. В программе используются стандартные функции пакета (таблица 5.1).

Таблица 5.1 – Функции, использованные при создании программы

|  |  |
| --- | --- |
| Функция | Назначение |
| sum(x) | Вычисление суммы элементов массива x |
| length(x) | Длина объекта x |
| reglin(x, y) | Вычисление коэффициентов линейной регрессии для последовательностей x и у |
| plot2d(x, y, -1) | Построение двумерного графика функции у от аргумента x. Параметр -1 обозначает вид отображения «+». |
| plot(x, a\*x+b) | Построение двумерного графика функции a\*x+b от аргумента x. |

Структура программы приведена на рисунке 5.1

|  |
| --- |
| Вычисление коэффициента корреляции  Вычисление коэффициентов регрессии  Вычисление коэффициентов регрессии  Вывод результатов в командную строку  Построение графика заданных последовательностей и линии регрессии |
| Рисунок 5.1 – Структура программы |

Результаты работы программы приведены на рисунке 5.2.

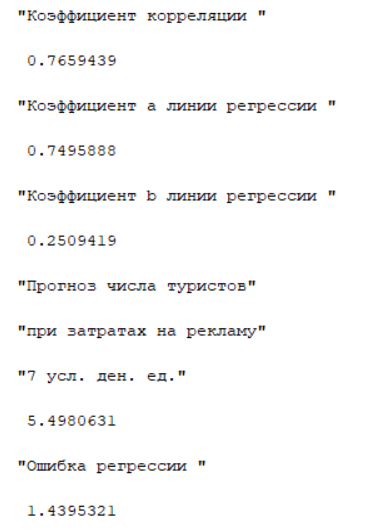


Рисунок 5.2 – Результаты работы программы

График заданных последовательностей и прямой регрессии приведён на рисунке 5.3.

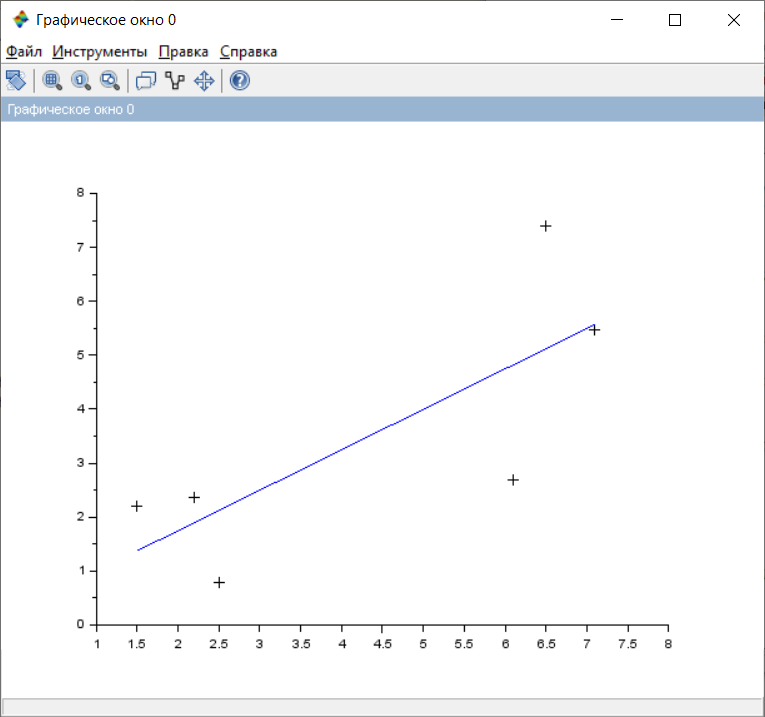


Рисунок 5.3 – Заданные последовательности экспериментальных данных и прямая регрессии

ЗАКЛЮЧЕНИЕ

В данной расчетно-графической работе был произведён анализ экспериментальных данных, полученных в виде набора значений двух зависимых величин. Была установлена связь этих величин на основании величины коэффициента корреляции. Построено уравнение линейной регрессии. Вычислено значение зависимой величины по заданному значению независимой, определена ошибка регрессии. Написана программа в пакете Scilab 5.5.1 для проведения указанных вычислений и построены графики заданных последовательностей и линии регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. Шикин, Е.В. Математические методы и модели в управлении: Учеб. пособие [Текст] / Е.В. Шикин, А.Г. Чхартишвили. – М.: Дело, 2004. - 440 с.
2. Тропин, И.С. Численные и технические расчеты в среде Scilab (ПО для решения задач численных и технических вычислений): Учеб. пособие [Текст] / И.С. Тропин, О.И. Михайлова, А.В. Михайлов. – Москва: 2008. – 65 с.
3. Оформление студенческих работ [Текст] : стандарт / Д. Н. Афоничев, Д. Ю. Капитонов, Н. Н. Харченко, А. С. Черных ; М-во образования и науки РФ, ГОУ ВПО «ВГЛТА». – Воронеж, 2011. – 59 с.