Изображение выглядит как текст, Шрифт, логотип, символ

Автоматически созданное описание**Национальный исследовательский университет «Высшая школа экономики»**

Высшая школа бизнеса

Бизнес информатика

Отчет о выполнении модульного домашнего задания №2

«Двумерные данные»

по майнору «Прикладной статистический анализ»

«Статистические методы анализа двумерных данных на примере данных дохода и уровня образования людей по Иркутской области»

Работу выполнила

студентка 2 курса

Зайдес Анна Александровна

Москва, 2023 г.

Содержание

[ВВЕДЕНИЕ 3](#_Toc152435192)

1. [ОПИСАНИЕ ПОКАЗАТЕЛЕЙ 3](#_Toc152435193)
2. [КОРРЕЛЯЦИОННЫЙ АНАЛИЗ 4](#_Toc152435194)
   1. [Корреляционное облако 4](#_Toc152435195)
   2. [Гипотеза 5](#_Toc152435196)
   3. [Вывод 6](#_Toc152435197)
3. [РЕГРЕССИОННЫЙ АНАЛИЗ 7](#_Toc152435198)
   1. [Построение линейной регрессионной модели 7](#_Toc152435199)
   2. [Построение нелинейных регрессионных моделей 8](#_Toc152435200)
4. [ВЫВОДЫ 10](#_Toc152435201)

[ПРИЛОЖЕНИЯ 11](#_Toc152435202)

[ЛИТЕРАТУРА 11](#_Toc152435203)

# ВВЕДЕНИЕ

**Актуальность:** учеба занимает значимую часть жизни людей, а также является одним из сложных периодов. Поэтому некоторые молодые люди не хотят тратить на нее время, не понимая, как она будет влиять на их будущее. Чтобы доказать школьникам и студентам, что все обучение не зря, я решила провести анализ зависимости уровня образования на доход, ведь деньги также являются одной из главных частей жизни. Была взята область, в которой я прожила большую часть своей жизни, чтобы показать своим знакомым как важно учится на примере наших же земляков.

**Цель:** проверить значимость уровня образования на доход

**Задачи:**

1. Охарактеризовать данные
2. Провести корреляционный анализ
3. Провести регрессионный анализ

# ОПИСАНИЕ ПОКАЗАТЕЛЕЙ

Был выбран регион с кодом по ОКАТО – 25, то есть Иркутская область. В качестве независимой переменной Х был взять уровень образования людей Иркутской области на момент 2022 года, которая имеет номинальную шкалу и разбивается на 9 значений:

1 – кадры высшей квалификации (послевузовское) - аспирантура

2 – высшее - специалитет, магистратура

3 – высшее - бакалавриат

4 – неполное высшее (незаконченное высшее) – оконченные 3 курса

5 – среднее профессиональное по программе подготовки специалистов

6 – среднее профессиональное по программе подготовки квалифицированных кадров

7 – среднее общее (среднее полное общее)

8 – основное общее (неполное среднее)

9 – не имеете основного общего

В качестве зависимой переменной Y был взят доход населения Иркутской области, которая имеет шкалу отношений. Значения выражаются в рублях и показывают доход человека за 2022 год.

# КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

# Корреляционное облако

Изначально дано 1843 пар данных, то есть 1843 человека. При построение графика для всех пар данных получаем график 1.

График 1.

По такому графику сложно определить характер связи признаков, так как данных много, имеются аномальные значения, например при уровне образования 5, то есть среднее профессиональное по программе подготовки специалистов, имеется один человек с доходом гораздо выше, чем в среднем получают люди с таким образованием. Такие аномалии будут всегда и их легко объяснить: на доход также влияют и сторонние факторы, например доход родителей, удача, образовательное учреждение и так далее. По данному графику сразу можно отметить, что людей без основного общего образования очень мало, поэтому будет неправильно сравнивать их с остальными, так как они дадут некорректные данные для статистики.

Исходя из вышенаписанного, я решила взять средний доход по каждому уровню образования, так как с таким подходом анализ получится более информативный. Поэтому находим средний доход для каждого значения уровня образования и опять строим корреляционное облако, график 2.

График 2.

По данному графику уже четко видно, что присутствует полиномиальная зависимость 6 степени. Полином почти перекрывает исходные данные, это значит, что тут четкая зависимость, единственные очень маленькие отклонения от полинома присутствуют на 4 уровне образования.

# Гипотеза

По графику 2 без сомнений можно выдвинуть конкурирующую гипотезу Н1. Теперь нужно доказать, что это действительно так, найдя фактический и теоретический критерии Стьюдента.

Для того чтобы найти фактический критерий Стьюдента, нужно рассчитать коэффициент корреляции. Так как мы уже поняли, что нужно рассматривать полиномиальную модель 6 степени, то найдем уравнение зависимости вида

С помощью дифференциальных уравнений можно составить систему линейных уравнений, составить по ней матрицу коэффициентов и найти решения с помощью метода Крамера. Все расчеты представлены в Приложении 1. После нахождения уравнения зависимости находим новый у и рассчитываем коэффициент корреляции. Уровень значимости для расчета теоретического критерия Стьюдента нам был дан – 0,05. В итоге tфакт и tтеор получились следующие:

Таблица 1.

|  |  |
| --- | --- |
|  | 0,05 |
| tфакт | -10,16858343 |
| tтеор | 2,364624252 |

=> Нулевая гипотеза Н0 отвергается с вероятностью ошибки . Изначальная гипотеза подтверждена.

Вернемся к началу: к выбранной модели корреляции. Наглядно покажем, почему был выбран именно полином 6 степени. Для этого найдем уравнение параболическое и линейное, а затем найдем для них коэффициенты корреляции и среднее стандартное отклонение:

Таблица 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| коэффициент корреляции линейной модели | коэффициент корреляции параболической модели | коэффициент корреляции полиномиальной модели 6 степени | Sлин | Sпарабол | Sполином |
| -0,96744 | -0,99932 | -0,96778 | 24540,588 | 24295,282 | 2549,252 |

Коэффициенты корреляции примерно все равны, но стандартное отклонение у полиномиального наименьшее, это значит, что оно ближе всего соответствует изначальным данным.

# Вывод

Коэффициент линейной корреляции оказался сильно значимым и отрицательным, это значит, что чем ниже цифра по шкале уровня образования (**то есть чем выше уровень образования**), тем выше доход. У параболической и полиномиальной модели также сильно значимый отрицательный коэффициент корреляции – из этого можно сделать уже более детальные выводы: максимум дохода достигается все-таки не при самом высоком уровне образования, но рядом с ним, а 6ая степень возникает из-за тех данных людей, у которых нет образования вообще. Из-за того, что их изначально было мало, можно сказать, что все-таки более правильно данные будет описывать параболическая модель, что показывает и самый высокий по модулю коэффициент корреляции.

Мои гипотезы, почему все-таки наибольший доход достигается не при высшем уровне образования: людей заканчивающих в наше время аспирантуру в Иркутской области довольно мало, так как появилась возможность уезжать в большие города, соответственно в представленных данных скорее всего люди, окончившие аспирантуру это пенсионеры или люди, близкие к пенсионному возрасту, так как в их студенческие годы не было тенденции уезжать из родных городов. Их доход, естественно, ниже, так как большинство уже получает только пенсию (но можно заметить, что люди науки получают большую пенсию). Но стоит заметить, что на графике 1 видно, что минимальный доход людей с наивысшем образованием выше минимальных доходов людей из других групп. Это говорит о том, что образование ценится и с ним всегда будет хороший доход.

# РЕГРЕССИОННЫЙ АНАЛИЗ

# Построение линейной регрессионной модели

Вернемся к исходным данным. Высчитаем по формулам уравнение линейной модели:

Коэффициент b1 получился отрицательным => связь между доходом и образованием обратная. (Тут не стоит забывать, что чем меньше цифра уровня образования, тем выше образование)

Сделав регрессионный анализ в MS EXCEL, получаем коэффициент детерминации равный 0,094759666, что говорит о не очень хорошей работе модели, то есть о слабой связи. По данным таблицам в Приложении 1 Регрессионный анализ проверим значимость по критерию Фишера:

Таблица 3.

|  |  |
| --- | --- |
| Qрегр < Qост | F-стат > F-теор |
| F-статистика | 192,7140658 |
| F-теор | 3,84651488 |

Qрегр < Qост говорит о слабой значимости, то есть почти подтверждает нулевую гипотезу, так как они влияют на R-квадрат. Но F-стат > F-теор говорит о значимости уравнения и опровергает нулевую гипотезу.

Теперь проверим значимость уравнения по значимости F. P-value вышло 9,54962E-42, что гораздо меньше 0,05, а значит уравнение значимо с вероятность ошибки 0,05. Проверим значимость коэффициентов:

Таблица 4.

|  |  |  |
| --- | --- | --- |
| b0 | 1,0033E-280 | <0,05 |
| b1 | 9,54962E-42 | <0,05 |

Оба коэффициента значимы => интерпретация возможна. Приведена выше. Выведем исходные данные и линейную регрессионную модель на график:

График 3.

Можем видеть, что с ухудшением уровня знаний на 1 падает доход.

# Построение нелинейных регрессионных моделей

Теперь рассмотрим другие две регрессионные модели: гиперболическую и степенную:

Все расчеты уравнений приведены в Приложении 1. Добавим их на общий график:

График 4.

По данной диаграмме довольно сложно понять, какая модель лучше всего описывает исходные данные, поэтому давайте рассчитаем стандартное отклонение:

(суммы отклонений приведены для каждой модели в Приложении 1)

Таблица 5.

|  |  |
| --- | --- |
| Модель | S |
| Линейная | 242960,3934 |
| Гиперболическая | 246278,8125 |
| Степенная | 253452,9909 |

Можем заметить, что линейная модель меньше всего отклоняется от исходных данных, а значит лучше всего подходит, то есть более оптимальна.

# ВЫВОДЫ

В данной работе была проделан большая работа над вопросом связи между уровнем образования и доходом населения Иркутской области. В первую очередь были выбраны данные с официального сайта Росстата. Затем провелся корреляционный анализ, для более понятного представления были взяты средние значения дохода для каждого уровня образования, то есть MS(Y|X=x). Корреляционный анализ дал нам следующие выводы: зависимость между уровнем образования и средним доходом представляет собой полином 6 степени, то есть наивысший средний доход достигается в точке Х = 2, то есть у людей, закончивших специалитет или магистратуру. Также при первоначальных данных мы выяснили, что людей без образования очень мало, а значит некорректно рассматривать их в статистике. Таким образом, корреляционная связь между образованием и доходом очень значима и представляет собой параболу с вершиной в точке 2 и ветвями вниз. Значимость корреляции была также доказана с помощью критерия Стьюдента. То есть учится важно и нужно, если вы хотите жить в достатке и ни в чем себе не отказывать. Возникает тогда вопрос, почему же люди с наилучшем образованием получают меньший доход? На этот вопрос легко ответить: это либо люди пенсионного возраста, так как в Иркутской области мало кто остается в наше время в аспирантуре и после нее, либо люди, которые пошли в науку, где доход и правда не сильно высокий.

Далее был проведен регрессионный анализ с изначальными данными, то есть с 1843 парами данных. Были рассчитаны 3 модели: линейная, степенная и гиперболическая и по данным стандартным отклонениям стало понятно, что наилучшим образом связь описывает линейная модель. Значимость уравнения была проверена коэффициентом корреляции, критериям Фишера и F-значимостью. Все способы указали, что значимость есть, а коэффициент корреляции показал, что она обратная, то есть с увеличением Х уменьшается Y, но так как уровень образования был он высшего с низшему, то связь между привычным нам уровнем образования и доходом прямая.

# ПРИЛОЖЕНИЯ

Приложение 1:тфайл EXCEL “Модульная\_работа\_2\_приложений”

# ЛИТЕРАТУРА

Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С.

Мхитаряна. — Москва : Издательство Юрайт, 2020 — 490 с. — (Высшее образование).

ISBN 978-5-534-00616-2. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: http://biblio-online.ru/bcode/450166