Министерство науки и высшего образования Российской Федерации
Калужский филиал
федерального государственного бюджетного образовательного
учреждения высшего образования
**«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»**
(КФ МГТУ им. Н.Э. Баумана)

И.И. Ерохин

ЛИНЕЙНАЯ РЕГРЕССИЯ.

Методические указания к выполнению лабораторной работы

по курсу «Технологии анализа данных»

Калуга – 2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Настоящие методические указания составлены в соответствии с программой проведения лабораторных работ по курсу «Технологии анализа данных» на кафедре «Программное обеспечение ЭВМ, информационные технологии» факультета «Информатика и управление» Калужского филиала МГТУ им. Н.Э. Баумана.

Методические указания, ориентированные на студентов 4-го курса направления подготовки 09.03.04 «Программная инженерия», содержат краткую теоретическую часть, описывающую область применения линейной регрессии.

Методические указания составлены в расчете на всестороннее ознакомление студентов с линейной регрессией.

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ, ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ЕЕ ВЫПОЛНЕНИЯ

Целью выполнения лабораторной работы является формирование практических навыков работы с линейной регрессией.

Основными задачами выполнения лабораторной работы являются:

1. Ознакомиться с линейной регрессией.

Результатами работы являются:

1. Построенный график линейной регрессии.
2. Подготовленный отчет.

КРАТКАЯ ХАРАКТЕРИСТИКА ОБЪЕКТА ИЗУЧЕНИЯ, ИССЛЕДОВАНИЯ

Линейная регрессия

Линейная регрессия — используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной *y* от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) *x* с линейной функцией зависимости.

Модель называется линейной регрессией, если функция регрессии $f(x,b)$ имеет вид

$$f(x,b)=b_0+b_1x_1+b_2x_2+\ldots++b_kx_k$$

где b_j — параметры (коэффициенты) регрессии, x_j — регрессоры (факторы модели), k — количество факторов модели.

Когда фактор единственный, говорят о простейшей линейной регрессии

$$f(x,b)=b_0+b_1x_1$$

Линейная регрессия - это способ построить модель системы из реального мира в мир статистики. В нашем примере в системе мы хотим отразить взаимосвязь между заработной платой и опытом. Формула $Salary=b_0+b_1x_{\text{exp}}$ это и есть модель. Где b_0 это стартовая точка при отсутсвии опыта — 30 тысяч условных единиц. b_1x_{exp} коэффициент на опыт, который помогает ответить на вопрос на сколько увеличится заработная плата при увеличении опыта на 1 год.

Плюс регрессии — интерпретируемость. Вы можете легко инетрпретировать данные, что поможет вам ответить на вопросы какие факторы и как влияют на заработную платую. В нашем случае фактор только один.

В задаче линейной регрессии мы подбираем такую модель, чтобы сумма квадратов разницы была минимальной.

Для построения модели будем использовать метод наименьших квадратов.

Метод наименьших квадратов (МНК) — математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных. МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

ОБРАЗЕЦ ВЫПОЛНЕНИЯ ЗАДАНИЯ

```
# Загружаем библиотеки
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import datasets

# загружаем данные
df = pd.read_csv('Salary_Data.csv')

#y = заработная плата, X = опыт
print(df.columns)
X = df['YearsExperience']
y = df['Salary']
print(X)
print(y)
X = X.values.reshape(len(X),1)
y = y.values.reshape(len(y),1)

# Разделим данные на тренировочную и тестовую выборку
X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=1/3, random_state=0)

#Создаем регрессора как метод, не указываем
дополнительные параметры
regressor = LinearRegression()#МНК
#Тренеруем модель на тренировочных данных.
regressor.fit(X_train,y_train)

#Прогнозируем результаты тестовой выборки
#Разница между y_pred и y_test в том, что y_pred это
прогнозные значения
#Теперь мы можем сравнить их с тестовыми значениями
y_pred = regressor.predict(X_test)

#Визуализация результатов тренировочного набор данных
plt.scatter(X_train,y_train,color = 'red')
plt.plot(X_train, regressor.predict(X_train), color =
'blue')
plt.title('Заработная плата vs Опыт(Тренировочные
данные)')
```

```
plt.xlabel('Опыт в годах')
plt.ylabel("Заработная плата")
plt.show()

#Визуализация результатов тестового набор данных
plt.scatter(X_test,y_test,color = 'red')
#линию регресии не меняем. Мы получим тестовые и оценим
как линия регрессии
# описывает тестовый набор
plt.plot(X_train, regressor.predict(X_train), color =
'blue')
plt.title('Заработная плата vs Опыт(Тестовые данные)')
plt.xlabel('Опыт в годах')
plt.ylabel("Заработная плата")
plt.show()
```

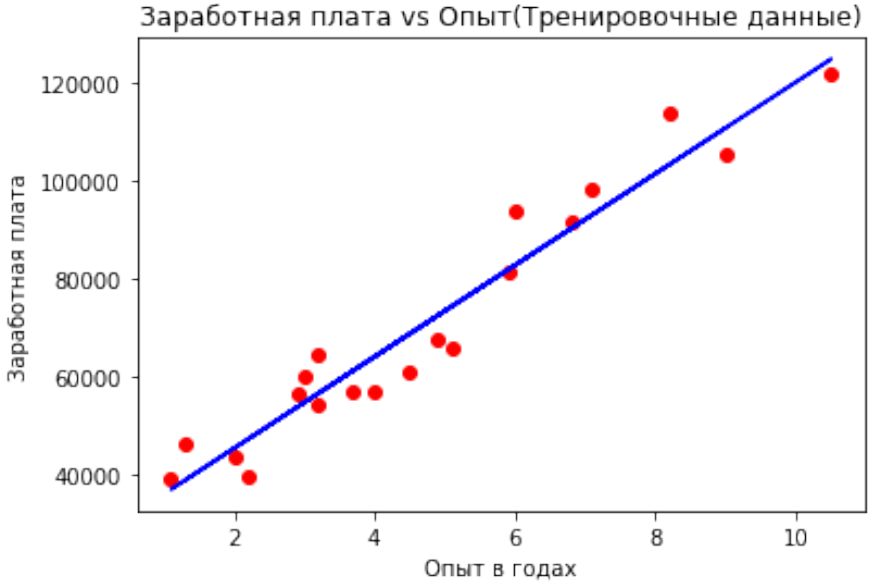

Рис. 1. График линейной регрессии тренировочных данных

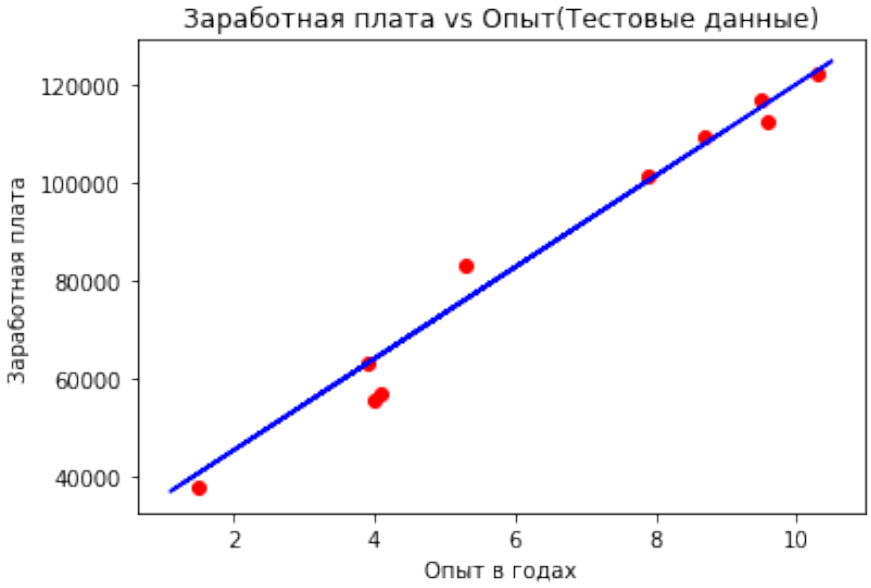

Рис. 2. График линейной регрессии тестовых данных

ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ

Создать дерево принятия решений согласно варианту, полученному у преподавателя.

ТРЕБОВАНИЯ К РЕАЛИЗАЦИИ

В качестве результата работы необходимо построить дерево принятия решений. По завершении готовится отчёт.

ВАРИАНТЫ ЗАДАНИЙ

Вариант 1

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать цену, в качестве x использовать размер помещения.

Вариант 2

Загрузить данные boston из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве y использовать уровень преступности, в качестве x использовать расстояния до центров занятости.

Вариант 3

Загрузить данные diabetes из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать возраст, в качестве x использовать массу тела.

Вариант 4

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать цену, в качестве x использовать количество ванных комнат.

Вариант 5

Загрузить данные boston из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве y использовать уровень преступности, в качестве x использовать среднюю стоимость домов.

Вариант 6

Загрузить данные diabetes из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать возраст, в качестве x использовать среднее артериальное давление.

Вариант 7

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать цену, в качестве x использовать количество спален.

Вариант 8

Загрузить данные boston из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве y использовать уровень преступности, в качестве x использовать среднее количество комнат в жилом помещении.

Вариант 9

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать размер помещения, в качестве x использовать количество этажей.

Вариант 10

Загрузить данные boston из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве y использовать среднее количество комнат в жилом помещении, в качестве x использовать ставку на налог.

Вариант 11

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать размер помещения, в качестве x использовать количество спален.

Вариант 12

Загрузить данные boston из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве y использовать среднюю стоимость домов, в качестве x ставку на налог.

Вариант 13

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать размер помещения, в качестве x использовать количество ванных комнат.

Вариант 14

Загрузить данные boston из библиотеки sklearn.datasets в структуру DataFrame. Построить график линейной регрессии. В качестве y использовать среднее количество комнат в жилом помещении, в качестве x использовать среднюю стоимость домов.

Вариант 15

Считать данные из файла Housing.csv в структуру DataFrame. Построить график линейной регрессии. В качестве yиспользовать цену, в качестве x использовать количество этажей.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Дайте определение линейной регрессии.
2. Дайте определение простой линейной регрессии.
3. Главный плюс регрессии.
4. Как подбирается модель.
5. Объясните принцип МНК.

ФОРМА ОТЧЕТА ПО ЛАБОРАТОРНОЙ РАБОТЕ

На выполнение лабораторной работы отводится 1 занятие (2 академических часа: 1 час на выполнение и сдачу лабораторной работы и 1 час на подготовку отчета).

Отчет на защиту предоставляется в печатном виде.

Структура отчета (на отдельном листе(-ах)): титульный лист, формулировка задания, описание процесса выполнения лабораторной работы, результаты выполнения работы, выводы.

ОСНОВНАЯ ЛИТЕРАТУРА

1. Маккинли, Уэс Python и анализ данных / Пер. с англ. Слинкин А.А. - М.: ДМК Пресс, 2015. - 482 с.:ил.
2. Грас, Дж. Data Science. Наука о данных с нуля / Пер. с англ. - СПБ.: БХВ -Петербург, 2017. - 336с.: ил.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

3. Henley, A.J. Learn Data Analysis with Python: Lessons in Coding / A.J. Henley, Dave Wolf ISBN 978-1-4842-3486-0

Электронные ресурсы:

4. Научная электронная библиотека http://eLIBRARY.RU
5. Электронно-библиотечная система http://e.lanbook.com