Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.03.03 Прикладная информатика.**

ОТЧЕТ

по лабораторной работе № 1

Название: Прогнозирование моделью линейной регрессии

Дисциплина: Прикладной анализ данных

Студент ИУ6-55Б _____ Л.Э. Барсегян
(Группа) (Подпись, дата) (И.О. Фамилия)

Преподаватель _____ М.А. Кулаев
(Подпись, дата) (И.О. Фамилия)

Москва, 2023

Вариант 8

Цель работы: изучить методы построения и оценки моделей линейной регрессии.

Задание:

1. Нормирование (масштабирование) исходных данных.
2. Расчет весов линейной регрессии по аналитической формуле.
3. Построение и интерпретация корреляционной матрицы. Определение степени мультиколлинеарности на основе числа обусловленности.
4. Анализ регрессионных остатков.
5. Определение весов линейной регрессии градиентным методом. Проанализировать изменение ошибки от итерации к итерации.
6. Сравнение результатов по аналитическому и градиентному методу.
7. С помощью библиотеки sklearn сделать fit-predict модели линейной регрессии. Сравнить результаты с ранее полученными.
8. С помощью библиотеки statmodels получить «эконометрический» результат обучения модели линейной регрессии. Проинтерпретировать все его составляющие (в т.ч. те, которые изучались только теоретически), сравнить с предыдущими результатами.
9. Сравнить качество получаемых моделей на основе коэффициента детерминации и MSE.
10. Сделать итоговый вывод касательно причин различия в результатах при выполнении работ разными методами, а также по получаемым моделям в целом. Провести сравнительный анализ.

Ход выполнения работы

1. Нормирование (масштабирование) исходных данных.

Для варианта 8 требовалось выделить данные по следующим районам: Северный, Центральный, Волго-Вятский, Северо-Кавказский, Восточно-Сибирский, Дальневосточный районы.

С помощью библиотеки Pandas данные были преобразованы в фреймы данных. Полученный фрейм был разделён на две части: целевые признаки X и целевые значения Y.

Так как разброс данных целевых признаков большой, использовалась Z-нормализация, также известная как "Стандартизация". Этот метод подходит тем, что централизует данные и менее чувствителен к выбросам, в отличие от метода "Min-Max Scaler".

2. Расчет весов линейной регрессии по аналитической формуле.

С помощью аналитической формы расчета весов $(X^T X)^{-1} X^T Y$ были получены веса для модели линейной регрессии. При расчете весов в качестве X был взят z-нормированный набор, к которому был добавлен единичный столбец x0. Полученные веса представлены на рисунке 1.

```
Полученные веса:
[[58.23617021]
 [-2.10483027]
 [-2.42128122]
 [ 0.95750726]
 [-2.14875258]
 [-0.32268881]
 [-0.30566366]
 [-1.15868994]
 [-0.84045299]
 [-0.79407424]]
```

Рисунок 1 — Полученные аналитически веса регрессии

3. Построение и интерпретация корреляционной матрицы. Определение степени мультиколлинеарности на основе числа обусловленности.

С помощью метода .corr() библиотеки Pandas была построена корреляционная матрица для нормированного набора X.

```
Корреляционная матрица:
    0         1         2         3         4         5         6         7  \
0 NaN       NaN       NaN       NaN       NaN       NaN       NaN       NaN
1 NaN  1.000000 -0.493003 -0.116073 -0.577910  0.250840 -0.376163 -0.392244
2 NaN -0.493003  1.000000  0.106909  0.069805 -0.066190  0.238550  0.044261
3 NaN -0.116073  0.106909  1.000000  0.392078 -0.086839  0.336084  0.164214
4 NaN -0.577910  0.069805  0.392078  1.000000  0.124522  0.195631  0.293789
5 NaN  0.250840 -0.066190 -0.086839  0.124522  1.000000 -0.240909 -0.145890
6 NaN -0.376163  0.238550  0.336084  0.195631 -0.240909  1.000000  0.555050
7 NaN -0.392244  0.044261  0.164214  0.293789 -0.145890  0.555050  1.000000
8 NaN  0.646093 -0.211056 -0.258291 -0.548800  0.165603 -0.503291 -0.542994
9 NaN -0.164474  0.046720 -0.312169  0.226796  0.191490 -0.111555  0.260572

          8         9
0       NaN       NaN
1  0.646093 -0.164474
2 -0.211056  0.046720
3 -0.258291 -0.312169
4 -0.548800  0.226796
5  0.165603  0.191490
6 -0.503291 -0.111555
7 -0.542994  0.260572
8  1.000000  0.026612
9  0.026612  1.000000

Число обусловленности: 4.805461631488785
```

Рисунок 2 — корреляционная матрица и

число обусловленности

Число обусловленности отражает, насколько чувствительна функция к изменениям или ошибкам на входе. В нашем случае число меньше 10, что говорит об отсутствии мультиколлинеарности в данных, а значит, полученные на их основе модели линейной регрессии с меньшей вероятностью будут выдавать неадекватные значения.

Для более наглядного представления с помощью библиотеки Seaborn была построена тепловая карта, представленная на рисунке 3.

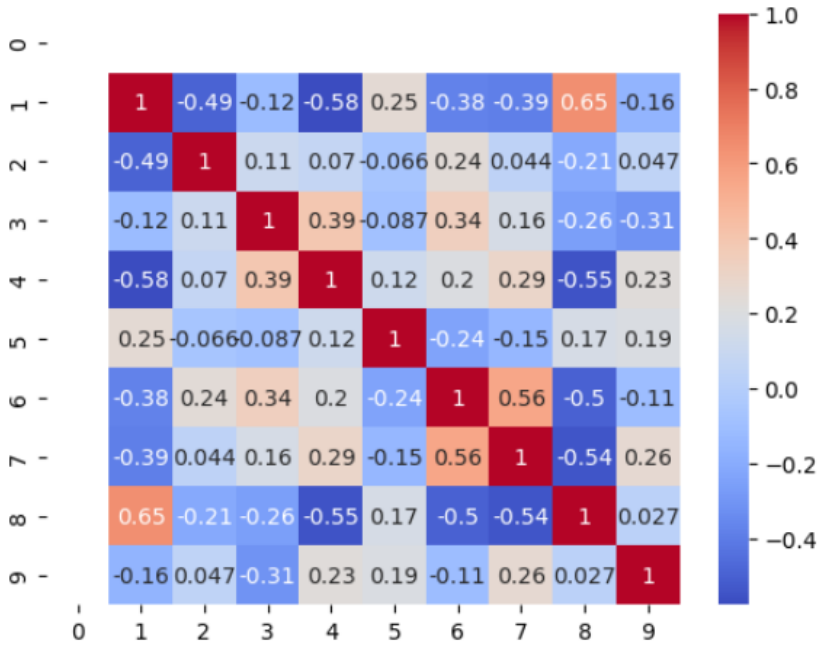

Рисунок 3 — Тепловая карта на основе корреляционной матрицы X

Полученные матрицы и тепловая карта показали, что параметры слабо коррелируют между собой. Также было замечено, что следующие признаки имеют средние коэффициенты корреляции, но наибольшие среди полученных:

- x_1 и x_4 — -0.58;
- x_4 и x_8 — -0.55;
- x_6 и x_8 — -0.5;
- x_7 и x_8 — -0.54;
- x_1 и x_8 — 0.65.
- x_6 и x_7 — 0.56.

Было учтено, что:

- x1 – рождаемость населения на 1000 человек;

- x4 – число разводов на 1000 человек;
- x6 – соотношение денежного дохода и прожиточного минимума, %;
- x7 – соотношение средней оплаты труда и прожиточного минимума трудоспособного населения, %;
- x8 – численность населения с денежными доходами ниже прожиточного минимума в % от численности населения (x8);

Таким образом, чем выше рождаемость, тем ниже число разводов и выше численность населения с денежными доходами ниже прожиточного минимума. Также высокие проценты соотношения денежного дохода и прожиточного минимума и соотношения средней оплаты труда и прожиточного минимума приводят к уменьшению процента численности населения с доходами ниже прожиточного минимума.

4. Анализ регрессионных остатков.

С помощью ранее вычисленных предсказанных значений модели были получены регрессионные остатки. На основе полученных остатков были посчитаны: MSE (Mean Square Error), RMSE (Root MSE) и R2.

Средняя квадратическая ошибка: 3.2850653276126236

Среднеквадратическое отклонение: 1.8124749177885535

Коэффициент детерминации: 0.7001679445457348

RMSE показывает среднее расстояние между наблюдаемыми значениями признака и прогнозируемыми значениями. Малое значение RMSE относительно среднего значения целевого признака (58.24) говорит об удовлетворительной точности модели.

5. Определение весов линейной регрессии градиентным методом. Проанализировать изменение ошибки от итерации к итерации.

На листинге 1 представлена функция, реализующая алгоритм регрессионного спуска.

Листинг 1. Определение весов линейной регрессии градиентным методом.

```
# 1. Инициализация весов
weights_i = np.ones((X.shape[1], 1))
learning_rate = 0.1

for i in range(200):
    # 2. Расчет таргета по весам
    Y_predicted_i =  np.matmul(X, weights_i)
    delta = Y - Y_predicted_i

    # 3. Расчет ошибки
    S_i = 0;
    for j in range(Y.shape[0]):
        S_i +=  (delta[j] ** 2 / Y.shape[0])[0]
    print(f"Итерация {i + 1}: \tОшибка: {S_i}")

    # 4. Расчет градиента функции потерь
    dS_dw = (- 2 / Y.shape[0]) * np.matmul(np.transpose(delta), X)
    # 5. Установка новых значений весов
    weights_i -= learning_rate * np.transpose(dS_dw)

    if i == 199:
        print(f"\t\tRMSE: {S_i**0.5}")
        R2_grad = r2_score(Y, Y_predicted_i)
        print(f"\t\tКоэффициент детерминации: {R2_grad}")
```

На каждой из итераций фиксировалась среднеквадратическая ошибка. Основной вклад в формирование весов и уменьшение ошибки внесли первые 30 итераций, далее ошибка уменьшалась незначительно, но при увеличении количества итераций постепенно приближалась к числу, полученному в п. 4.

6. Сравнение результатов по аналитическому и градиентному методу.

Результаты аналитического метода:

- Средняя квадратическая ошибка: 3.2850653276126236
- Среднеквадратическое отклонение: 1.8124749177885535
- Коэффициент детерминации: 0.7001679445457348

Результаты градиентного метода:

- Ошибка: 3.285107199403924
- Среднеквадратическое отклонение: 1.8124864687505735
- Коэффициент детерминации: 0.7001641228545357

Аналитический метод оказался немного точнее, но стоит учитывать, что точность градиентного метода была получена за 200 итераций. При 30 итерациях ошибка была немного больше.

7. С помощью библиотеки sklearn сделать fit-predict модели линейной регрессии. Сравнить результаты с ранее полученными.

С помощью библиотеки sklearn была обучена модель данных. На вход модели был подан исходный набор X, поскольку библиотека самостоятельно применяет методы нормализации к данным. Полученные значения RMSE, MSE, R2 практически совпадают с полученными ранее.

```
MSE: 3.2850653276126227
RMSE: 1.8124749177885533
R2: 0.700167944545735
```

Рисунок 4 — Анализ регрессионных остатков

8. С помощью библиотеки statmodels получить «эконометрический» результат обучения модели линейной регрессии. Проинтерпретировать все его составляющие (в т.ч. те, которые изучались только теоретически), сравнить с предыдущими результатами.

С помощью библиотеки statsmodels был построен “эконометрический” результат обучения модели линейной регрессии.

```
MSE: 3.285065327612621
RMSE: 1.8124749177885526
R2: 0.700167944545735
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.700
Model:                            OLS   Adj. R-squared:                  0.627
Method:                 Least Squares   F-statistic:                     9.600
Date:                Tue, 03 Oct 2023   Prob (F-statistic):           2.27e-07
Time:                        21:11:07   Log-Likelihood:                -94.641
No. Observations:                  47   AIC:                             209.3
Df Residuals:                      37   BIC:                             227.8
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         58.2362      0.298    195.444      0.000      57.632      58.840
x1            -2.1048      0.572     -3.681      0.001      -3.263      -0.946
x2            -2.4213      0.395     -6.138      0.000      -3.221      -1.622
x3             0.9575      0.407      2.350      0.024       0.132       1.783
x4            -2.1488      0.530     -4.056      0.000      -3.222      -1.075
x5            -0.3227      0.350     -0.922      0.363      -1.032       0.387
x6            -0.3057      0.411     -0.743      0.462      -1.139       0.527
x7            -1.1587      0.440     -2.636      0.012      -2.049      -0.268
x8            -0.8405      0.484     -1.737      0.091      -1.821       0.140
x9            -0.7941      0.392     -2.025      0.050      -1.589       0.001
==============================================================================
Omnibus:                        9.315   Durbin-Watson:                   2.222
Prob(Omnibus):                  0.009   Jarque-Bera (JB):               22.239
Skew:                           0.019   Prob(JB):                     1.48e-05
Kurtosis:                       6.370   Cond. No.                         4.81
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Рисунок 5 — Результат “эконометрического” расчета

Ниже представлена расшифровка полученных данных.

Dep. Variable - имя зависимой переменной – Y;

Model - вид обученной модели – OLS (Ordinary least squares), при поиске оптимальных весов стремимся уменьшить квадрат ошибки.

Method - используемый метод обучения – least squares, см. пункт выше.

Date, Time - дата и время обучения

No. Observations - количество наблюдений в выборке (n) – 47, соответствует входному набору X.

Df Residuals - степень свободы *(количество наблюдений (n) - количество переменных (k) - 1)* – 47 - 9 - 1= 37.

Df Model - количество переменных (регрессоров) (k) – 9 (x_1,x_2)

Covariance Type - Тип ковариации – по умолчанию

R-squared - коэффициент детерминации R2 – 0.7, соответствует ранее обученным моделям

Adj. R-squared - R2, с внесенным “штрафом” за большое количество зависимых переменных (корреляции).

F-statistic - критерий Фишера - метрика, значение которой применяется при проверке гипотезы о равенстве дисперсий.

Prob. (F-statistic) - вероятность того, что коэффициенты при всех переменных равны нулю.

Log-likelihood - метрика, показывающая, насколько хорошо данные описываются моделью.

AIC BIC - используются для сравнения различных моделей. Сами по себе значения этих показателей не несут никакой информации. Однако, они могут быть использованы для выбора наилучшей модели, исходя из степени ее переобученности. Чем меньше значение AIC или BIC, тем лучше модель, при этом BIC сильнее штрафует модели за наличие дополнительных параметров, которые не влияют на качество предсказания.

Omnibus описывает нормальность распределения остатков w_0 – 9.315.

Prob (Omnibus) - является статистическим тестом, который измеряет вероятность того, что остатки (то есть разница между предсказанными и реальными значениями) имеют нормальное распределение. Если результат равен 1, то это означает, что распределение остатков идеально нормальное.

Skew (Перекос) - это мера симметрии распределения остатков, где 0 означает идеальную симметрию.

Kurtosis (Эксцесс) - измеряет остроту распределения остатков (в экстремуме) или его концентрацию около 0 на нормальной кривой. Более высокий эксцесс означает меньшее количество выбросов.

Durbin-Watson - критерий для проверки наличия автокорреляции. При отсутствии автокорреляции значение критерия находится между 1 и 2 – коэффициент равен 2.222, следовательно корреляция в данных присутствует.

Jarque-Bera (JB) и **Prob (JB)** - альтернативные методы измерения того же значения, что и Omnibus и Prob (Omnibus), с использованием асимметрии (перекос) и эксцесса. Нулевая гипотеза – распределение является нормальным, асимметрия равна нулю, а эксцесс равен трем. При небольших выборках тест Jarque-Bera склонен отклонять нулевую гипотезу когда она верна.

Cond. No (число обусловленности) — показывает чувствительность выходных данных к изменению входных данных (в т.ч. указывает на наличие мультиколлинеарности) – 4.81, аналогично полученным для других моделей данным.

Исходя из полученных метрик, можно сделать вывод, что применение модели линейной регрессии для приведенных данных (и, возможно, для всех аналогичных данных) несёт серьёзные риски.

9-10. Сравнить качество получаемых моделей на основе коэффициента детерминации и MSE. Сделать итоговый вывод касательно причин различия в результатах при выполнении работ разными методами, а также по получаемым моделям в целом. Провести сравнительный анализ.

Для каждой из ранее полученных моделей была посчитаны метрики MSE, RMSE, R2.

Таблица 1 – сравнительный анализ моделей линейной регрессии

	Аналитическая модель	Градиентная модель	sklearn-модель
MSE	3.2850653276126236	3.285107199403924	3.2850653276126227
RMSE	1.8124749177885535	1.8124864687505735	1.8124749177885533
R2	0.7001679445457348	0.7001641228545357	0.700167944545735

Незначительные различия коэффициентов детерминации можно объяснить погрешностью системы при различных методах вычисления. MSE и RMSE, полученные градиентным методом, были достаточно точными, но за счет большого количества итераций. При меньшем количестве точность снижается и уступает точности аналитического метода.

Это может быть связано с наличием корреляций в данных, хотя они и незначительны, а число обусловленности в пределах нормы, метод градиентного спуска может быть более чувствительным к этому. Кроме того, метод градиентного спуска лучше всего показывает себя на больших наборах данных, в то время как в лабораторной работе использовался входной набор в 47 записей.

Вывод: В результате выполнения лабораторной работы были изучены методы построения и оценки моделей линейной регрессии.