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
Вариант 15

Название: Прогнозирование моделью линейной регрессии

Дисциплина: Прикладной анализ данных

Студент ИУ6-55Б _____ В.К. Полубояров
(Подпись, дата) (И.О. Фамилия)

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

Москва, 2023

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.

Исходя из варианта были подготовлены исходные данные - отобраны строки, соответствующие регионам: Центральный, Центрально-Черноземный, Северо-Кавказский, Западно-Сибирский, Восточно-Сибирский, Дальневосточный районы.

С помощью библиотеки Pandas были получены данные в виде массивов, где X - целевые признаки, Y - целевые значения.

Для корректной работы с данными необходимо нормировать данные. Для нормирования была выбрана “Стандартизация”, которую часто называют Z-оценкой. Она рассчитывается с помощью формулы, отдельно для каждого x.

После нормализации необходимо добавить единичный столбец (x0), так как мы работаем с моделью линейной регрессии.

```
X = pd.DataFrame(X)
for column in X.columns:
  X[column] = (X[column] - X[column].mean()) / X[column].std()
ones_column = np.ones((X.shape[0], 1))
X = np.hstack((ones_column, X))
X
```

Рисунок 1 - Z-нормализация и добавление единичного столбца.

2. Расчет весов линейной регрессии по аналитической формуле.

Для получения весов линейной регрессии используется аналитическая формула:

$$w = (X^T X)^{-1} X^T Y$$

Для подсчета используется нормированная матрица целевых признаков X.

```
# получение ковариационной матрицы
cov_mat = np.matmul(np.transpose(X), X)

# получение матрицы -1
inv_mat = np.linalg.inv(cov_mat)

# расчет весов
mat = np.matmul(inv_mat, np.transpose(X))
weights = np.matmul(mat, Y) # получение матрицы весов

print("Полученные веса:", "\n", weights)
```

```
Полученные веса:
 [[58.13061224]
 [-2.40255485]
 [-2.4500539 ]
 [ 0.67582994]
 [-2.13092791]
 [-0.59362218]
 [-0.39965211]
 [-1.06941919]
 [-1.09418984]
 [-0.98486064]]
```

Рисунок 2 - расчет весов по аналитической формуле.

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

Для построения корреляционной матрицы был использован метод .corr() библиотеки Pandas. Для большей наглядности полученного результата корреляционная матрица была представлена в виде тепловой карты с использованием библиотеки seaborn.

```
corr_plot = sns.heatmap(pd.DataFrame(X).corr(), cmap="coolwarm", annot=True)
```

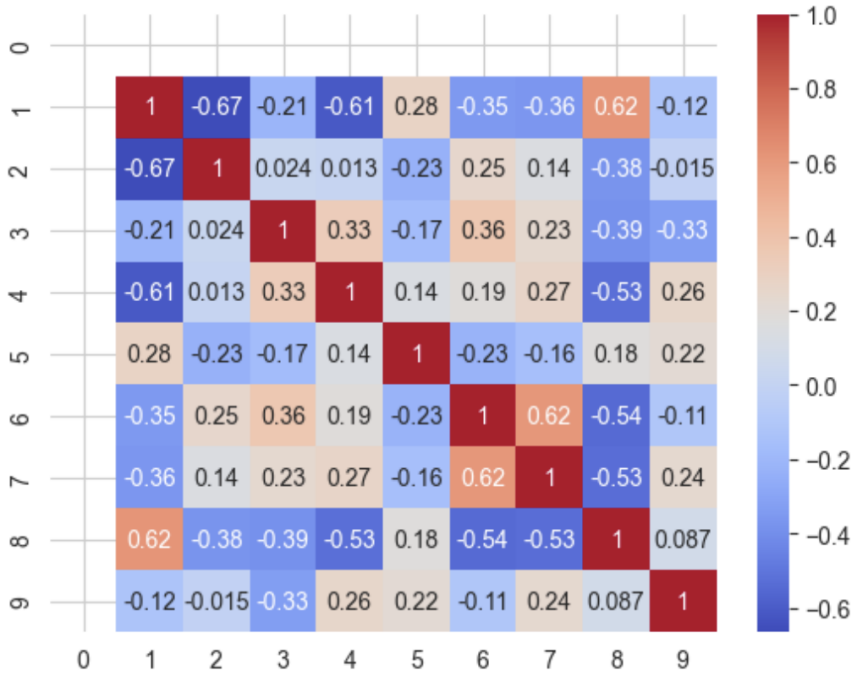

Рисунок 3 - корреляционная матрица признаков.

Посмотрим, какие градации степени корреляции существуют.

Таблица анализа силы связи между переменными

Значение	Интерпретация
от 0 до 0,3	очень слабая
от 0,3 до 0,5	слабая
от 0, 5 до 0,7	средняя
от 0,7 до 0, 9	высокая
от 0,9 до 1	очень высокая

Рисунок 4 - сила связи между переменными

Самая высокая корреляция в наших данных между признаками x1 и x2 = -0.67, что считается средней силой взаимосвязи.

x1 - Рождаемость населения на 1000 человек.

x2 - Смертность населения на 1000 человек.

В нашем случае, чем выше рождаемость, тем ниже смертность.

Также существует взаимосвязь между x1 и x8 = 0.62, где x8 - численность населения с денежными доходами ниже прожиточного минимума в % от численности населения. Получается, что чем выше рождаемость, тем больше людей в регионе имеют доход меньше прожиточного минимума.

Исходя из полученных корреляций можно предположить, что так как у нас более высокая рождаемость, чем смертность, то происходит рост населения. Скорее всего, в регионах недостаточное количество рабочих мест, что приводит к тому, что все бОльшее количество людей имеют доход ниже прожиточного минимума.

Затем было подсчитано число обусловленности, с помощью метода cond().

```
print(f"Число обусловленности: {np.linalg.cond(X, p=2)}")
pd.DataFrame(cov_mat)
```

```
Число обусловленности: 7.167690720885988
```

Рисунок 5 - число обусловленности

Полученное число обусловленности - 7.17. Считается, что мультиколлинеарности нет, если число обусловленности меньше 10.

Отсутствие мультиколлинеарности говорит о том, что между переменными отсутствует критическая корреляция. Следовательно, наши данные можно считать подходящими для обучения линейной регрессии.

4. Анализ регрессионных остатков.

Для анализа регрессионных остатков требуется найти предсказанные значения целевой переменной (Y_pred). Для этого необходимо умножить полученные веса на этапе 2 на матрицу нормализованных признаков X. На основе полученных предсказаний посчитаем RMSE - Среднеквадратическое отклонение.

```
y_pred = np.matmul(X, weights)
MSE = mean_squared_error(Y, y_pred)
RMSE = MSE ** 0.5
print("RMSE: ", RMSE)

R2_analytics = r2_score(Y, y_pred)
print(f"Коэффициент детерминации: {R2_analytics}")
```

```
RMSE:  1.159729049686305
Коэффициент детерминации: 0.8234681200381894
```

Рисунок 6 - среднеквадратическое отклонение

Полученный результат RMSE = 1.16. Среднее значение реального значения Y = 58.13.

Это значит, что результат нашей модели, полученный с помощью аналитической формулы, является достаточно точным, так как предсказанное значение попадает в диапазон +- 1.16 от реального значения, что является не очень большим отклонением даже от минимального значения Y в наших данных (49.7).

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

Для определения весов линейной регрессии градиентом методом была составлена функция, описывающая алгоритм:

1. Инициализация весов (единичный вектор)
2. Расчет предсказанных значений Y
3. Расчет градиента функции потерь
4. Пересчет весов
5. Повторение этапа 2-4 по кругу n раз.

В нашем случае критерием остановки стало количество итераций, равное 1000. Скорость обучения была взята равной 0.1.

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

for i in range(1000):
  # 2. Расчет предсказанного значения y_pred по весам w
  y_pred = np.matmul(X, weights)
  delta = Y - y_pred

  # 3. Расчет градиента функции потерь
  curr = 0
  for i in range(Y.shape[0]):
    curr += (pow(Y[i] - y_pred[i], 2))

  S.append((1 / X.shape[0]) * curr)

  # 4. Установка новых  весов
  dS = (-2 / X.shape[0]) * np.transpose(np.matmul(np.transpose(delta), X))
  weights -= learning_rate * dS

for i in range (len(S)):
  print(f"Итерация {i + 1} - {S[i][0]}")
```

Рисунок 7 - расчет весов градиентным методом.

Существенное изменение ошибки прекратилось на 233 итерации. Итоговым значение на 1000 итерации S = 1.3449714686990657.

RMSE = 1.15972905.

6. Сравнение результатов по аналитическому и градиентному методу.

	Аналитический метод	Градиентный метод
RMSE	1.15972904968631	1.15972905
Коэффициент детерминации	0.8234681200381894	0.8234681200365137

Исходя из полученных значений RMSE и коэффициента детерминации можно сделать вывод, что результаты двух методов идентичны, а

небольшие различия в числах можно считать погрешностью при вычислениях.

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

Для обучения модели была использована библиотека sklearn. Для вычисления предсказанных значений Y был использован метод библиотеки predict. Для сравнения полученной модели с результатами предыдущих этапов был посчитан параметр RMSE.

	Аналитический метод	Градиентный метод	sklearn
RMSE	1.15972904968631	1.15972905	1.15972904968306

Полученный результат можно считать идентичным двум предыдущим методам построения модели.

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.

Для получения “эконометрического” результата обучения модели воспользуемся библиотекой statmodels и методом ols().

```
stat_model = sm.OLS(Y, X).fit()
y_pred_from_stat = stat_model.predict(X)
MSE = mean_squared_error(Y, y_pred_from_stat)
RMSE = MSE ** 0.5
print('RMSE: ', RMSE)
print(stat_model.summary())
```

Рисунок 8 - код построения эконометрического результата.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.783
Method:	Least Squares	F-statistic:	20.21
Date:	Wed, 04 Oct 2023	Prob (F-statistic):	4.44e-12
Time:	20:57:44	Log-Likelihood:	-76.789
No. Observations:	49	AIC:	173.6
Df Residuals:	39	BIC:	192.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	58.1306	0.186	313.026	0.000	57.755	58.506
x1	-2.4026	0.510	-4.712	0.000	-3.434	-1.371
x2	-2.4501	0.376	-6.523	0.000	-3.210	-1.690
x3	0.6758	0.242	2.798	0.008	0.187	1.164
x4	-2.1309	0.432	-4.937	0.000	-3.004	-1.258
x5	-0.5936	0.236	-2.518	0.016	-1.070	-0.117
x6	-0.3997	0.267	-1.498	0.142	-0.939	0.140
x7	-1.0694	0.289	-3.704	0.001	-1.653	-0.485
x8	-1.0942	0.314	-3.481	0.001	-1.730	-0.458
x9	-0.9849	0.252	-3.913	0.000	-1.494	-0.476

Omnibus:	1.704	Durbin-Watson:	1.790
Prob(Omnibus):	0.426	Jarque-Bera (JB):	1.583
Skew:	-0.331	Prob(JB):	0.453
Kurtosis:	2.419	Cond. No.	7.17

Рисунок 9 — Результат “эконометрического” расчета

Исходя из полученного отчета видим:

- Имя зависимой переменной - Y.
- Вид модели - Ordinary least squares, уменьшающий квадрат ошибки.
- Метод обучения - метод наименьших квадратов.
- Время создания отчета.
- Количество наблюдений - 49
- Степень свободы - 39 (разница между количеством наблюдений и количеством признаков).
- Количество признаков - 9 (без учета единичного столбца)

- Вид ковариации - ковариация, которая неустойчива к гетероскедастичности ошибки.) Гетероскедастичность - неоднородность наблюдений, выражающаяся в непостоянной дисперсии случайной ошибки регрессионной модели, т.е. заключается в разбросе значений случайной величины относительно её математического ожидания.
- Коэффициент детерминации - 0.89 (Он показывает, какая доля дисперсии результативного признака объясняется влиянием независимых переменных). В нашем случае модель считается хорошей, так как коэффициент близок к единице. Наша модель описывает вариативность около 90% данных.
- Скорректированное значение коэффициента детерминации - 0.78. (R2, с штрафом за большое количество зависимых переменных и соответственно, корреляций)
- Критерий Фишера - метрика, значение которой применяется при проверке гипотезы о равенстве дисперсий.
- p-уровень для критерия Фишера - вероятность того, что коэффициенты при всех переменных равны нулю. Обычно уровень значимости равен 0.05. Так как наше число меньше 0.05, значит мы можем отвергать нулевую гипотезу о значимости модели.
- Log-likelihood – метрика, показывающая, насколько хорошо данные описываются моделью.
- AIC и BIC – используются для сравнения различных моделей. Параметры 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.

- Следующая таблица описывает гипотезы о значимости коэффициентов. В столбце P > |t| если < 0.05, то гипотеза о значимости коэффициента принимается (отвергается нулевая). В случае нашей модели мы не можем отвергнуть гипотезу для x6, следовательно данная переменная не вносит вклад в нашу модель.
- Omnibus описывает нормальность распределения наших остатков. Чем ближе его значение к 0, тем ближе к нормальности распределения остатков.
- Prob (Omnibus) - это статистический тест, измеряющий вероятность нормального распределения остатков. Значение 1 означает совершенно нормальное распределение. В нашем случае этот показатель равен 0.42
- Skew (Перекос) - это мера симметрии распределения остатков, где 0 означает идеальную симметрию.
- Kurtosis (Эксцесс) - измеряет остроту распределения остатков (в горбу) или его концентрацию около 0 на нормальной кривой. Более высокий эксцесс означает меньшее количество выбросов.
- Durbin-Watson - это критерий для проверки наличия автокорреляции При отсутствии автокорреляции значение критерия находится между 1 и 2. В нашем случае автокорреляция отсутствует.
- 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. В нашем случае данный показатель **нерепрезентативен**, так как выборка **мала**.
- Cond. No (число обусловленности) — это мера чувствительности нашей модели по отношению к входящим данным. То есть при

малейшем изменении данных коэффициенты сильно меняются. В нашем случае коэффициент равен 7, что может говорить об отсутствии мультиколлинеарности.

$$\mu(X) = \frac{\lambda_{max}}{\lambda_{min}}$$

Рисунок 10 - формула расчета числа обусловленности.

9. Сравнить качество получаемых моделей на основе коэффициента детерминации и MSE.

Для всех методов обучения модели коэффициент детерминации был равен примерно 0.82, RMSE – 1.16.

Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 0,5, поэтому полученные модели можно считать достаточно хорошими.

Малое значение RMSE относительно среднего значения целевого признака говорит об хорошей точности модели.

10. Сделать итоговый вывод касательно причин различия в результатах при выполнении работ разными методами, а также по получаемым моделям в целом. Провести сравнительный анализ.

	Аналитический метод	Градиентный метод	sklearn
RMSE	1.15972904968631	1.15972905	1.159729049686306

R^2	0.8234681200381894	0.8234681200365137	0.823468120038189

По таблице видно, что метрики, рассчитанные с помощью аналитического метода и полученные для Skylearn-модели, отличаются друг от друга очень мелкими величинами. Данные числовые различия могут быть объяснены погрешностями системы при вычислениях.

Вывод: в ходе выполнения лабораторной работы были изучены методы построения линейных регрессионных моделей с использованием аналитического метода и регрессионного спуска. Также был проведен анализ моделей на основе различных метрик.