Дисциплина «Информационные технологии» Рабочая тетрадь № 4

Регрессия – это зависимость среднего значения какой-либо величины от некоторой другой величины или от нескольких других величин. В отличие от чисто функциональной зависимости 𝑦 = 𝑓(𝑥), когда каждому значению независимой переменной 𝑥 соответствует одно определённое значение зависимой переменной у, при регрессионной связи одному и тому же значению независимой переменной (фактору) 𝑥 могут соответствовать в зависимости от конкретного случая различные значения зависимой переменной (отклика) у.

Изучение регрессии основано на том, что случайные величины Х и 𝑌 связаны между собой вероятностной зависимостью: при каждом конкретном значении 𝑋 = 𝑥 величина 𝑌 является случайной величиной с вполне определённым распределением вероятностей. Зависимость зависимой переменной – отклика от одной независимой переменной – фактора или нескольких факторов называется уравнением регрессии. По количеству факторов выделяют парную (однофакторную) и множественную (многофакторную) регрессию. Для парной будем рассматривать следующие методы регрессии: линейную, показательную, экспоненциальную, гиперболическую и параболическую.

Регрессионный анализ – это раздел математической статистики, изучающий регрессионную зависимость между случайными величинами по статистическим данным. Цель регрессионного анализа состоит в определении общего вида уравнения регрессии, вычислении оценок неизвестных параметров, входящих в уравнение регрессии проверке статистических гипотез о регрессионной связи.

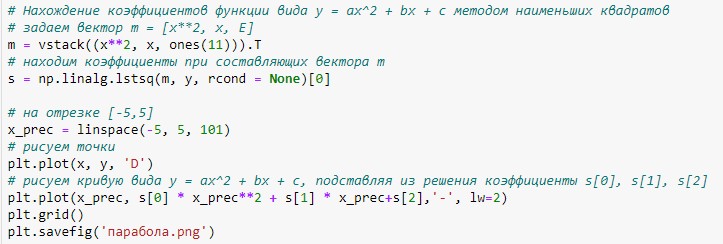
Таким образом, регрессионный анализ – набор статистических методов исследования влияния одной или нескольких независимых переменных 𝑋1, … , 𝑋𝑛 на зависимую переменную 𝑌. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные – критериальными переменными.

|  |
| --- |
| **1.1. Теоретический материал – Линейные регрессионные модели** |
| Линейная регрессия  Линейная регрессия (Linear regression) – модель зависимости переменной x от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости. Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных и стала простым предшественником нелинейных методов, которые используют для обучения нейронных сетей. |

|  |  |
| --- | --- |
| Цель линейной регрессии — поиск линии, которая наилучшим образом соответствует этим точкам. Напомним, что общее уравнение для прямой есть  𝑓 (𝑥) = 𝑏 + 𝑚 ⋅ 𝑥 +, где 𝑚 – наклон линии, а 𝑏 – его сдвиг.  **Функция потерь — метод наименьших квадратов**  Функция потерь – это мера количества ошибок, которые наша линейная регрессия делает на наборе данных. Хотя есть разные функции потерь, все они вычисляют расстояние между предсказанным значением 𝑦(х) и его фактическим значением.  Одна очень распространенная функция потерь называется средней квадратичной ошибкой MSE. Чтобы вычислить MSE, мы просто берем все значения ошибок, считаем их квадраты длин и усредняем.  **Задача экраполяции**  Допустим у нас есть много экспериментальных точек. Необходимо через них провести кривую, которая как можно ближе проходила к этим точкам. При этом необходимо минимизировать среднюю квадратичную ошибку (MSE).  Для решения данной задачи в Python есть множество библиотек. Самыми распостраненными выступают:  **numpy - numpy.linalg.lstsq**  **scipy - scipy.linalg** (содержит все функции из numpy.linalg плюс часть новых функций, которых нет в numpy.linalg). | |
| **1.1.1 Пример** | |
| ***Задача:*** | |
|  | Проведем прямую 𝑦 = 𝑚𝑥 + 𝑏 через экспериментальные точки. |
| ***Решение:*** | |
|  |  |
|  |  |
|  |  |
| ***Ответ:*** | |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
| **1.1.2 Пример** | |
| ***Задача:*** | |
|  | Пусть 𝑥, 𝑦 – вектора длиной 𝑛 > 3 (точек > 3). Задача заключается в построении эстраполяционного полинома второго порядка (параболы). Таким образом, необходимо найти такие коэффициенты поринома 𝑎, 𝑏, 𝑐 по методу наименьших квадратов. Данные мтогут быть получены в результате измерений. Покажем пример генерации данных случайным образом и  загрузки их из файла. |
| ***Решение:*** | |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  | |
| ***Ответ:*** | |
|  |  |
|  |  |
|  | **1.1.3 Пример** |
|  | ***Задача:*** |
|  | По данным предыдущего примера постройте эстраполяционного полинома  третьего порядка |
|  | ***Решение:*** |





|  |  |
| --- | --- |
|  |  |
|  |  |
| ***Задание:*** | |
|  | Представьте собственные данные и постройте эктраполяцию полиномами  первой, второй и третьей степени. |

# Решение:



***Ответ:***

|  |
| --- |
| **1.1.4 Пример** |
| ***Задача:*** |
| Необходимо проверить гипотезу, что наши точечно заданная функция ложится  на кривую вида 𝑓(𝑥, 𝑏) = 𝑏0 + 𝑏1𝑒𝑥𝑝(−𝑏2𝑥2) |
| ***Решение:*** |
|  |

|  |
| --- |
|  |
|  |
|  |
| ***Ответ:*** |
|  |
|  |
|  |

|  |
| --- |
|  |

|  |
| --- |
| **1.1.5 Пример** |
| ***Задача:*** |
| Необходимо проверить гипотезу, что наши точечно заданная функция ложится на кривые вида:  1) 𝑓(𝑥, 𝑏) = 𝑏0 + 𝑏1𝑥  2) 𝑓(𝑥, 𝑏) = 𝑏0 + 𝑏1𝑥 + 𝑏2𝑥2  3) 𝑓(𝑥, 𝑏) = 𝑏0 + 𝑏1𝑙𝑛(𝑥)  4) 𝑓(𝑥, 𝑏) = 𝑏0 𝑥𝑏1 |
| ***Решение:*** |
|  |

|  |
| --- |
|  |
|  |
|  |

|  |
| --- |
|  |
|  |
|  |

|  |
| --- |
|  |
| ***Ответ:*** |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
|  |
|  |
|  |
| ***Задание:*** |
| Подставьте собственные данные и поэкспериментируйте с представленными  функциями. Проанализируйте динамику изменения данных. |
| ***Решение:*** |
| ***Ответ:*** |

|  |
| --- |
| **1.2. Теоретический материал – Задачи регрессии** |
| **Линейная регрессия** - это широко используемый метод статистического анализа, который использует регрессионный анализ в математической статистике для определения количественной взаимосвязи между двумя или более переменными. Если регрессионный анализ включает две или более независимых переменных, а связь между зависимой и независимой |

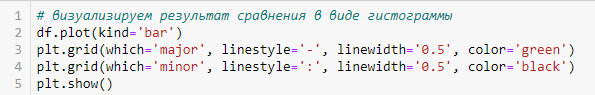
|  |
| --- |
| переменными является линейной, тогда имееи дело с множественной линейной регрессией.  В этом разделе мы увидим, как библиотеку Scikit-Learn в Python для машинного обучения можно использовать для реализации функций регрессии. Мы начнем с простой линейной регрессии с участием двух переменных, а затем перейдем к линейной регрессии с участием нескольких переменных. |

|  |  |
| --- | --- |
|  |  |
| **1.2.1 Пример** | |
| ***Задача:*** | |
|  | Построим простую линейную регрессию в Python с использованием  библиотеки scikit-learn |
| ***Решение:*** | |
|  |  |
|  |  |
|  |  |
|  |  |
| ***Ответ:*** | |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |
|  | |
|  | После того как мы получили представление о данных, разделим информацию на «атрибуты» и «метки». Атрибуты – это независимые переменные, а метки  – это зависимые переменные, значения которых должны быть предсказаны. В нашем наборе всего два столбца и необходимо предсказать оценку в зависимости от количества часов. Чтобы извлечь атрибуты и метки, выполните следующий скрипт: |
| ***Решение:*** | |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
| ***Ответ:*** | |
|  |  |
|  |
|  |
|  | |
|  | Получившийся результат можно интерпретировать следующим образом: с каждым затраченным часом на обучение результат экзамена повышается приблизительно на 17 баллов. Далее можно построить прогнозы. Для этого мы будем использовать наши тестовые данные и посмотрим, насколько точно наш алгоритм предсказывает процентную оценку. Чтобы сделать прогноз на  тестовых данных необходимо выполнить следующий код: |
|  | |

# Решение:



|  |
| --- |
|  |
| ***Ответ:*** |
|  |
|  |
| ***Задание:*** |
| Постройте модель линейной регрессии для произвольных данных из двух столбцов. Для примера можно взять точечную зависимость заработной платы от опыта работы:  (https://raw.githubusercontent.com/AnnaShestova/salary-years-simple-linear- regression/master/Salary\_Data.csv).  Найдите коэффициенты линии регрессии. Постройте прогноз. |
| ***Решение:*** |
| ***Ответ:*** |

|  |
| --- |
| **1.3. Теоретический материал – Множественная регрессия** |
| В предыдущем примере мы проиллюстрировали линейную регрессию с двумя переменными. Однако, почти все реальные задачи имеют больше параметров. Линейная регрессия с участием нескольких переменных называется «множественной линейной регрессией» или многомерной линейной регрессией. Шаги для выполнения множественной линейной регрессии аналогичны шагам для простой . Разница заключается в оценке. Вы можете использовать множественную регрессию, чтобы узнать, какой фактор оказывает наибольшее влияние на прогнозируемый результат или  как различные переменные связаны друг с другом. |

|  |  |
| --- | --- |
| **1.3.1 Пример** | |
| ***Задача:*** | |
|  | Для решения задачи множественной регрессии можно задействовать уже  известный метод numpy.linalg.lstsq. |
| ***Решение:*** | |
|  |  |
| ***Ответ:*** | |
|  |  |
|  | |
|  | Кроме этого можно использовать возможности библиотеки sсikit-learn.  Рассмотрим пример. |
|  | |
| **1.3.2 Пример** | |
| ***Задача:*** | |
|  | Для данных из предыдущей задачи построить модель множественной  линейной регрессии с использованием средств библиотеки sсikit-learn. |
| ***Решение:*** | |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| ***Ответ:*** | |

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | | |
| **Задание** | | |
| ***Задача:*** | | |
|  | | Постройте модель множественной линейной регрессии для произвольных данных из нескольких столбцов. Для примера можно взять потребления газа (в миллионах галлонов) в 48 штатах США или набор данных о качестве красного вина (1) и (2) соответственно. Найдите коэффициенты  множественной регрессии. Постройте прогноз. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1.  https://raw.githubusercontent.com/likarajo/petrol\_consumption/master/data/pe trol\_consumption.csv  2. https://raw.githubusercontent.com/aniruddhachoudhury/Red-Wine-  Quality/master/winequality-red.csv | | | | | | | |
| ***Решение:*** | | | | | | | | |
|  | | | | | | | | |
| **Задание\*** | | | | | | | | |
| ***Задача:*** Экспериментально получены N − значений величины Y при различных значениях величины X. Построить полиномы первой и второй степени, аппроксимирующие результаты эксперимента, с применением метода наименьших квадратов. Результаты выводятся в виде таблиц  значений и графиков, полученных полиномов. | | | | | | | | |
|  | | **Варианты заданий:**  Вариант 1 Вариант 2  x y x y  0,0 3,0 0,0 5,0  0,2 6,0 0,2 5,0  0,4 3,0 0,4 4,0  0,6 6,0 0,6 4,0  0,8 4,0 0,8 6,0  1,0 3,0 1,0 6,0  Вариант 3 Вариант 4  x y x y  3,0 2,0 3,0 6,0  3,2 3,0 3,2 2,0  3,4 3,0 3,4 6,0  3,6 3,0 3,6 4,0  3,8 2,0 3,8 3,0  4,0 4,0 4,0 4,0  Вариант 5 Вариант 6 | | | | | | |
|  | x | y |  | x | y |  |
| 5,0 | 2,0 | 4,0 | 4,0 |
| 5,2 | 4,0 | 4,2 | 3,0 |
| 5,4 | 4,0 | 4,4 | 6,0 |
| 5,6 | 3,0 | 4,6 | 6,0 |
| 5,8 | 3,0 | 4,8 | 4,0 |
| 6,0 | 3,0 | 5,0 | 4,0 |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Вариант 7 | | |  | Вариант 8 | | |  |  |
| x | | | y | x | | | y |
| 1,0 | | | 2,0 | 5,0 | | | 3,0 |
| 1,2 | | | 6,0 | 5,2 | | | 2,0 |
| 1,4 | | | 4,0 | 5,4 | | | 5,0 |
| 1,6 | | | 4,0 | 5,6 | | | 2,0 |
| 1,8 | | | 2,0 | 5,8 | | | 2,0 |
| 2,0 | | | 5,0 | 6,0 | | | 3,0 |
| Вариант 9 | | |  | Вариант 10 | | |  |
|  | x | y | |  | x | y | |  |
| 2,0 | 4,0 | | 0,0 | 6,0 | |
| 2,2 | 2,0 | | 0,2 | 3,0 | |
| 2,4 | 4,0 | | 0,4 | 2,0 | |
| 2,6 | 2,0 | | 0,6 | 6,0 | |
| 2,8 | 5,0 | | 0,8 | 2,0 | |
| 3,0 | 2,0 | | 1,0 | 5,0 | |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |