**Лабораторная работа № 1. Ч.2**

**Часть 2. Обобщенный метод наименьших квадратов.**

Цель: закрепить теоретические сведения о линейной регрессии, получить практические навыки построения модели линейной регрессии.

Задачи:

* Вспомнить и усвоить основные понятия моделей линейной регрессии;
* Научиться строить модели линейной регрессии с использованием обобщенного метода Наименьших квадратов;
* Научиться использовать библиотеку statsmodels;

Обобщенный метод наименьших квадратов

Обобщенный метод наименьших квадратов (ОМНК, или Generalized Least Squares, GLS) — это метод оценки параметров линейной регрессионной модели, который используется, когда предположения о стандартном методе наименьших квадратов (МНК) не выполняются. В частности, ОМНК применяется в случаях, когда ошибки модели имеют гетероскедастичность (разные дисперсии) или коррелированы.

Для нахождения коэффициентов регрессии при ОМНК мы используем следующую формулу:

где – ковариационная матрица ошибок модели

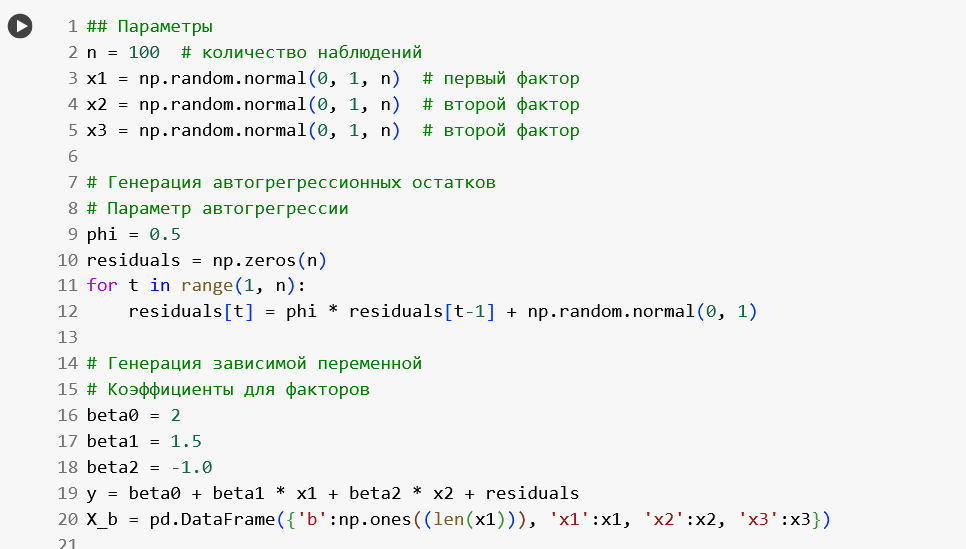
Для построения Данной модели необходимо рассчитать матрицу ковариаций.

Для начала импортируем все необходимые библиотеки:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рассмотрим пример. Сгенерируем данные с авторегрессией остатков:



Далее построим регрессию стандартным методом наименьших квадратов:

Изображение выглядит как текст, Шрифт, снимок экрана, белый

Автоматически созданное описание

Получим остатки этой модели.



Далее необходимо получить оценку характера автокорреляционной связи остатков. Обычно предполагается, что эта зависимость может быть выражена моделью авторегрессии первого порядка (AR(1)). Это означает, что ошибку можно выразить следующим образом:

где - коэффициент автокорреляции ошибки первого порядка.

Для нахождения построим модель линейной регрессии для остатков.

*Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание*

Коэффициент будет параметром при независимой переменной. Запишем его в переменную:

Изображение выглядит как текст, Шрифт, дизайн

Автоматически созданное описание

Как мы знаем, процесс AR(1) означает, что близкие соседи имеют более сильную связь, поэтому мы можем задать эту структуру, используя матрицу Тёплица.



Матрица Тёплица — это квадратная матрица, в которой элементы зависят только от разности индексов. То есть, элемент матрицы aij​ зависит только от i−j. Это означает, что значения на каждой диагонали, параллельной главной диагонали, одинаковы.

Для построения этой матрицы необходимо импортировать ее из библиотеки scipy.

Далее строим матрицу ковариаций:



В данном примере мы возвели коэффициент автокорреляции в степени из матрицы Тёплица и записывает в массив.

Тогда коэффициенты регрессии, согласно ОМНК, можно определить так:



То же самое можно выполнить с использованием библиотеки statsmodels

Изображение выглядит как текст, Шрифт, снимок экрана, линия

Автоматически созданное описание

Задание

1. Создайте набор данных с зависимой переменной y и независимыми переменными x1 и x2. Добавьте автокорреляцию в остатки.
2. Используйте метод наименьших квадратов для оценки модели и выведите результаты.
3. Используйте метод обобщенных наименьших квадратов для оценки модели и выведите результаты.
4. Постройте график остатков для обеих моделей, чтобы визуально оценить их распределение.
5. Загрузите данные boston.csv. Создайте DataFrame библиотеки pandas, используя следующий код:

import pandas as pd

df = pd.read\_csv('boston.csv')

print(df.head())

1. Выполните изучение распределения данных. Визуализируйте зависимости. Для этого рассчитайте коэффициенты корреляции и визуализируйте их с помощью тепловых карт. Для этого используйте следующий код:

import seaborn as sns

sns.heatmap(df.corr().round(1), annot = True, cbar=False)

plt.show()

1. Разделите данные на X и y:

X = df[[‘CRIM’, ‘ZN’, ‘INDUS’, ‘CHAS’, ‘NOX’, ‘RM’, ‘AGE’, ‘DIS’, ‘RAD’, ‘TAX’, ‘PTRATIO’, ‘B’, ‘LSTAT’]]

y = df[‘MEDV’]

1. Постройте модель множественной линейной регрессии с использованием OLS и GLS с использованием библиотеки statsmodels.
2. Постройте модель с использованием GLS для данных по недвижимости.
3. Сделайте выводы о качестве построенной модели.

1. **Для синтетических данных**:

* **OLS**:
  + R2=0.695: Около 69.5% дисперсии зависимой переменной объясняется независимыми переменными.
  + Коэффициент автокорреляции (ρ=0.398) показывает, что в остатках есть положительная автокорреляция, что нарушает предположения OLS.
* **GLS**:
  + R2=0.748: GLS объясняет больше дисперсии, так как учитывает структуру автокорреляции.
  + Остатки модели GLS более стабильны, что делает её результаты более надёжными.

2. **Для данных Boston**:

* **OLS**:
  + R2=0.741: 74.1% дисперсии цен на жильё (MEDV) объясняется независимыми переменными.
  + Основные факторы:
    - Положительное влияние: RM (количество комнат), B, ZN.
    - Отрицательное влияние: LSTAT (процент низкого социального класса), DIS (удалённость от центра), TAX (налоги).
* **GLS**:
  + R2=0.680: GLS объясняет немного меньше дисперсии, но лучше справляется с автокорреляцией остатков.

3. **Общий вывод**:

* **OLS** подходит для данных без автокорреляции или слабо выраженной автокорреляцией.
* **GLS** является предпочтительным методом, если в данных присутствует автокорреляция, так как обеспечивает более надёжные оценки параметров.