Курсовая работа

«Регрессионный анализ»

1. Теоретическая часть

Разрывный дизайн

https://colorful-smelt-ead.notion.site/Regression-Discontinuity-Design-a6b019a0c0ca42769961ee2849cc110b

2. Практическая часть

Работа выполняется с помощью R, Python, Matlab, C++, Java. Можно использовать готовые функции из библиотек или написать свои.

Линейная регрессия. Библиотека statsmodels

Построение простой линейной регрессии в библиотеке statsmodels

2.1. Модельная часть

Смоделировать данные самостоятельно в соответствии с вариантом

$$X_k = f(h_k) + e_k, \quad k=1,\dots,60$$

где e_k — независимый случайные величины с распределением $N(0,\sigma^2)$.

Точки внутри носителя для h выбирать равномерно.

Смоделировать тестовую выборку объема 40, половина значений правее наблюдаемых значений, половина левее.

2.2. Метод наименьших квадратов

Для регрессии вида

$$X_k = \theta_0 + \theta_1 h_k + e_k, \quad k=1,\dots,60 \qquad (1)$$

- Найти МНК-оценки неизвестных параметров.
- Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.
- Вычислить коэффициент детерминации и найти оценку ковариационной матрицы МНК-оценки.
- Найти значения информационных критериев [3]
- С помощью критерия Фишера проверить гипотезу $\theta_0=0$, $\theta_1=0$

- Построить доверительный интервал надежности 0.95 и 0.8 для полезного сигнала $X = \theta_1 + \theta_2 h$ при h из исходного носителя $\pm 50\%$.
- Построить оценку метода наименьших модулей, отобразить ее на графике
- Оценить качество построенных регрессий на тестовой выборке

Оценить качество построенных регрессий с помощью LOO_CV
https://robjhyndman.com/hyndsight/loocv-linear-models/

Для остатков $\hat{e}_k = X_k - \hat{X}_k$

- Построить гистограмму, ядерную оценку плотности распределения
- По остаткам проверить гипотезу, что $\hat{e}$ имеет гауссовское распределение с помощью одного из критериев
  - критерий Шапиро-Уилка (Shapiro–Wilk) [1];
  - критерий D'Agostino K^2 [1];
  - критерий Харке–Бера (Jarque–Bera) [1].
- Проверить наличие автокорреляции с помощью критерия Дарбина-Уотсона.
- Проверить наличие гетероскедастичности с помощью одного из критериев.

Выводы.

2.3. Полиномиальная регрессия

Построить регрессию с помощью МНК

$$X = \theta_0 + \theta_1 h + \theta_2 h^2 + \ldots + \theta_p h^p$$

Порядок полинома p подбирать несколькими способами:

- по значению среднеквадратической погрешности МНК-оценки (на обучающей и/или тествой)
- по значению статистики критерия Фишера для гипотезы $\theta_p = 0$
- по MSE на тестовой выборке
- ваш способ

Выбираем единственное значение p.

Провести анализ остатков по схеме из пункта 2.2.

Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.

Проверить для подобранной модели является ли матрица $H^T H$ мультиколлиниарной, если да, то построить оценку параметров с помощью метода редукции (ридж-оценка).

Выводы

2.4. Регрессия для наблюдений с выбросами

Смоделировать ошибки для модели регрессии (1) с помощью распределения Тьюки, приняв долю выбросов $\boldsymbol{\delta = 0.08}$, номинальную дисперсию $\boldsymbol{\sigma_0^2 = \sigma^2}$, дисперсию аномальных наблюдений $\boldsymbol{\sigma_1^2 = 100\sigma^2}$.

Построить МНК-оценку неизвестных параметров для модели (1) и оценить ее качество.

Провести анализ остатков по схеме из пункта 2.2.

Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.

Провести отбраковку выбросов и пересчитать МНК-оценку и оценить качество оценки.

Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.

Провести анализ остатков по схеме из пункта 2.2.

Построить оценку метода наименьших модулей.

Построить график, на котором отобразить наблюдения, исходную функцию и линию регрессию.

Провести анализ остатков по схеме из пункта 2.2.

* Построить робастную оценку Хубера [3] (дополнительное задание)

Выводы

2.5. Квантильная регрессия

Смоделировать несимметричные ошибки для исходных данных, заменив у 90% отрицательных ошибок знак с минуса на плюс.

Построить МНК и МНМ оценки для получившихся наблюдений и регрессии (1).

Построить несколько квантильных регрессий (для разных значений параметра $\boldsymbol{\alpha}$) и оценить их качество.

Построить график, на котором отобразить наблюдения, исходную функцию и линии регрессии.

Выводы

Дополнительная литература

[1] Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006.

[2] Построение простой линейной регрессии в библиотеке statsmodels
https://habr.com/ru/articles/690414/

[3]
https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F_%D0%BF%D0%BE%D1%82%D0%B5%D1%80%D1%8C_%D0%A5%D1%8C%D1%8E%D0%B1%D0%B5%D1%80%D0%B0

[4]
https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9 (разд. информационные критерии в частных случаях)

[5] Метод борьбы с выбросами https://en.wikipedia.org/wiki/Random_sample_consensus

[6] Учебная литература по теме «Регрессионный анализ»

Содержание отчета

1. Титульный лист.
2. Оглавление.
3. Текст задания с вариантом
4. Теоретическая часть
5. Практическая часть
   5.1. Вычисления
   5.2. Графики
   5.3. Выводы
6. Список использованной литературы.

Требования к оформлению отчета

1. Оформление курсовой работы на стандартных листах формата А4(210х297) в печатном или рукописном виде.
2. Листы должны быть пронумерованы, сложены в один файл не скрепленными.
3. На титульном листе должны быть: название "Курсовая работа", тема работы, ФИО студента, группа, ФИО преподавателя, у которого выполняется работа.
4. При оформлении работы в печатном виде графики выполняются в линейном масштабе из расчета один график на один лист формата А4, ориентация листа альбомная (landscape). При оформлении работы в рукописном виде графики выполняются на листах миллиметровой бумаги в формате А4.
5. Материал работы должен располагаться только с одной стороны листа бумаги.