Санкт-Петербургский государственный университет
Прикладная математика, программирование и искусственный интеллект

Отчет по учебной практике 1 (научно-исследовательской работе) (семестр 2)

Временные ряды, тренд, прогноз (метод наименьших квадратов)

Работа выполнена на очень хорошем уровне и может быть зачтена с оценкой А

Выполнил:

Баранов Иван Александрович, группа 23.Б05-мм

Научный руководитель:

д.ф.-м.н., профессор

Голяндина Нина Эдуардовна

Кафедра статистического моделирования

Санкт-Петербург
2024

Содержание

1 Введение

В этой работе исследуется возможность применения модели линейной регрессии для выявления закономерностей и трендов во временных рядах.

Анализ временных рядов играет ключевую роль в понимании динамики данных, позволяя прогнозировать будущие значения, обнаруживать аномалии.

Применение регрессионного анализа, как одного из самых доступных и эффективных методов, позволяет определить зависимость между зависимой переменной и независимой переменной, что способствует более глубокому пониманию временных рядов и принятию более обоснованных решений в различных областях.

2 Теория линейной регрессии

2.1 Линейная регрессия с двумя параметрами

Пусть задан временной ряд последовательностями значений (x_i, y_i). Для начала построим функцию вида y = ax + b, которая лучше всего приближает временной ряд. В качестве меры того, насколько хорошо наша функция описывает данный временной ряд, будем считать MSE - Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{ipred})^2$$

Или подробнее

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

Наилучшее приближение достигается при минимизации MSE по параматрам a и b

$$MSE(a, b) \to min$$

Необходимое условие экстремума в точке (a, b):
Частные производные по переменным a и b должны равняться нулю

$$\begin{cases} \frac{\delta F}{\delta a} = 2\sum_{i=1}^{n}(ax_i^2 + bx_i - x_i y_i) \\ \frac{\delta F}{\delta b} = 2\sum_{i=1}^{n}(ax_i + b - y_i) \end{cases}$$

Приравняем к 0 и перепишем в удобном виде

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Откуда методом крамера для СЛУ найдем a и b

$$\Delta a = \begin{bmatrix} \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & n \end{bmatrix}$$

$$\Delta b = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \end{bmatrix}$$

$$\Delta = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$a = \frac{\Delta a}{\Delta} \quad b = \frac{\Delta b}{\Delta}$$

2.2 Линейная регрессия с центрированием

Линейную регрессию можно построить по усредненным данным, это позволяет упростить формулы, уменьшить масштаб переменных, сводить более сложные задачи построения регрессии к более простым. Покажем что новые формулы для линейной регрессии с центрированием дают тот же результат, что и старые.
Вычтем из данных среднее, посчитаем коэффициенты как и раньше

$$x_{ci} = x_i - \hat{y}$$
$$y_{ci} = y_i - \hat{y}$$

Где средние вычисляются как

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Тогда формула для a

$$\Delta a_c = \begin{bmatrix} \sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{x}) & 0 \\ 0 & n \end{bmatrix}$$

$$\Delta_c = \begin{bmatrix} \sum_{i=1}^{n}(x_i - \hat{x})^2 & 0 \\ 0 & n \end{bmatrix}$$

$$a = \frac{\Delta a_c}{\Delta_c} = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sum_{i=1}^{n}(x_i - \hat{x})^2}$$

Преобразуем ее

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \hat{y}\sum_{i=1}^{n} x_i - \hat{x}\sum_{i=1}^{n} y_i + n\hat{x}\hat{y}}{\sum_{i=1}^{n} x_i^2 - 2\hat{x}\sum_{i=1}^{n} x_i + n\hat{x}^2}$$

Откуда домножением на n получаем

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = a_{old}$$

Таким образом новое решение совпало со старым

2.3 Линейная регрессия высшего порядка

Рассмотрим линейную регрессию с произвольным количеством признаков.
Будем аппроксимировать наш ряд функцией

$$f(x, w_0, \ldots, w_k) = w_0 + w_1 x + w_2 x^2 + \cdots + w_k x^k$$

Запишем признаки в матрицу (поскольку признак в нашей задаче только один будем брать различные его степени)

$$X = \begin{bmatrix} 1 & x_1 & \ldots & x_1^k \\ 1 & x_2 & \ldots & x_2^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^k \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$B = \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix}$$

задача в матричном виде

$$Y = XB$$

МНК для такой задачи

$$MSE(w_0, \ldots, w_k) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, w_0, \ldots, w_k)) \to min$$

координаты минимума считаются по формуле

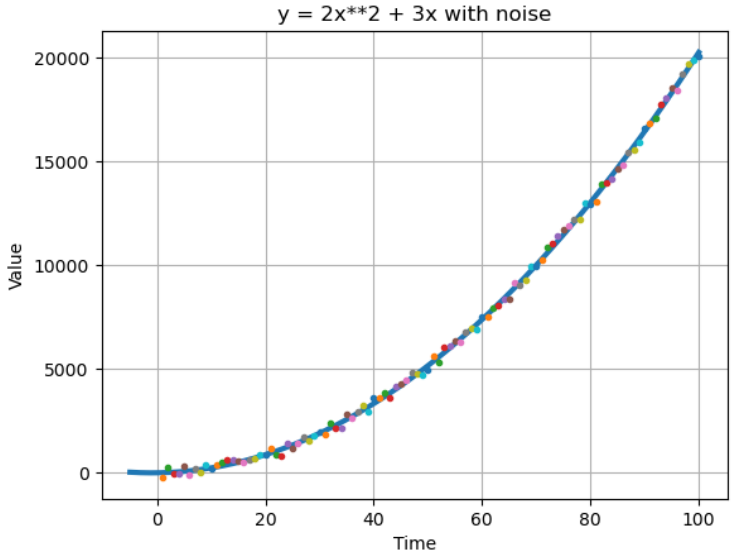
$$B = (X^T X)^{-1} X^T Y$$

3 Реализация на си

Реализованные функции:

- LinReg2Dim - линейная регрессия с двумя параметрами
- LinReg2DimCenter - линейная регрессия с двумя параметрами и центрированием
- Mul, Tr, Minor, Det, Inv - функции для работы с матрицами
- LinRegPolynom - линейная регрессия произвольного порядка
- Predict - предсказать значения по признакам и коэффициентам
- Mse - считает MSE по target и предсказанным значениям
- LinF - генерирует линейную выборку с шумом
- SqF - генерирует квадратичную выборку с шумом

Рис. 1: пример работы с синтетическим датасетом



4 Тест на реальных данных

В качестве данных выбраны продажи в трех сферах бизнеса в США по месяцам с 1992 по 2023 год. Данные с 2020 по 2023 год я исключил, из-за сильного влияния covid, так же выбрал те сферы, на которые не сильно влиял кризис 2008 года.

Изначально я выделял промежуток времени 2016-2019 под тестовую выборку, строил регрессии различного порядка на данных 1992-2015. При таком подходе единственный параметр для оценки качества построенной модели - MSE на тренировачных данных. На практике он работает плохо, поскольку модель подгоняется под тренировочные данные с ростом степени полинома, что ухудшает предсказание на test.

Другой подход - разбить данные на 3 части: train (1992-2012), valid (2013-2015), test (2016-2019). Строим модель по train, считаем MSE по предсказаниям на valid. Выбираем такую из построенных, чтобы

MSE на valid был минимальным, выбранной моделью делаем предсказания на test. Этот подход показал себя намного лучше.

Рис. 2: пример построенной регрессии и ее прогноза, красными чертами отделены train (1992-2012), valid (2013-2015), test (2016-2019)

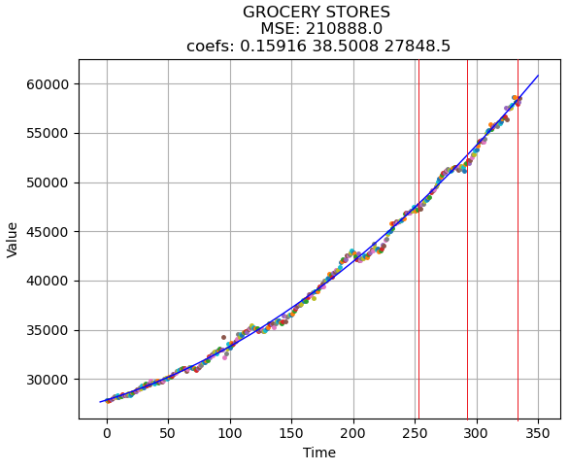


Рис. 3: пример построенной регрессии и ее прогноза, красными чертами отделены train (1992-2012), valid (2013-2015), test (2016-2019)

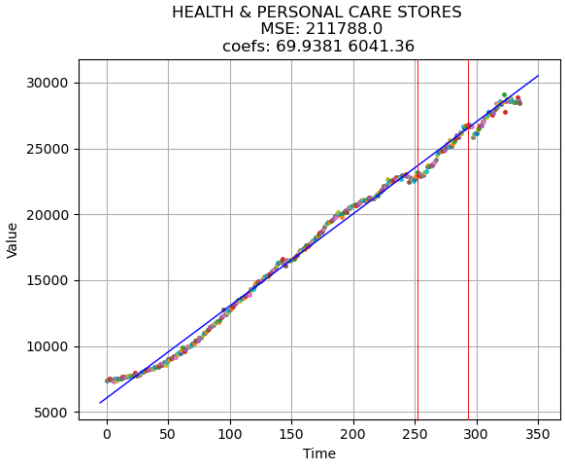


Рис. 4: пример построенной регрессии и ее прогноза, красными чертами отделены train (1992-2012), valid (2013-2015), test (2016-2019)

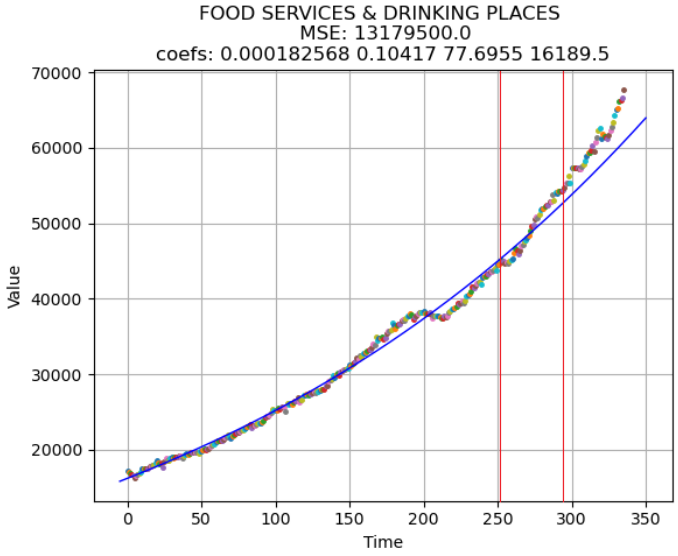


Рис. 5: график MSE на valid и test

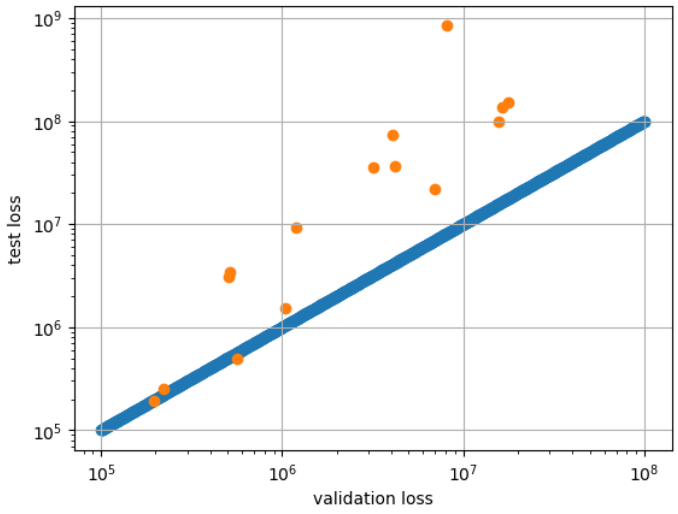


Рис. 6: MSE на valid и test на различных данных

GROCERY STORES

degree	mse train	mse valid	mse test
1	7.173278e+05	7.000507e+06	4.023901e+06
2	1.848661e+05	1.962644e+05	1.860947e+05
3	1.847610e+05	2.201929e+05	1.952669e+05
4	1.836070e+05	5.062557e+05	6.427377e+05

HEALTH & PERSONAL CARE STORES

degree	mse train	mse valid	mse test
1	2.456009e+05	5.635153e+05	2.847312e+05
2	2.369213e+05	1.035655e+06	4.382478e+05
3	4.055942e+04	3.190301e+06	5.616981e+06
4	3.572123e+04	1.187346e+06	1.477776e+06

FOOD SERVICES & DRINKING PLACES

degree	mse train	mse valid	mse test
1	9.877797e+05	1.554213e+07	1.625075e+07
2	5.990574e+05	4.198061e+06	6.223914e+06
3	4.411362e+05	1.775340e+07	2.451582e+07
4	4.408523e+05	1.627045e+07	2.172669e+07

Также я рассмотрел информационные критерии AIC и BIC, которые накладывают штраф на количество параметров. Они считаются на соединенном датасете train + valid.

$$AIC = \frac{2k}{n} + ln(n * MSE)$$
$$BIC = \frac{k * ln(n)}{n} + ln(n * MSE)$$

Рис. 7: AIC и BIC

GROCERY STORES

degree	AIC	BIC	mse test
1	1.396143e+01	1.398104e+01	1.149930e+07
2	1.212053e+01	1.215974e+01	2.030522e+05
3	1.211739e+01	1.217621e+01	2.482833e+05
4	1.211755e+01	1.219598e+01	2.751409e+05

HEALTH & PERSONAL CARE STORES

degree	AIC	BIC	mse test
1	1.248064e+01	1.250025e+01	2.156954e+05
2	1.247248e+01	1.251169e+01	1.557149e+05
3	1.175243e+01	1.181125e+01	5.064807e+06
4	1.092236e+01	1.100078e+01	1.297699e+06

FOOD SERVICES & DRINKING PLACES

degree	AIC	BIC	mse test
1	1.457568e+01	1.459528e+01	6.614376e+07
2	1.360962e+01	1.364883e+01	1.697693e+07
3	1.359803e+01	1.365685e+01	1.236120e+07
4	1.318083e+01	1.325926e+01	9.202550e+06

Оказалось, что их штраф в сравнении с ошибкой слишком мал.

Я пробовал распознать экспоненциальный рост на этих данных. Вместо предсказания исходной целевой переменной построим регрессию на предсказание ее логарифма. На новых данных просто возведем e в степень предсказания и получим исходное предсказание.

Рис. 8: предсказание логарифма целевой переменной

GROCERY STORES

degree	mse train	mse valid	mse test
1	2.959407e+05	1.756089e+06	3.937050e+06
2	1.909468e+05	2.510131e+05	1.828630e+06
3	1.845004e+05	2.377046e+05	4.262392e+05
4	1.837297e+05	6.338215e+05	5.725399e+06

HEALTH & PERSONAL CARE STORES

degree	mse train	mse valid	mse test
1	1.035157e+06	2.250086e+07	7.745309e+07
2	1.635430e+05	5.820805e+05	2.991281e+05
3	8.038618e+04	1.004510e+07	9.496715e+07
4	4.839323e+04	4.306133e+05	5.063254e+07

FOOD SERVICES & DRINKING PLACES

degree	mse train	mse valid	mse test
1	9.130235e+05	1.711420e+06	1.591794e+06
2	6.114501e+05	3.268444e+06	3.085373e+07
3	4.596284e+05	2.889279e+07	2.765133e+08
4	4.409419e+05	2.098746e+07	1.785229e+08

В целом, получилось чуть хуже чем в базовом варианте, хотя отдельные предсказания получились точнее.

5 Заключение

В данной курсовой работе была проведена оценка эффективности линейной регрессии и метода наименьших квадратов для анализа временных рядов и выявления трендов. Исследование охватило основные методы построения моделей линейной регрессии и оценки тренда с помощью метода наименьших квадратов.

Результаты работы показали, что линейная регрессия и метод наи-

меньших квадратов могут быть эффективными инструментами для анализа временных рядов. Однако для достижения большей точности необходимо учитывать специфику данных, проводить дополнительный анализ и рассматривать применение других моделей.

6 Список литературы

А. Емелин. Математика для заочников.
http://www.mathprofi.ru/metod_naimenshih_kvadratov.html
«cleverstudents.ru» - доступная математика.
http://www.cleverstudents.ru/articles/mnk.html
сайт кафедры СтатМод
https://statmod.ru/wiki/