МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ С ИСПОЛЬЗОВАНИЕМ ОРТОГОНАЛЬНЫХ ФУНКЦИЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Муллиной Арины Александровны

Научный руководитель
доцент, к. ф.-м. н., доцент _____ В. В. Новиков

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2022

ВВЕДЕНИЕ

Термин «регрессия» был предложен Ф. Гальтоном в конце XIX в. Он обнаружил, что дети родителей с высоким или низким ростом обычно не наследуют выдающийся рост, и назвал этот феномен «регрессия к посредственности». Сначала этот термин использовался исключительно в биологическом смысле. После работ К. Пирсона его стали использовать и в статистике. Регрессионный анализ - метод моделирования измеряемых данных и исследования их свойств.

Актуальность темы. В эконометрике широко используются методы статистики. В практических задачах аппроксимации и прогнозирования, изучая различные связи в экономических, производственных системах, необходимо на основании экспериментальных данных выразить зависимую переменную в виде некоторой математической функции от независимых переменных - регрессоров, то есть построить регрессионную модель.

Целью магистерской работы является изучение математических основ построения и анализа регрессионных моделей и теории ортогональных систем функций применительно к регрессионному анализу, знакомство с основами среды программирования R применительно к задачам регрессионного анализа.

Объект исследования регрессионное моделирование с использованием ортогональных функций.

Предмет исследования практическая реализация задач регрессионного моделирования с использованием ортогональных функций.

Для достижения поставленной цели в работе необходимо решить следующие **задачи:**

— изучить модели линейной и полиномиальной регрессий; |item рассмотреть основные понятия теории временных рядов и модели авторегрессии;
— изучить матрицы плана с ортогональными столбцами;
— рассмотреть полиномы, ортогональные на системе точек;
— определить основные понятия среды программирования R;
— реализовать на практике аппроксимацию данных с использованием ор-

тогональных функций;

— построить прогноз, используя пакет языка R.

Практическая значимость исследования состоит в том, что на основе практических реализаций изученных формул можно аппроксимировать не только данные заболеваемости коронавирусом, но и другие данные обладающие плохой обусловленностью при использовании МНК для оценивания.

Структура и содержание магистерской работы. Работа состоит из введения трех разделов, заключения, списка использованных источников, состоящего из двадцати наименований, и двух приложений.

Работа прошла апробацию на ежегодной студенческой конференции «Актуальные проблемы математики и механики», которую проводил механико-математический факультет СГУ в апреле 2022 года, в секции «Анализ данных», в X Международной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», проводимой в ноябре 2021 года.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы работы, формируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

В первом разделе рассмотрены математические основы построения и анализа регрессионных моделей.

В математической статистике под регрессией случайной величины y на случайную величину x понимают условное математическое ожидание $E(y|x)$.

Предполагается, что условное математическое ожидание y при заданном X является линейной функцией, неизвестной с точностью до коэффициентов $\alpha = (\alpha_1, \ldots, \alpha_m)' \in R^m : E(y|X) = X\alpha$.

Предположения регрессии:

1 Изначально ничего не известно о параметре α, но иногда определенная информация все же существует.

2 Зависимая переменная имеет стохастическую природу.

3 Схема "математической регрессии" обходится без определения отклонений ε.

4 Ковариационная матрица y при фиксированной матрице X имеет вид $\sigma^2 I_n$, то есть

$$\mathrm{cov}(y|X) = E\{(y - X\alpha)(y - X\alpha)'/X\} = \sigma^2 I_n.$$

5 X - случайная матрица.

6 $\mathrm{rank}X = m$ с вероятностью 1.

Распределение матрицы X может быть известно или неизвестно с точностью до конечного числа неизвестных параметров.

Модель простой линейной регрессии

$$y = \alpha_0 + \alpha_1 x + \varepsilon,$$

где α_0 - отступ и α_1 - наклон являются неизвестными постоянными, а ε - случайная ошибка.

Предполагается, что ошибки некоррелированы и пренебрежимо малы.

Среднее $E(y|x) = \alpha_0 + \alpha_1 x$ и дисперсию $V(y|x) = V(\alpha_0 + \alpha_1 x + \varepsilon) = \sigma^2$. То есть среднее y есть линейная функция x. Параметры α_0 и α_1 называют коэффициентами регрессии. Наклон α_1 - это изменение в среднем значении распределения y, произведенное единичным изменением в x. Если диапазон данных по x содержит $x = 0$, то отступ α_0 - это среднее распределения отклика y при $x = 0$. Если диапазон x не содержит нуль, то α_0 не имеет практической интерпретации.

Если в общей многомерной линейной модели регрессии положить $x_{ij} = x_i^j$, то получится полиномиальная модель k-й степени:

$$Y_i = \beta_0 + \beta_1 x_i^1 + \ldots + \beta_k x_i^k + \varepsilon_i,\ (i = \overline{1,n},\ k \leq n-1).$$

При предположении, что x_i приблизительно равномерно распределить на отрезке $[0;1]$, тогда при больших n матрица $X'X$ плохо обусловлена. Степень плохой обусловленности матрицы $X'X$ измеряется ее числом обусловленности $K[X]$, которое определяется как отношение наибольшего сингулярного значения матрицы X к ее наименьшему ненулевому сингулярному значению. Сингулярное значение матрицы X- это положительные квадратные

корни из собственных значений матрицы $X'X$. Свойства чисел $K[X]$:

1 $K[X'X] = (K[X])^2$,

2 Так как $X'X = U'U$, имеем $K[U] = K[X]$.

Процесс регистрации исходных статистических данных происходит по времени t и время фиксируется наряду со значениями анализируемых характеристик $x_i^j(t_k)(j = 1, 2, \ldots, p; i = 1, 2, \ldots, n; k = 1, 2, ..., N)$, то говорят о статистическом анализе панельных данных. Если зафиксировать номер переменной j и номер статистически обследуемого объекта i, то расположенную в хронологическом порядке последовательность значений

$$x_i^j(t_1), x_i^j(t_2), \ldots, x_i^j(t_k)$$

называют одномерным временным рядом. Если же одновременно рассматривать p одномерных временных рядов, т. е. исследовать закономерности во взаимосвязанном поведении временных рядов для $j = 1, 2, \ldots, p$, характеризующих динамику переменных, измеренных на каком-то одном (i-м) объекте, тогда говорят о статистическом анализе многомерного временного ряда

$$X(t) = \left(x^1(t_k), x^2(t_k), \ldots, x^p(t_k)\right)^T \ (k = 1, 2, \ldots, N).$$

Для удобства представим временной ряд в виде

$$x(1), x(2), \ldots, x(N),$$

где $x(t)$ - значение исследуемого показателя, зарегистрированное в $t-$ моменте времени $(t = 1, 2, \ldots, N)$.

Так же временным рядом называется ряд наблюдений $x(t_1), x(t_2), \ldots, x(t_N)$ анализируемой случайной величины $\xi(t)$, произведенных в последовательные моменты времени $t_1, t_2, \ldots, t_N$.

Принципиальные отличия временного ряда от последовательности наблюдений $_1, _2, \ldots, x_n$, образующих случайную выборку:

— в отличие от элементов случайной выборки члены временного ряда не являются статистически независимыми;

— члены временного ряда не являются одинаков о распределенными, то

есть $P\{x(t_1) < x\} \neq P\{x(t_2) < x\}$ при $t_1 \neq t_2$.

Значит нельзя распространять свойства и правила статистического анализа случайной выборки на временные ряды. Но при этом взаимозависимость членов временного ряда создает свою специфическую базу для построения прогнозных значений анализируемого показателя.

Во втором разделе описывались приложения ортогональных систем функций к регрессионному анализу.

Ряд способов преодоления плохой обусловленности связан с использованием в том или ином виде ортогональных систем. Рассматривается регрессионную модель

$$Y_i = a_0\varphi_0(x_i) + a_1\varphi_1(x_i) + \ldots + a_m\varphi_m(x_i) + \varepsilon_i, \qquad (1)$$

где $\varphi_l(x_i)$ - полином l-й степени от $x_i \quad (l = 0, 1, ..., m)$. Предполагается, что полиномы ортогональны на множестве значений переменной x, справедливо следующее равенство

$$\sum_{i=1}^{n} \varphi_l(x_i)\varphi_k(x_i) = 0 \quad \text{для всех } l, k; l \neq k.$$

Тогда получается регрессионную модель $Y = Xa + \varepsilon$, где матрица X имеет следующий вид

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_m(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \ldots & \varphi_m(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_m(x_n) \end{pmatrix}$$

столбцы матрицы X взаимно ортогональны и матрица $X'X$ выглядит следу-

ющим образом

$$X'X = \begin{pmatrix} \sum_i \varphi_0^2(x_i) & 0 & \dots & 0 \\ 0 & \sum_i \varphi_1^2(x_i) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sum_i \varphi_m^2(x_i) \end{pmatrix}.$$

Таким образом матрица $X'X$ оказывается диагональной, а такие матрицы, как правило, хорошо обусловлены.

Находятся выражения для ортогональных многочленов Чебышева при заданных точках $x_1, x_2, \dots, x_n$.

Пусть $\varphi_0(x) = 1$. Из условия

$$\sum_{i=1}^{n} \varphi_l(x_i)\varphi_k(x_i) = 0 \quad \text{для всех } l, k; l \neq k,$$

положив $l = 0$, $k = 1$ для многочлена $\varphi_1(x)$, получено

$$\sum \varphi_1(x_i) = 0.$$

Из $\varphi_l = x^l + \alpha_l^{(1)} x^{l-1} + \dots$ следует, что многочлен $\varphi_1(x) = x + \alpha_1$, таким образом

$$\varphi_1(x) = x - \frac{1}{n}\sum x_i.$$

Далее

$$\varphi_2(x) = (x + \beta_2)\varphi_1(x) + \gamma_2\varphi_0(x).$$

Получено

$$\beta_2 = \frac{-\sum x_i[\varphi_1(x_i)]^2}{\sum[\varphi_1(x_i)]^2},\ \gamma_2 = -\frac{1}{n}\sum x_i\varphi_1(x_i).$$

Затем приводится рекуррентную формулу, позволяющая вычислять по двум предыдущим многочленам - следующий.

Предполагается, что $\sum x_i^k \varphi_r(x_i) = 0$ для всех $k < r$. Все многочлены

$\varphi_0(x)$, $\varphi_1(x), \ldots, \varphi_r(x)$ построены

$$\begin{cases} \varphi_0(x) = 1, \\ \varphi_1(x) = x + \alpha_1, \\ \varphi_2(x) = x^2 + \alpha_2^{(1)} x + \alpha_2^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \varphi_k(x) = x^k + \alpha_k^{(1)} x^{k-1} + \ldots + \alpha_k^{(k)}, \\ \varphi_r(x) = x^r + \alpha_r^{(1)} x^{r-1} + \ldots + \alpha_r^{(r)}. \end{cases}$$

Отсюда находится

$$x = \varphi_1(x) - \alpha_1 \varphi_0(x),$$

$$x^2 = \varphi_2(x) - \alpha_2^{(1)} \varphi_1(x) + (\alpha_2^{(1)} \alpha_1 - \alpha_2^{(2)}) \varphi_0(x), \text{ и т.д.}$$

В результате

$$x^k = \varphi_k(x) + b_{k-1} \varphi_{k-1}(x) + \ldots + b_0 \varphi_0(x).$$

Доказывается, что многочлен $\varphi_{r+1}(x_i)$ можно представить рекуррентной формулой

$$\varphi_{r+1}(x) = (x + \beta_{r+1}) \varphi_r(x) + \gamma_{r+1} \varphi_{r-1}(x),$$

если $k < r - 1$, то

$\sum \varphi_k(x_i) \varphi_{r+1}(x_i) = \sum (x_i + \beta_{r+1}) \varphi_k(x_i) \varphi_r(x_i) + \gamma_{k+1} \sum \varphi_k(x_i) \varphi_{r-1}(x_i) = 0$. Откуда

$$\beta_{r+1} = -\frac{\sum x_i [\varphi_r(x_i)]^2}{\sum [\varphi_r(x_i)]^2}, \quad \gamma_{r+1} = -\frac{\sum x_i \varphi_{r-1}(x_i) \varphi_r(x_i)}{\sum [\varphi_{r-1}(x_i)]^2}.$$

Если искать приближающий многочлен в виде

$$y = a_0 \varphi_0(x) + a_1 \varphi_1(x) + \ldots + a_m \varphi_m(x),$$

то значения коэффициентов a_0, $a_1, \ldots, a_m$ находятся по формулам

$$a_r = \frac{\sum y_i \varphi_r(x_i)}{\sum [\varphi_r(x_i)]^2} \quad (r = 0,\, 1,\, 2, \ldots).$$

Используя написанные выше формулы, построена полиномиальная аппроксимация тренда заболеваемости коронавирусом COVID-19 в мире, России и Саратовской области за все время пандемии.

Третий раздел посвящен моделированию и анализу регрессионных зависимостей средствами языка R.

Прогнозирование — это наиболее распространенная задача, возникающая при работе с временными рядами.

В 2017 году специалисты компании Facebook объявили о разработанном ими новом пакете для прогнозирования временных рядов — prophet ("пророк").

В основе методологии пакета лежит процедура подгонки аддитивных регрессионных моделей (GAM) вида:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t,$$

где $g(t)$ и $s(t)$ -функции,соответственно аппроксимирующие тренд ряда и сезонные колебания, $h(t)$ - отражение эффектов праздников и других влиятельных событий, а ε_t - нормально распределенные случайные возмущения. Методы, используемые для аппроксимации перечисленных функций:

— тренд: кусочная линейная регрессия или кусочная логистическая кривая роста;

— годовая сезонность: частичные суммы ряда Фурье, число членов которого определяет гладкость функции;

— недельная сезонность: представлена в виде индикаторной переменной;

— "праздники" представлены в виде индикаторных переменных.

Оценивание параметров аппроксимируемой модели выполняется с использованием принципов байесовской статистики (либо методом нахождения апостериорного максимума, либо путем полного байесовского вывода). Для этого применяется платформа вероятностного программирования Stan. Па-

кет prophet - это удобный интерфейс для работы с этой платформой из среды R.

Используя функции из рассмотренного пакета, был получен прогноз для заболеваемости коронавирусом в мире, России и Саратовской области.

В заключении описаны результаты проделанной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Определены основы регрессионного анализа.
2. Изучены линейная и полиномиальная регрессионная модели.
3. Рассмотрены основные понятия теории временных рядов и модели авторегрессии.
4. Рассмотрены матрицы плана с ортогональными столбцами и полиномы, ортогональные на системе точек.
5. Изучены основы среды программирования R.
6. Построена полиномиальная аппроксимация тренда заболеваемости коронавирусом.
7. Построен прогноз заболеваемости коронавирусной инфекцией.