Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций
и стохастического анализа

ИДЕНТИФИКАЦИЯ ВРЕМЕННЫХ РЯДОВ И ВЫБОР ОПТИМАЛЬНОЙ МОДЕЛИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 412 группы
направления 01.03.02 Прикладная математика и информатика
Ариповой Камиллы Сейткалиевны

Научный руководитель
к.ф-м.н, доцент ______________________________ Н.Ю. Агафонова

Зав. кафедрой
д.ф-м.н, профессор ______________________________ С.П. Сидоров

Саратов, 2018

ВВЕДЕНИЕ

Актуальность темы. Ежедневно мы сталкиваемся с явлениями, которые изменяются во времени, т.е. с временными рядами. Примерами таких рядов являются метеорологические условия, цены на тот или иной товар, деловая активность и т.д. Это создает необходимость изучения методов анализа, которые предназначены для выявления структуры временных рядов, а также для их прогнозирования.

Целью бакалаврской работы является исследование и программная реализация методов, используемых для идентификации временных рядов и выбора оптимальной модели.

Объект исследования — временные ряды макроэкономических показателей.

Предмет исследования — методы определения математических моделей временных рядов по результатам их экспериментальных исследований.

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

— определить основные понятия, необходимые для анализа временных рядов;

— изучить стационарные стохастические модели $AR(p)$ и $MA(q)$;

— рассмотреть частные случаи модели $AR(p)$ — процесс Маркова $AR(1)$ и процесс Юла $AR(2)$;

— описать структуру моделирования временного ряда $AR(p)$;

— рассмотреть ряд, построенный по реальным данным макроэкономических показателей;

— рассчитать основные характеристики рассматриваемых временных рядов;

— исследовать качество построенной модели для искусственных и реальных данных;

— найти прогнозные значения.

Практическая значимость проводимого исследования состоит в том, что построенные модели временных рядов можно использовать для:

— прогнозирования поведения временного ряда в будущем на основании прошлых и настоящих значений;

— управления процессом, порождающим ряд;

— выяснения механизма, лежащего в основе процесса;

— удаления компонент временного ряда, которые затемняют его динамику;

— описания характерных особенностей ряда.

Структура и содержание бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников, содержащего 20 наименований, и двух приложений. Общий объем работы составляет 52 страницы.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

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

В **первом** разделе приводятся основные понятия, используемые в анализе рядов.

Случайной величиной называется действительная измеримая относительно борелевской σ-алгебры функция $X(\omega)$.

Функцией распределения случайной величины X называется функция $F_X(x)$, определяющая для каждого значения x вероятность того, что случайная величина X примет значение меньшее, чем x, то есть

$$F_X(x) = P(X < x),\ x \in R.$$

Отображение $\xi(\omega, t) : \Omega \times T \to R$ называется случайным процессом, если $\forall t_0 \in T\ \xi(\omega, t_0)$ удовлетворяет определению случайной величины.

Сечением случайного процесса называется любая случайная величина вида $\xi(\omega, t_0); t_0 \in T$ (фиксируется время). Сечение дает информацию о вероятностном распределении для конкретного момента времени.

Математическим ожиданием случайного процесса $\xi(\omega, t)$ называется функция $m(t) = E[\xi(\omega, t)] : T \to R$, которая каждому $t \in T$ сопоставляется математическое ожидание соответствующего сечения как случайной величины.

Корреляционной функцией случайного процесса $\xi(\omega, t)$ называется функция

$$K_\xi(t_1,t_2) = cov(\xi(\omega,t_1),\xi(\omega,t_2)) =$$
$$= E[(\xi(\omega,t_1) - E(\xi(\omega,t_1))(\xi(\omega,t_2) - E(\xi(\omega,t_2))],\ t_1,t_2 \in T$$

Временной ряд – это последовательность значений некоторого протекающего во времени процесса.

Спецификация модели временного ряда состоит из:

— систематической составляющей (детерминированной последовательности $\{d_t\}_{t\geq 1}$, элементы которой являются функциями времени);

— случайной (иррегулярной) составляющей $\{\varepsilon_t\}_{t\geq 1}$.

Во **втором** разделе рассмотрены стационарные модели авторегрессии и скользящего среднего.

Временной ряд принято моделировать как случайный (стохастический) процесс, который представляет собой статистическое явление, развивающееся во времени по законам теории вероятностей.

Статистический процесс называется строго стационарным (или стационарным в узком смысле), если взаимное распределение вероятностей m наблюдений является инвариантным относительно общего сдвига временного аргумента. Это означает, что при любых целых значениях сдвига k совместная плотность распределения случайных величин $x_{t_1}, x_{t_2}, \ldots, x_{t_m}$ такая же, какая для величин $x_{t_1+k}, x_{t_2+k}, \ldots, x_{t_m+k}$.

Статистический процесс называется слабо стационарным (или стационарным в широком смысле), если случайные величины x_t удовлетворяют следующим условиям:

1. $E(x_t) = \mu = const$;
2. $Var(x_t) = E(x_t - \mu)^2 = \sigma^2 = const$;
3. Ковариация зависит только от величины сдвига k и не зависит от t:

$$cov(x_t, x_{t+k}) = \gamma_k.$$

Величину γ_k называют автоковариацией с лагом (задержкой) k и вычисляют по формуле:

$$\gamma_k = E(x_t - \mu)(x_{t+k} - \mu),$$

Последовательность автоковариаций $\{\gamma_k\}_{k=-\infty,\ldots,+\infty}$ называется автоковариационной функцией, которая является симметричной относительно нуля: $\gamma_k = \gamma_{-k}$, т.е. достаточно рассматривать $k = 0, 1, 2, \ldots$

Величину $\rho_k = \frac{\gamma_k}{\gamma_0}$ называют автокорреляцией k-го порядка, а последовательность $\{\gamma_k\}_{k=-\infty,\ldots,+\infty}$ – автокорреляционной функцией, которая также симметрична, причем $\rho_0 = 1$, поэтому рассматривают $k = 0, 1, 2, \ldots$

Стационарный стохастический процесс x_t с нулевым математическим ожиданием представим в виде линейной комбинации последовательности возмущений $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$, а именно:

$$x_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \ldots = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \tag{1}$$

Выражение (1) можно представить в виде:

$$x_t = (1 + \psi_1 L + \psi_2 L^2 + \ldots)\varepsilon_t,$$

где L – лаговый оператор: $L\varepsilon_t = \varepsilon_{t-1}, L^m \varepsilon_t = \varepsilon_{t-m}$, а также $\psi_0 = 1$ и справедливо условие:

$$\sum_{i=0}^{\infty} |\psi_i| < \infty, \tag{2}$$

т.е. ряд абсолютных значений коэффициентов обязан сходиться.

Выражение вида (1) называют моделью линейного фильтра. Линейный оператор

$$\psi(L) = 1 + \psi_1 L + \psi_2 L^2 + \ldots = \sum_{i=0}^{\infty} \psi_i L_i,$$

называется оператором линейного фильтра.

Для того чтобы модель линейного фильтра имела смысл, должно выполняться условие (2). Тогда оператор линейного фильтра $\psi(L)$ называется устойчивым, а процесс стационарным.

Частным случаем модели линейного фильтра является **модель скользящего среднего MA(q)**:

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q},$$

т.е. x_t линейно зависит от конечного числа q предыдущих значений ε.

Данную модель можно представить в более сжатой форме:

$$x_t = \theta(L)\varepsilon_t,$$

где $\theta(L)$ – оператор скользящего среднего, который вычисляется по формуле:

$$\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_q L^q.$$

Процесс $MA(q)$ стационарен без каких-либо ограничений на параметры θ_j.

Другим частным случаем модели линейного фильтра является **модель авторегрессии AR(p)**. В модели авторегрессии текущее значение процесса x_t представляют как линейную комбинацию конечного числа предыдущих значений процесса и белого шума ε_t:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} + \varepsilon_t, \tag{3}$$

причем текущее значение ε_t не коррелировано с лагами x_t. Такую модель называют авторегрессией p-го порядка и обозначают $AR(p)$.

Процесс Маркова AR(1).

Авторегрессионный процесс первого порядка $AR(1)$, который еще называют процессом Маркова (марковским процессом), имеет следующий вид:

$$x_t = \varphi x_{t-1} + \varepsilon_t, \tag{4}$$

где ε_t - белый шум, который не коррелирует с x_{t-1}.

Условие стационарности процесса Маркова имеет вид:

$$|\varphi| < 1$$

При $0 < \varphi < 1$ автокорреляционная функция экспоненциально затухает, при $-1 < \varphi < 0$ она имеет форму затухающей знакопеременной экспоненты.

При $\varphi > 1$ марковский процесс превращается во «взрывной» процесс.

В случае $\varphi = 1$ наблюдается процесс случайного блуждания, который относят к разряду нестационарных.

Процесс Юла AR(2).

Процессом Юла называется авторегрессия второго порядка $AR(2)$, которая имеет следующий вид:

$$x_t = \varphi x_{t-1} + \varphi_2 x_{t-2} + \varepsilon_t,$$

или, с использованием лагового оператора:

$$(1 - \varphi_1 L - \varphi_2 L^2) x_t = \varepsilon_t.$$

Для того чтобы процесс авторегрессии $AR(2)$ был стационарным, необходимо, чтобы корни λ_1, λ_2 характеристического уравнения

$$1 - \varphi_1 z - \varphi_2 z^2 = 0,$$

которые, вообще говоря, могут быть комплексными, удовлетворяли условиям $|\lambda_1| > 1, |\lambda_2| > 1$. Другими словами, они должны лежать вне единичного круга на комплексной плоскости.

Частная автокорреляционная функция позволяет измерить «чистую» корреляцию между уровнями временного ряда x_t и x_{t-k} при исключении влияния промежуточных уровней ряда и определяется как функция частной автокорреляции от задержки k, где $k = 1, 2, \ldots$

Для $AR(p)$ частная автокорреляционная функция φ_{kk} не равна нулю при $k \leq p$ и имеет нулевое значение при $k > p$, то есть обрывается на задержке, следующей за порядком p. В этом состоит характеристическое свойство процесса авторегрессии $AR(p)$.

В **третьем** разделе построена математическая модель временного ряда $AR(1)$, а также исследовано качество модели для искусственных и реальных данных.

На языке C++ написана программа, которая моделирует временной ряд $AR(1)$ и вычисляет характеристики, необходимые для оценки качества построенной модели.

Входными данными программы являются:

— сгенерированный набор чисел $n = 100$, распределенных по нормальному закону;

— коэффициент авторегрессии $\varphi = 0.5$.

Выходными данными программы являются следующие:

— автокорреляционная функция (АКФ) 1-го и 2-го порядков;

— частная автокорреляционная функция (ЧАКФ) 1-го и 2-го порядков;

— посчитанный коэффициент авторегрессии b;

— значение Q-статистики Бокса-Пирса.

Для реализации модели сгенерирован набор нормальных величин на отрезке [0,1] методом, основанным на центральной предельной теореме. Согласно данной теореме, при сложении достаточно большого количества независимых случайных величин с произвольным законом распределения получится случайная величина, распределенная по нормальному закону.

Для того чтобы определить значение нормально распределенной случайной величины с математическим ожиданием, равным нулю, и дисперсией, равной единице, необходимо сложить равномерно распределенные случайные величины y_i, а из суммы вычесть 6, т.е.:

$$\xi = \sum_{i=1}^{12} y_i - 6.$$

Для выборки из n=100 нормально распределенных случайных чисел по формуле (4) построена модель $AR(1)$ с заданным коэффициентом b=0,5. Полученный временной ряд был проверен на стационарность методом средних уровней.

Для этого был применен следующий алгоритм:

1. Ряд разбивается на две равные по числу членов n_1 и n_2, каждая из которых рассматривается как частная выборка;

2. Для каждой из частей считается среднее значение и дисперсия;

3. Проверяется гипотеза о равенстве дисперсий частных выборок при помощи F-критерия Фишера:

$$F = \frac{s_1^2}{s_2^2} \approx F(n_1 - 1, n_2 - 1)$$

Если $F < F_{\text{кр}}$, то гипотеза о равенстве дисперсий не отвергается. Метод разности средних уровней может быть применен.

4. Вычисляется разность средних: $R = \bar{Y_1} - \bar{Y_2}$

5. Проверяется гипотеза о равенстве средний значений частных выборок при помощи t-критерия Стьюдента:

$$t_R = \frac{R}{S_R} = \frac{R}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Если $t_R < t_{\text{кр}}$, то средние различаются несущественно, т.е. исследуемый процесс является стационарным.

В случае стационарности временного ряда считаются его выборочные АКФ и ЧАКФ 1-го и 2-го порядков по следующим формулам:

АКФ:

$$r = \frac{\sum\limits_{i=2}^{n}(y_t - \overline{y}_1)(y_{t-1} - \overline{y}_2)}{\sqrt{\sum\limits_{i=2}^{n}(y_t - \overline{y}_1)^2 \sum\limits_{i=2}^{n}(y_{t-1} - \overline{y}_2)^2}}$$

ЧАКФ:

$$R_1 = r_1$$

$$R_2 = \frac{r_2 - r_1^2}{1 - r_1^2}$$

Проверка адекватности полученной модели.

В модели, содержащей констату, среднее остатков должно иметь нулевое значение. Выборочная АКФ остатков считается по формуле:

$$r = \frac{\sum\limits_{t=k+1}^{n} e_t e_{t-k}}{\sum\limits_{t=1}^{n} e_t^2}, \quad k = 1, 2, \ldots$$

Q-статистика Бокса-Пирса проверяет гипотезу равенства нулю сразу K первых значений АКФ остатков:

$$Q = n\sum_{k=1}^{K} r_k^2$$

При нулевой гипотезе отсутствия автокорреляции Q имеет распределение $\chi^2(K - p)$, где $K = 3$, $p = 1$. Нулевая гипотеза отвергается, если полученное значение Q больше соответствующего критического значения (в нашем случае, если $Q > 6$).

Для $AR(1)$ значение АКФ первого порядка равно коэффициенту авторегрессии.

В случае нестационарности исследуемого процесса составляется ряд первых разностей:

$$u_t = y_t - y_{t-1}$$

Если полученный временной ряд проходит проверку на стационарность, то исследуемый процесс является интегрированным временным рядом первого порядка ARIMA(1,1,0).

В результате из десяти смоделированных временных рядов восемь являются стационарными процессами $AR(1)$ с коэффициентом, приблизительно равным заданному ($b = 0,5$). Остальные два ряда соотвествуют модели ARIMA(1,1,0). Отсюда можно сделать вывод, что программа работает верно, модель $AR(1)$ построена.

Работоспособность программы была протестирована на реальных значениях временного ряда, составленного по данным из отчета о динамике экспортных цен в США в период с марта 2014 г. по март 2018 г. Данный ряд содержит $n = 70$ членов. В качестве сравнения использованы результаты, построенные в Gretl.

Для исследования стационарности временного ряда был применен тест Дики - Фуллера, определяющий значение коэффициента φ в авторегрессионном уравнении (4).

Тестируя гипотезу $H_0 : \varphi = 1$ против альтернативной гипотезы $H_1 : \varphi = 1$, проверяется предположение о наличии единичного корня. Расчётное значение (-5,56405) оказалось меньше критического (-2,93). Следовательно, рассматриваемый ряд является стационарным.

Согласно построенной коррелограмме процесс соответствует модели авторегрессии первого порядка. В соответствии с построенной моделью в Gretl временной ряд описывается следующим уравнением:

$$Y_t \approx -0,59 + 0,374 \cdot Y_{t-1}$$

Из сравнительного анализа результатов работы программного продукта с результатами анализа в Gretl, видно, что характеристики временного ряда имеют приблизительно одинаковые значения (таблица 1).

Таблица 1. Сравнение результатов анализа работы программы с результатами анализа в Gretl

Хар-ки	Результаты работы программы	Результаты анализа в Gretl
АКФ1	0,373	0,364
АКФ2	0,134	0,131
ЧАКФ1	0,373	0,364
ЧАКФ2	-0,0055	-0,0015
b	0,373	0,374

Основной целью анализа временных рядов является построение прогноза. Прогнозные значения $y_i, 70 < i < 74$, приведены в таблице 2. Как видно из таблицы 2, прогнозируемые данные отличаются от реальных не более чем на 10 %, т.е. погрешность мала. Следовательно, их можно использовать в дальнейшем анализе.

Таблица 2. Прогнозные значения модели $AR(1)$

N	y-реальные	y-прогнозные	Погрешность
71	1,34628	1,28564	0,06064
72	1,07112	1,08006	0,00894
73	0,95613	0,97956	0,02343

Таким образом, созданный программный продукт выполняет идентификацию и построение модели $AR(1)$.

В **заключении** представлены результаты бакалаврской работы.

Основные результаты

1. Определены основные понятия, необходимые для анализа временных рядов.

2. Изучены модели временных рядов $AR(p)$ и $MA(q)$. Рассмотрены частные случаи модели $AR(p)$ — процесс Маркова $AR(1)$ и процесс Юла $AR(2)$.

Изучено понятие стационарности временного ряда, а также приведены условия, при которых стохастические модели являются стационарными. Описано поведение автокорреляционной и частной автокорреляционной функций для рассматриваемых моделей временных рядов.

3. Разработана программа, моделирующая временной ряд $AR(1)$ и позволяющая вычислять основные характеристики ряда. Разработанный программный продукт позволяет также вычислять основные характеристики временного ряда $AR(1)$ на основе теоретических формул. Работоспособность программы была протестирована на реальных данных макроэкономических показателей. В качестве сравнения были рассчитаны основные характеристики ряда в Gretl. Проведен сравнительный анализ полученных характеристик исследуемого процесса. Найдены прогнозные значения рассматриваемого временного ряда.