МОДЕЛИРОВАНИЕ ПРОЦЕССА АВТОРЕГРЕСИИ С ПОМОЩЬЮ ЯЗЫКА ПРОГРАММИРОВАНИЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

Безбочина А.А. Email: Bezbochina1168@scientifictext.ru

Безбочина Александра Андреевна – бакалавр,
кафедра прикладной математики и экономико-математических методов,
Санкт-Петербургский государственный экономический университет, г. Санкт-Петербург

Аннотация: *в статье анализируются возможности статистического пакета анализа данных R для прогнозирования временного ряда. Идентификация модели производится двумя способами: определение стационарности ряда, дифференцирование, анализ автокорреляционной и частной автокорреляционной функций и автоматический подбор статистическим пакетом. Полученные модели анализируются на наличие всех предпосылок (соответствие остатков белому шуму, отсутствие гетероскедастичности и т.д.). Наилучшая модель выбирается исходя из показателей точности и критериев Акаике и Шварца.*
Ключевые слова: *временной ряд, прогнозирование, авторегрессия, методология Бокса-Дженкинса, оптимизация модели.*

MODELING THE AUTOREGRESSION PROCESS USING THE STATISTICAL DATA PROCESSING PROGRAMMING LANGUAGE

Bezbochina A.A.

Bezbochina Aleksandra Andreevna – Bachelor,
DEPARTMENT OF APPLIED MATHEMATICS AND ECONOMIC AND MATHEMATICAL METHODS,
SAINT-PETERSBURG STATE UNIVERSITY OF ECONOMICS, SAINT-PETERSBURG

Abstract: *the article analyzes the capabilities of the statistical data analysis package R for forecasting the time series. Identification of models based on two products: determination of stationary series, differentiation, analysis of autocorrelation and partial autocorrelation functions and automatic selection of a statistical package. The obtained models are analyzed for the presence of all possible ones (correspondence to residual white noise, lack of heteroskedasticity, etc.). The best model is selected based on accuracy indicators and Akaike and Schwartz criteria.*
Keywords: *time series, forecasting, autoregression, Box-Jenkins methodology, model optimization.*

УДК 331.225.3

Очень часто специалистам по анализу данных приходится сталкивать с необходимостью спрогнозировать тот или иной временной ряд. Существует большое количество простых и трудоемких способов это сделать: среднее, экспоненциальное сглаживание, тренд, адаптивные модели и т.д. В данной статье мы рассмотрим другой метод прогнозирования временного ряда – модель ARIMA(p,d,q), реализуем это на языке программирования R.

Модели ARMA(p,q) и ARIMA(p,d,q) стремятся описать автокорреляцию в данных, при этом моделирование процесса условно можно разделить на следующие составляющие: подготовка данных (важный этап, которым не стоит пренебрегать), идентификация модели, т.е. определение параметров p, d и q, анализ модели на наличие всех предпосылок и качество остатков, оценивание с помощью выбранных показателей точности, прогнозирование.

Для моделирования возьмем ряд выручки компании N. Предварительная обработка данных заключалась в удалении выбросов и пропущенных значений, а также делении данных на две части: для построения модели (5 лет или 60 уровней) и сравнения прогноза модели с фактическими данными (1 год или 12 уровней).

Первым шагом в методологии Бокса-Дженкинса является определение стационарности/ нестационарности ряда, взятие последовательных разностей (дифференцирование) нестационарного временного ряда (т.е. определение параметра d – порядка разности) [2].

Проведенные критерии Льюнга – Бокса и критерий KPSS говорят о нестационарности исходного ряда «Выручка», этот вывод можно подтвердить, построив автокорреляционную и частную автокорреляционную функции ряда.

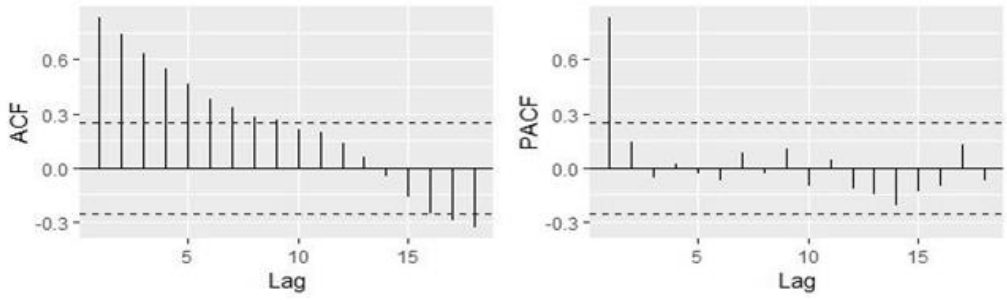

Рис. 1. Автокорреляционная и частная автокорреляционная функции для ряда

P-value теста Льюнга – Бокса составило p-value=3,49*10^{-11}, следовательно, отвергается нулевая гипотеза об отсутствии автокорреляции во временном ряду. Статистика KPSS составила 0,433 при критическом значении 0,347 (на уровне значимости 0,1), следовательно, также отвергается гипотеза о стационарности ряда. Таким образом, исходные ряд не является стационарным, и необходимо его дифференцировать.

Визуально дифференцированный ряд похож на белый шум. Функции ACF и PACF представлены на рисунке 2.

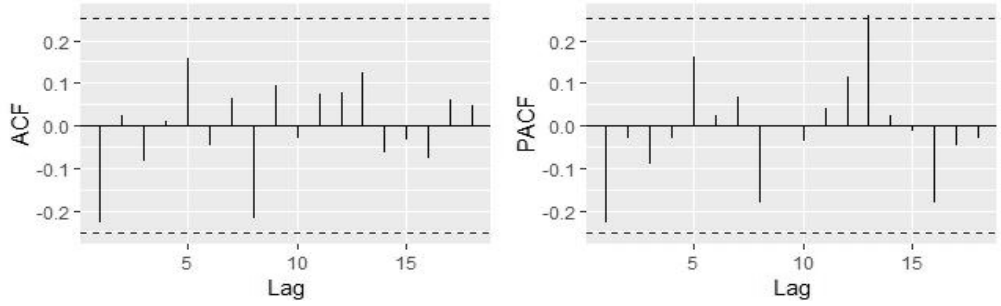

Рис. 2. Автокорреляционная и частная автокорреляционная функции для дифференцированного ряда

Расширенный тест Дики – Фуллера дифференцированного ряда (альтернативная гипотеза о стационарности ряда) показал, что на уровне значимости 0,01 нулевая гипотеза о наличии единичного корня отвергается.

P-value критерия Льюнга – Бокса составило p-value=0,07, следовательно, отвергается гипотеза о наличии автокорреляции во временном ряду.

Статистика теста KPSS меньше критического значения на уровне значимости 0,01, тогда гипотеза о стационарности не отклоняется на уровне значимости 0,01.

Проведенные тесты и анализ ACF и PACF показали, что при нестационарности исходного ряда дифференцированный ряд является стационарным, поэтому параметр d модели ARIMA(p,d,q) можно принять d=1.

Порядок скользящей средней определен как число значимых лагов автокорреляционной функции, а порядок авторегрессии – как число ненулевых лагов частной автокорреляционной функции. Поэтому порядок процесса AR(p): p=1,5,8, в то время как порядок MA(q): q=1,5,8 [3]. С разными комбинациями параметров были построены модели, затем определялся прогноз на 2019 г. и сравнивался с реальными значениями тестовой выборки. К рассматриваемым моделям была также добавлена ARIMA(0,1,1), которая была автоматически подобрана пакетом R [1].

Среди построенных моделей (таблица Д.1) на основании показателей точности прогноза (ME, RMSE, MAE, MPE, MAPE и др.) и информационных критериев Акаике и Шварца были отобраны следующие 5 моделей (модели и их ошибки представлены в таблице 1).

Таблица 1. Основные показатели точности исследуемых моделей

Показатель	**ARIMA(0,1,1)**	**ARIMA(1,1,1)**	**ARIMA(1,1,5)**	**ARIMA(5,1,1)**	**ARIMA (5,1,5)**
ME	0.825	0.825	0.824	0.807	0.984
RMSE	0.874	0.874	0.871	0.855	1.038
MAE	0.825	0.825	0.824	0.807	0.984
MPE	7.138	7.135	7.126	6.978	8.516
MAPE	7.138	7.138	7.126	6.978	8.516
MASE	2.899	2.898	2.894	2.834	3.456
AIC	40.72	42.72	47.45	48.77	46.76

BIC	44.87	48.95	61.99	63.32	69.62

У построенных моделей были проанализированы остатки – они должны соответствовать белому шуму (нулевая автокорреляция, постоянная дисперсия). Сравнение моделей проводилось с помощью теста Льюнга – Бокса, графиков автокорреляционной функции, гистограмм распределения остатков, визуального анализа остатков [4].

Соответствие остатков белому шуму проверяется с помощью теста Льюнга – Бокса (таблица 2), тест проверяет статистически значимые отличия от нуля значения первых коэффициентов автокорреляции.

Таблица 2. Значение статистики Льюнга – Бокса для отобранных моделей

Показатель	**ARIMA(0,1,1)**	**ARIMA(1,1,1)**	**ARIMA(1,1,5)**	**ARIMA(5,1,1)**	**ARIMA (5,1,5)**
p-value	0,758	0,669	0,4691	0,377	0,163

Значения p-values свидетельствуют о том, что у всех представленных моделей гипотеза о случайности остатков не отвергается.

Утверждение об отсутствии автокорреляции в остатках подтверждается также анализом коррелограмм построенных моделей.

При отсутствии корреляции в остатках может присутствовать гетероскедастичность (изменение дисперсии). Визуальный анализ остатков построенных моделей показал, что остатки гомоскедастичны и не имеют тенденции.

Кроме того, гистограммы моделей ARIMA(0,1,1), ARIMA(1,1,1) и ARIMA(5,1,5) свидетельствуют о нормальном распределении остатков, остатки симметрично распределены относительно нуля.

Модели имеют маленький разброс в значения критериев Акаике и Шварца. Среди моделей, у которых распределение остатков схоже с нормальным распределением, для прогноза была выбрана модель ARIMA (0,1,1), т.к. она имеет наименьший критерий Акаике и не отличается большим количеством параметров (как, например, ARIMA (5,1,5)). Стоит отметить, что модель представляет собой простое экспоненциальное среднее для ряда первых разностей ряда.

Таким образом, ряд описывается следующим уравнением: $Y_t=u_t - 0{,}2214u_{t-1}$. Средняя абсолютная процентная ошибка MAPE=2,43 - ошибка составила 2,43% от фактических значений (по тестовой выборке).

Ошибки модели, как уже сказано не коррелируют между собой (коэффициенты автокорреляции в пределах нормы), распределение похоже на нормальное: гистограмма демонстрирует небольшую ассимметрию относительно нуля (представлено на рисунке 3).

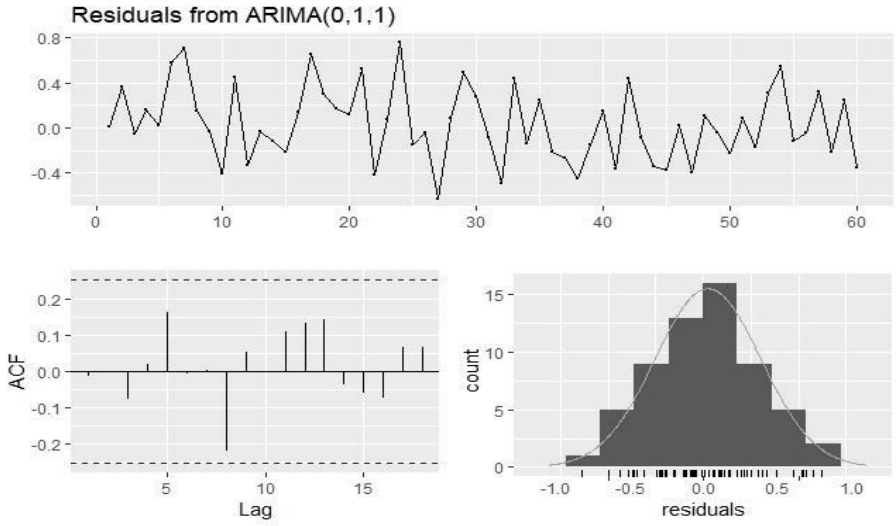

Рис. 3. Анализ остатков модели ARIMA(0,1,1)

Для сравнения распределения остатков с нормальным распределением был построен график квантиль-квантиль, который представлен на рисунке 4.

Можно видеть поквантильное сравнение фактических значений остатков с нормальным распределением: наблюдения, находящиеся на границах распределения, немного «выбиваются», но в целом можно считать распределение близким к нормальному.

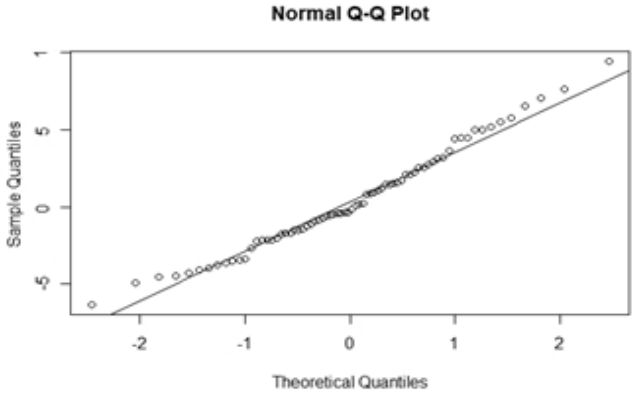

Рис. 4. График квантиль-квантиль остатков модели ARIMA(0,1,1)

Таким образом, убедившись, что модель ARIMA (0,1,1) удовлетворяет всем требованиям, можно использовать ее для прогноза. Прогноз на 2019-2020 гг. можно видеть на рисунке 5.

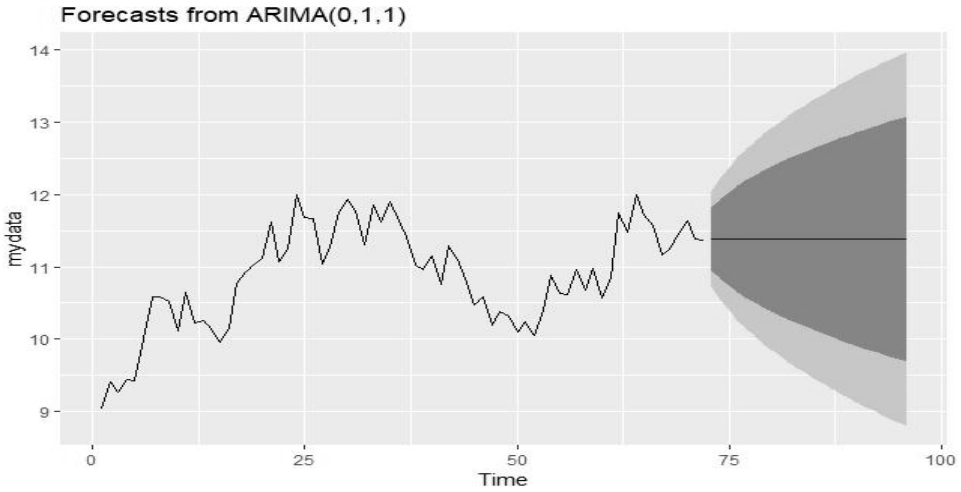

Рис. 5. Прогноз по модели ARIMA(0,1,1)

Эти прогнозные значения могут быть использованы для принятия управленческого решения, т.к. точность прогноза хорошая, и модель соответствует всем необходимым требованиям.

Стоит отметить, что помимо ручного подбора параметров, используемый язык программирования «предложил» свои параметры модели исходя из минимизации критериев Акаике и Шварца.

Список литературы / References

1. *Hyndman R.* Automatic time series forecasting: the forecast package for R // Journal of Statistical Software, 2007. № 6. 31 p.
2. *Трегуб А.* Методика построения модели ARIMA для прогнозирования динамики временных рядов // Лесной вестник. 2011. №5. С. 179-183.
3. *Елисеева И.И.* Эконометрика: учебник для магистров. М.: Юрайт, 2012. 453 с.
4. *Канторович Г.* Анализ временных рядов // Экономический журнал ВШЭ, 2002. № 3. С. 379-401.