[1] Department of Sociology, Anthropology, and Social Work, Texas Tech University. Texas, USA

[2] American Society for Emergency Contraception

Correspondence to: B Wagner
brandon.wagner@ttu.edu (or @BrandonGWagner on Twitter/X)
https://orcid.org/0000-0002-1023-4762

Cite this as: *BMJ* 2023;383:p2739
http://dx.doi.org/10.1136/bmj.p2739
Published: 20 December 2023

FAST FACTS

Using autoregressive integrated moving average models for time series analysis of observational data

This article discusses the use of autoregressive integrated moving average (ARIMA) models for time series analysis. Rather than forecasting future values, we focus here on examining change across time in outcomes of interest and how this change is related to relevant variables.

Brandon Wagner, [1] Kelly Cleland[2]

Time series data

Much of the data that we collect about the world around us—stock prices, unemployment rates, party identification—are measured repeatedly over time. By failing to account for the linked and time dependent nature of these data, common analytic techniques may misrepresent their internal structure. If we wish to describe patterns over time or forecast values beyond the observation period, we need to account for how current values may depend on previous values, trends may exist in the data, or data may vary seasonally. To visualize this, consider an electrocardiogram. The readings expected at a given moment depend not only on the preceding values but also on the position within the entire cycle. For example, following the P wave, we would expect to see the QRS complex. The assumption that each reading is unaffected by preceding values would be valid only in the most distressing circumstances (that is, during fibrillation or after death).

Description of ARIMA model

To incorporate this complex nature of time series data into models, Box and Jenkins introduced the autoregressive integrated moving average (ARIMA) model.[1] As the name implies, this model contains three different components: an autoregressive (AR) component, a differencing for stationarity (I) component, and a moving average (MA) component. The first component allows the outcome at a given moment to depend on previous values of the outcome. As this model requires a time series with properties that do not vary across time (that is, a stationary time series), the second model component (integrated) allows researchers to subtract previous observations to obtain a stationary time series, if needed. The third component (moving average) models the error term as a combination of both contemporaneous and previous error terms.

Box and Jenkins proposed an iterative process of modeling time series data that contains three steps. The first stage ("identification") involves transforming the data if needed, obtaining a stationary time series though differencing, and examining the data, autocorrelations, and partial autocorrelations to determine potential model specifications (that is, order of the autoregressive, integrated, and moving average components). The second step ("estimation") estimates the time series model with the sets of potential model parameters and then selects the best model. For example, in the linked paper (doi:10.1136/bmj-2023-077437),[2] we used the bayesian information criterion and Akaike information criterion to select the best fitting model from among candidate models. The model that best fitted the data, an ARIMA(1,1,1) model, had order one for each term (autoregressive, integrated, and moving average). This means that we model the change in sales between week t and week t-1, a first difference. The model also includes the previous week's value as a predictor of this change (autoregressive order 1) and an error term that is composed of the contemporary week's and previous week's errors (moving average order 1). Alternative specifications for the ARIMA model would correspond to the number of differences necessary to construct a stationary time series model, the number of previous values to include as predictors, or the combination of previous errors included in the error for a given observation. The third step ("diagnostic checking") examines the model for potential deficiencies and, if any are found, restarts the process. Although not without critiques,[3] this modeling approach remains popular today. Field specific texts can provide a helpful introduction to the topic for most readers. For example, we found a text by Becketti helpful in the preparation of the linked paper.[4]

The model and process described above allow researchers to explore change in an outcome over time. But what if you think some other variable is affecting your outcome of interest? In many modern computer packages, estimated from the ARIMA model described above can be adjusted with a set of exogenous X variables that also vary across time. The resulting model is often referred to as a regression with ARIMA errors, as the estimated regression includes an error term that is an ARIMA process.

When and why to use ARIMA model

ARIMA models have previously been used to explore time dependent processes in population health. For example, recent work has used ARIMA models to explore disease diagnosis or outcomes and demand for medical services.[5-8] ARIMA models or, more generally, regressions with ARIMA errors are commonly used for time series data for a few key reasons. Firstly, the model allows us to incorporate relations between observations. For example, the spread of an infectious disease through a population

likely depends on previous counts of infection in the population. Consequently, hundreds, if not thousands, of papers applied ARIMA models to counts of infection or death from the covid-19 pandemic, tracing the spread of the disease over time in settings around the world. Enabling data to incorporate dependencies in terms of lags or seasonality allows researchers to better fit such data. The second key benefit of estimating regressions with ARIMA errors is that it allows us to explore changes relative to the underlying background trends in the data. In our case, sales of levonorgestrel emergency contraception have been increasing over time in the United States.[9] A basic model exploring weekly sales as a function of the dichotomous holiday indicators might not correctly differentiate the sales increase following the New Year from such a background increase.

Limitations

ARIMA modeling remains popular today, although researchers must recognize some limitations. Firstly, these models may require relatively long time series, with a common rule of thumb being at least 50, or preferably 100, observations to estimate seasonal components. Although this is not a challenge to frequently measured values or long running time series, it may limit the acceptability of ARIMA models in some cases. Secondly, the described model estimation process fits a model form, specifically the order of autoregressive and moving average terms, to the observed data. Although useful in describing the observed time trend, fitting the ARIMA model this way may limit its utility in describing trends in other contexts. Finally, as with all models, ARIMA models should be examined as one possible model. In some cases, alternative models may better fit observed data,[10] so examination of the data and model specifications is essential before selecting a modeling approach.

Conclusion

Regressions with ARIMA errors can be useful tools to understand time series data. By incorporating linkages between observations, and exploring change across time, these models can both describe trends and explore how these trends vary with predictors of interest.

Funding and competing interests available in the linked paper on bmj.com.

Provenance and peer review: Commissioned; not externally peer reviewed.

We thank Circana Inc for allowing us to use its data for this project. All estimates and analyses in this paper based on Circana's data are by the authors and not by Circana Inc.

1 Box GEP, Jenkins GM. *Time Series Analysis: Forecasting and Control.* Holden-Day, 1970.

2 Wagner B, Cleland K. Retail demand for emergency contraception in United States following New Year holiday: time series study. *BMJ* 2023;383:e077437.

3 Hendry DF. *Dynamic econometrics.* Oxford University Press, 1995doi: 10.1093/0198283164.001.0001.

4 Becketti S. *Introduction to Time Series Using Stata.* Stata Press, 2020.

5 Xu B, Li J, Wang M. Epidemiological and time series analysis on the incidence and death of AIDS and HIV in China. *BMC Public Health* 2020;20:. doi: 10.1186/s12889-020-09977-8 pmid: 33317484

6 Earnest A, Evans SM, Sampurno F, Millar J. Forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using autoregressive integrated moving average (ARIMA) models. *BMJ Open* 2019;9:e031331. doi: 10.1136/bmjopen-2019-031331 pmid: 31431447

7 Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Brief* 2020;29:105340. doi: 10.1016/j.dib.2020.105340 pmid: 32181302

8 Eyles E, Redaniel MT, Jones T, Prat M, Keen T. Can we accurately forecast non-elective bed occupancy and admissions in the NHS? A time-series MSARIMA analysis of longitudinal data from an NHS Trust. *BMJ Open* 2022;12:e056523. doi: 10.1136/bmjopen-2021-056523 pmid: 35443953

9 Wagner B, Cleland K. Increases in Retail Sales of Levonorgestrel Emergency Contraception in the United States, 2016-2022.*SocArXiv* 2023. doi: 10.31235/osf.io/xs6zmosf.io/xs6zm

10 Nanda S. Forecasting: Does the Box-Jenkins Method Work Better than Regression? *Vikalpa J Decis Mak* 1988;13:-62.pmid: 35443953