

Autoregression as a means of assessing the strength of seasonality in a time series

Rahim Moineddin ¹

Ross E.G. Upshur ^{1,2,3,*}

Eric Crighton ²

Muhammad Mamdani ⁴

Address: ¹ Department of Family and Community Medicine, University of Toronto, 256 McCaul Street, 2nd Floor, Toronto, ON, Canada M5T 2W5

² Primary Care Research Unit, Sunnybrook and Women's College Health Sciences Centre, 2075 Bayview Avenue, #E-349, Toronto, ON, Canada M4N 3M5

³ Department of Public Health Sciences, University of Toronto, McMurrich Building, 12 Queen's Park Crescent W., Toronto, ON Canada, M5S 1A8

⁴ Institute for Clinical Evaluative Sciences, 2075 Bayview Avenue, Toronto, ON Canada M4N 3M5

* Corresponding Author

Email: Rahim Moineddin rahim.moineddin@utoronto.ca
Ross Upshur ross.upshur@sw.ca
Eric Crighton Eric.crighton@sw.ca
Muhammad Mamdani Muhammad.mamdani@ices.on.ca

Abstract

Background: The study of the seasonal variation of disease is receiving increasing attention from health researchers. Available statistical tests for seasonality typically indicate the presence or absence of statistically significant seasonality but do not provide a meaningful measure of its strength.

Methods: We propose the coefficient of determination of the autoregressive regression model fitted to the data ($R^2_{Autoreg}$) as a measure for quantifying the strength of the seasonality. The performance of the proposed statistic is assessed through a simulation study and using two data sets known to demonstrate statistically significant seasonality: atrial fibrillation and asthma hospitalizations in Ontario, Canada .

Results: The simulation results showed the power of the $R^2_{Autoreg}$ in adequately quantifying the strength of the seasonality of the simulated observations for all models. In the atrial fibrillation and asthma datasets, while the statistical tests such as BKS and Fisher's Kappa support statistical evidence of seasonality for both ($p < 0.01$), the $R^2_{Autoreg}$ quantifies the strength of that seasonality.

Corroborating the visual evidence that asthma is more conspicuously seasonal than atrial fibrillation, the calculated $R^2_{Autoreg}$ for atrial fibrillation is 0.355, indicating a weak to moderate seasonality, whereas for asthma, the calculated $R^2_{Autoreg}$ is 0.853, indicating a strong seasonality.

Conclusions: For the purposes of health services research, evidence of the statistical presence of seasonality is insufficient to determine the etiologic, clinical and policy relevance of findings.

Measurement of the strength of the seasonal effect, as can be determined using the $R^2_{Autoreg}$ technique, is also important in order to provide a robust sense of seasonality.

Key Words: Seasonality, autoregression, time series, asthma, atrial fibrillation

Background

Seasonality is an important component of disease manifestation. The presence of predictable seasonality is a clue to the possible etiology of disease, be it from microbial, environmental or social factors. Understanding seasonality is also essential for setting rational policy, particularly with respect to the planning for seasonal demands for health services.

For studying seasonality, several statistical methods are available ranging from simple graphical techniques to more advanced statistical methods. Additionally, autocorrelation functions can be examined to assess regularity of periodicity or seasonality. Several statistical tests have been introduced for studying the cyclical variation of time series data. For example, Edwards [1] developed a statistical test, which locates weights corresponding to the number of observed cases for each month at 12 equally spaced points on a circle. The test is said to be significant if the calculated centre of the mass significantly deviates from the circle's centre. Jones et al [2] developed a test for determining whether incidence data for two or more groups have the same seasonal pattern. Further, Marrero [3] compared the performance of several tests for seasonality by simulation, which can be used as a guideline for selecting appropriate tests for a given data set based on the size of the data set and the shape of the sinusoidal curve. To apply any of these tests, however, observations must be aggregated into 12 monthly data points.

Several alternative tests, which do not require aggregated data, have also been developed. These include Fisher's Kappa (FK), which tests whether the largest periodogram is statistically different from the mean of periodograms; Bartlett's Kolmogorov-Smirnov (BKS) test, which statistically compares the normalized cumulative periodogram with the cumulative distribution function of a uniform zero and one random variable; and the X-11 procedure as used by the census bureaus in the United States and Canada [2, 4-11]. These tests utilize the frequency and

time domain to detect seasonality. Each test provides an indication of the presence or absence of statistical significance of seasonality, however, they do not provide a sense of the magnitude of seasonality or how much variance is explained by seasonal occurrence in the data. This is particularly important in health care, as the presence of statistically significant seasonality may not translate into either etiologic or policy relevance.

In an effort to address the shortcomings in existing statistical methods, we propose the application of autoregressive regression models as a means for assessing the degree of accuracy to which a new observation can be predicted by stable (seasonal factors are constant over time) seasonal variation and use it for quantifying the strength of the seasonality within a set of serially correlated observations. In classical regression analysis the coefficient of determination, R^2 , is a standard statistical tool for estimating the proportion of total variation of the dependent variable, which can be explained by explanatory variables. A crucial point in standard regression is that observations are independent of one another. However, time series observations can be serially autocorrelated and this correlation must be taken into account.

Autoregressive regression models are a natural generalization of standard regression models for analyzing correlated data. For monthly data, one can use dummy variables for months in the autoregression model as a single predictor, and then, after correcting for the autocorrelation, calculate the coefficient of determination, $R^2_{Autoreg}$. When the time series is stationary and the trend is eliminated, the statistical significance of the dummy variables (months) indicates seasonality. The relationship between the stable seasonal factors and the estimates of the regression equation parameters are as follows: suppose there are k years monthly, $n = 12k$, trend removed and centred (mean deleted) observations. Let $\bar{y}_i, i = 1, 2, \dots, 12$ denote the monthly average. The monthly averages, \bar{y}_i 's, can be interpreted as

crude estimates of stable seasonal factors, therefore, the range of parameter estimates is a good estimate of the magnitude of seasonal variation. For estimating $R^2_{Autoreg}$ one defines 11 dummy variables $m_i = 1$ if month equals i , $m_i = 0$ otherwise and then regress m_i 's on y_t 's. It is not difficult to show that the ordinary least squared estimates of the parameters of the regression equation $y_t = \beta_0 + \sum_{i=1}^{11} \beta_i m_i + \varepsilon_t$ are $\hat{\beta}_i = \bar{y}_i, i = 1, \dots, 11$ and $\hat{\beta}_0 = \bar{y}_{12}$. In practice $R^2_{Autoreg}$ and parameters β_i 's are estimated simultaneously, therefore, the estimated parameters $\hat{\beta}_i$'s can be used as seasonal factor estimates.

The coefficient of determination, which lies between 0 and 1, can be used as a measure for the strength of the stable seasonality because it measures how well the next value can be predicted using month as the only predictor. When $R^2_{Autoreg}$ is zero there is no seasonality. When $R^2_{Autoreg}$ is equal to 1, observations can be perfectly predicted for each month, which means that the variable month explains 100% of the variation in the data. In other words, there is a perfect seasonality. In practice we may characterize the strength of the seasonality based on different ranges of values for $R^2_{Autoreg}$. Similar to other measures of correlation or goodness of fit we can interpret $R^2_{Autoreg}$ as follows: Values ranging from 0 to less than 0.4 may be characterized as non-existent to weak seasonality, 0.4 to less than 0.7 represent moderate to strong seasonality, and values ranging from 0.7 to 1 represent strong to perfect seasonality. The coefficient of determination, $R^2_{Autoreg}$, does not quantify the magnitude of the seasonal effect (the difference between peak and trough, which can be estimated by the difference between maximum and minimum parameter estimates of the regression equation) but rather it quantifies its strength (i.e., how well new observations can be predicted when month is the only predictor).

The purpose of this paper is to evaluate the utility of R-squared autoregression in explaining variance in assessing stable seasonality. To this end, we have examined the performance of the R-squared autoregression through a simulation study and using two data sets known to demonstrate statistically significant weak and strong seasonality: monthly hospitalizations for atrial fibrillation and asthma.

Methods

Statistical methods

The autoregressive linear regression model for monthly observations is defined as

$$\begin{aligned} Y_t &= X_t \beta + \varepsilon_t \\ \varepsilon_t &= -\phi_1 \varepsilon_{t-1} - \phi_2 \varepsilon_{t-2} - \dots - \phi_p \varepsilon_{t-p} + e_t \\ e_t &\sim N(0, \sigma_e^2) \end{aligned}$$

where Y_t is the observed time series, X_t is design matrix (a $k \times 12$ matrix of 0 and 1),

$\beta = (\mu, \beta_1, \dots, \beta_{11})'$ is vector of parameters, ε_t error term follows an autoregressive model of

order p and e_t is normally and independently distributed with mean zero and variance

σ_e^2 [Hamilton, 1994, Chapter 4 #12]. Standard statistical packages (e.g. SAS) can be used for

estimating the parameters and the coefficient of determination, $R_{Autoreg}^2$, after adjusting for

correlated error terms.

Simulation

In order to assess the performance of the proposed $R_{Autoreg}^2$ for measuring the strength of the seasonality of a time series, a simulation study was conducted. Following this, the proposed technique was applied to two real data sets. The SAS software, version 8.2 (SAS Institute Inc. Cary, North Carolina) is used for simulating monthly observations and calculating $R_{Autoreg}^2$.

We simulated 1000 replications of monthly observations over 10 years from the following model:

$$X_t = \alpha \cos\left(\frac{\pi t}{6}\right) + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_{12} X_{t-12} + e_t$$

and calculated the $R_{Autoreg}^2$ for each replication. By changing α, ϕ_1, ϕ_2 , and ϕ_{12} the coefficients, this model generates observations with a cyclical trend component of period 12, observations from a seasonal ARMA model with seasonal period of 12, and a combination of both. By changing the coefficients of the model we can generate pure white noise to highly correlated data with seasonal patterns. For example, a model with $\phi_1 = \phi_2 = \phi_{12} = 0$ generates a series of observations, which is a mixture of white noise and cyclical trend. The size of α controls the contribution of the seasonal trend in the generated observations. Parameters ϕ_1 and ϕ_2 control the correlation structure of the simulated data. When $\phi_1 = 0.9$, for example, highly correlated observations are generated. When parameter ϕ_{12} is nonzero, the model generates observations with a stochastic seasonal component. Similarly, nonzero ϕ_{12} combined with nonzero ϕ_1 or ϕ_2 generates a series with a stochastic seasonal component which is correlated with the non-seasonal components. When all parameters are nonzero the generated observations have a complex structure, which depends on a cyclical trend, a stochastic seasonal component, and its correlation structure. In our simulation, we set the order of the error terms in the autoregression model to 2 (for all simulated observations). The mean and standard deviation of 1000 calculated $R_{Autoreg}^2$ procedures are given in Table 1.

When α and ϕ_{12} are zero $R_{Autoreg}^2$ indicates no significant seasonality even for highly correlated data (e.g. $\phi_1 = -0.9$). The calculated $R_{Autoreg}^2$ increases as α increases regardless of the

sign and magnitude of the other parameters. When α is zero (there is no cyclical trend), the $R^2_{Autoreg}$ increases in all cases as ϕ_{12} increases. This demonstrates the ability and usefulness of $R^2_{Autoreg}$ in quantifying the strength of the seasonality in time series data.

In order to apply the autoregression procedure to actual data it is important to eliminate the nonstationarity in the mean and variance of the observations. For the data used in the study, a logarithmic transformation was used to stabilize the variance and then the trend was estimated and subtracted from the observations. The SAS procedure AUTOREG was used for calculating $R^2_{Autoreg}$. In all cases p , the order of autoregressive model for error terms is set equal to 2.

Data Sources

The data are derived from two retrospective population-based cross-sectional time series studies assessing temporal patterns in all discharge separations for asthma and atrial fibrillation for the population of Ontario from April 1, 1988 to March 31, 2000. Approximately 14 million residents of Ontario, Canada eligible for universal health care coverage during this time were included for analysis. The database used was the Canadian Institute for Health Information Discharge Abstract Database, which records discharges from all Ontario acute care hospitals. All records with a most responsible discharge diagnosis of atrial fibrillation (ICD-9 code: 427.3) and asthma (ICD-9 code: 493) were selected. The numerator consisted of the total number of discharge separations for each month. Denominators were constructed from annual census data provided by Statistics Canada for each age group for residents of Ontario. Monthly population estimates were created through linear interpolation.

Results

Atrial fibrillation

Overall, there were 96,506 (51,033 female and 45,473 male) discharge separations for all ages. Figure 1 shows the monthly rates of admission per 100,000 population. There is a conspicuous upward trend in admissions over the first four years (Figure 1). Visual inspection does not support conspicuous seasonality. However, after applying log transformation and eliminating the trend, the data doesn't show evident statistically significant seasonality (Bartlett's Kolmogorov-Smirnov Statistic (BKS) for both genders, female and male are 0.136, 0.155, 0.109 with a p-value 0.1424, 0.0650, 0.3695 respectively. The Fisher's Kappa (FK) test statistics are 16.83, 13.16, and 8.14, which are statistically significant with a p-value less than 0.01, 0.01, and 0.05 respectively). The calculated $R^2_{Autoreg}$ for both genders, female, and male are 0.40, 0.30, and 0.32. The difference between maximum and minimum months parameter estimates are 0.1357, 0.1184, and 0.1757 on log scale for both genders, female and male respectively. The small values for amplitude of seasonal factors, which correspond to 1-2 patients per 100,000, and low values of R-squared indicate a weak to non-existent seasonality.

Asthma

In total, there were 206,561 (104,283 female and 102,278 male) asthma hospital discharges for all ages. Figure 2 shows the monthly rates of asthma per 100,000 population. The visual inspection of Figure 2 shows a clear autumn peak and summer trough seasonal pattern occurring every year over the 12 year period. The results of the seasonality tests applied on the log-transformed rates demonstrate statistically significant seasonality. (Fisher's Kappa (FK) for both genders, female and male are 30.18, 27.00, and 30.35 with p-values less than 0.01; Bartlett Kolmogorov-Smirnov (BKS) are 0.432, 0.446, and 0.415 with p-values less than 0.0001). The

calculated $R^2_{Autoreg}$ are 0.92, 0.88, and 0.92 for both genders, female and male respectively. The difference between maximum and minimum months parameter estimates are 0.7388, 0.6968, and 0.8537 on log scale for both genders, female and male respectively. The large values for amplitude of seasonal factors (on log scale), which correspond to 7-8 patients per 100,1000, and high values of R-squared provide clear evidence of strong seasonality.

Discussion

The results of this study show that the $R^2_{Autoreg}$ can be useful in distinguishing weak from strong stable seasonal effects in both simulation and in actual data sets. While the statistical tests such as BKS and Fisher's Kappa support statistical evidence of seasonality in the data, the $R^2_{Autoreg}$ allows quantification of the strength of that stable seasonality as demonstrated by the simulation results. Regardless of the values of the parameters ϕ_1 and ϕ_2 , when the parameters α and ϕ_{12} were zero the $R^2_{Autoreg}$ was small and when one or both of those parameters increased, $R^2_{Autoreg}$ increased proportionally. This is important because any proposed statistics for measuring the strength of the seasonality must be invariant of the correlation structure. The simulation results showed the power of the $R^2_{Autoreg}$ in adequately quantifying the strength of the seasonality of the simulated observations for all models. The magnitude of $R^2_{Autoreg}$ shows how well the next value can be predicted by using month as the only predictor. In other words it shows the contribution of seasonality in the total variation of the data.

When the technique was applied to the two data sets, it corroborated the visual evidence that asthma is more conspicuously seasonal than atrial fibrillation. The seasonality of asthma has been conclusively demonstrated in several studies and is likely a key to understanding the

etiology of exacerbations of asthma [13-15]. The strength and consistency of the effect is likely of relevance to health policy and planning. The seasonality of atrial fibrillation has been reported, but outcomes were reported as relative risks [16]. The analysis provided here indicates that the seasonality of admissions for atrial fibrillation is not likely of policy or clinical significance as the magnitude is quite small.

One important question that remains to be answered is how the magnitude of seasonal factor changes over time affect the $R^2_{Autoreg}$. For non-stable seasonal variation a proper transformation such as log may be required to transfer the non-stable seasonal variation to the stable one. The value of $R^2_{Autoreg}$ may change if sampling period changes e.g. monthly data converted to weekly. By including year and month as predictors we can adjust for moving seasonality, however, further research is required.

Conclusions

The proposed autoregression method is a statistical technique well suited to the study of seasonality in health data. Although monthly data was used for this analysis, it can easily be applied to weekly or seasonal data. The approach allows researchers to quantify and compare the strength of the seasonality for different genders and age groups. The coefficient of determination is easy to calculate and interpret. And finally, it is well known to health care researchers and is frequently used as a measure for goodness of fit. For the purposes of health services research and population health measurement, evidence of the statistical presence of stable seasonality is insufficient to determine the etiologic, clinical and policy relevance of findings. Measurement of the strength of the seasonal effect is also required in order to provide a robust sense of seasonality. We believe that this autoregression technique, in concert with statistical testing,

graphical representation and measures of the absolute magnitude of seasonal effect, is an important component to this robust approach.

Competing Interests: None declared

Authors Contributions: RM and RU initiated the idea for the study. MM and EC contributed critical intellectual input to the project. RM wrote the initial draft, and all authors contributed to the subsequent drafts. All have read and approve of the final draft.

Acknowledgements: We would like to thank Shari Gruman for her expert assistance in formatting the manuscript.

References

1. Edwards J: **The recognition and estimation of cyclic trends.** *Ann Hum Genet* 1961, **25**:83-86.
2. Jones RH, Ford PM, Hamman RF: **Seasonality comparisons among groups using incidence data.** *Biometrics* 1988, **44**:1131-1144.
3. Marrero O: **The performance of several statistical tests for seasonality in monthly data.** *J Statist Comput Simulation* 1983, **17**:275-296.
4. Fuller W: *Introduction to Statistical Time Series.* New York: John Wiley; 1976.
5. Franses P: **Testing for seasonality.** *Econ Lett* 1992, **38**:259-262.
6. Marrero O: **L'analyse de la variation saisonniere quand l'amplitude et la taille sont faibles.** *Canad J Statist* 1999, **27**:857-882.
7. Marrero O: **Statistical testing for seasonality in data with multiple peaks and troughs.** *Biom J* 1984, **26**:591-608.
8. Marrero O: **The power of a nonparametric test for seasonality.** *Biom J* 1988, **30**:495-502.
9. Normolle D, Brown M: **Identification of aperiodic seasonality in non-Gaussian time series.** *Biometrics* 1994, **50**:798-812.
10. Fellman J, Eriksson AW: **Statistical analysis of the seasonal variation in demographic data.** *Hum Biol* 2000, **72**:851-876.
11. Harvey H: **Testing in unobserved components models.** *J Forecasting* 2001, **20**:1-19.
12. Hamilton J: *Time Series Analysis.* Princeton, NJ: Princeton University Press; 1994, Chapter 4.
13. Crighton EJ, Mamdani MM, Upshur RE: **A population based time series analysis of asthma hospitalisations in Ontario, Canada: 1988 to 2000.** *BMC Health Serv Res* 2001, **1**:7.
14. Kimbell-Dunn M, Pearce N, Beasley R: **Seasonal variation in asthma hospitalizations and death rates in New Zealand.** *Respirology* 2000, **5**:241-246.
15. Harju T, Keistinen T, Tuuponen T, Kivela SL: **Seasonal variation in childhood asthma hospitalisations in Finland, 1972-1992.** *Eur J Pediatr* 1997, **156**:436-439.

16. Frost L, Johnsen SP, Pedersen L, Husted S, Engholm G, Sorensen HT, Rothman KJ:
Seasonal variation in hospital discharge diagnosis of atrial fibrillation: a population-based study. *Epidemiology* 2002, **13**:211-215.

Figure Legends

Figure 1. Atrial fibrillation all ages

Figure 2. Asthma data all ages

Table 1: Simulation results

Parameters			$R^2_{Autoreg}$ (Std)			
ϕ_1	ϕ_2	ϕ_{12}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$	$\alpha = 4$
0	0	0	0.099 (0.041)	0.400 (0.070)	0.703 (0.051)	0.902 (0.020)
0.5	0	0	0.098 (0.040)	0.409 (0.070)	0.712 (0.049)	0.906 (0.019)
-0.9	0	0	0.103 (0.042)	0.396 (0.069)	0.698 (0.050)	0.899 (0.020)
0.5	-0.8	0	0.098 (0.039)	0.393 (0.065)	0.699 (0.044)	0.899 (0.017)
0	0	0.5	0.256 (0.088)	0.703 (0.064)	0.896 (0.026)	0.971 (0.007)
0	0	0.7	0.410 (0.114)	0.847 (0.043)	0.953 (0.014)	0.988 (0.004)
0	0	0.9	0.716 (0.101)	0.974 (0.009)	0.993 (0.002)	0.998 (0.001)
0.3	-0.2	0.3	0.171 (0.066)	0.602 (0.072)	0.847 (0.034)	0.957 (0.011)
0.3	-0.2	0.5	0.260 (0.095)	0.764 (0.060)	0.924 (0.021)	0.980 (0.006)
0.3	-0.2	0.7	0.470 (0.149)	0.934 (0.024)	0.982 (0.006)	0.995 (0.002)

Additional files provided with this submission:

Additional file 1: Moineddin Upshur Crighton Mamdani Figure 1.doc : 101KB
<http://www.pophealthmetrics.com/imedia/1499740815216748/sup1.doc>

Additional file 2: Moineddin Upshur Crighton Mamdani Figure 2.doc : 192KB
<http://www.pophealthmetrics.com/imedia/1423207212216748/sup2.doc>