Times Series Analysis of Monthly New York State Alcoholic Beverage Tax

——Huimeng Zhang, Jinjin He, Canyang Jin

Abstract:

We begin project by analysis of alcoholic and beverage tax. Apply decomposition, kernel smoother and simple linear regression to extract trend and seasonal patterns. Discovering the alcoholic and beverage tax associated with the alcoholic and beverage tax at time t-6 and time t-12 from lagged scatterplot matrix, we fit a regression model with lagged variables and get desirable results. We also apply Holt-Winters model, SARIMA model to predict future alcoholic beverage tax and compare these models.

Introduction:

Alcohol-related problems have a series of impact to a country's public health. 88,000 deaths each year in the USA are directly attributable to alcohol consumption. Additionally, some studies show that deaths related to alcohol are up 35 to 50 since 2000. Since alcohol consumption is a contributing factor for the transport accidents, violent crime suicide and a number of other causes of injury and death, alcoholic beverages have been taxed at a relatively high rate. We attempt to figure out the monthly New York State alcoholic beverage tax by time series analyses. There are two main goals of our project:

(1) Identify trend and seasonal patterns of our response, the monthly New York State alcoholic beverage tax.

(2) Analyze relationships between predictors (monthly cigarette tax, personal income tax and unemployment rate) and the response (monthly alcoholic beverage tax), and make prediction on the monthly alcoholic beverage tax based on the gathered information.

Methods and Materials

We dataset is from https://data.world/stateofny/taxesfees-collected-monthly https://labor.ny.gov/stats/laus.asp. The Department of Taxation and Finance monthly produces a compilation of those state and local and local purpose taxes and fees collected by the Department. The taxes and fees information provided in this data set are primarily taxes imposed by the Tax Law, but also includes fees that are imposed by other state laws but are administered and collected by the Department. Collections are net of refunds and other processing and accounting adjustments. The data set provides a history of these collections by month beginning with April 1996. (Taxes/Fees Collected Monthly, 2016) Our dataset includes monthly cigarette tax, personal income tax, unemployment rate and alcoholic beverage tax, containing 245 observations collected by month from April 1996

to August 2016. Our response is monthly alcoholic beverage tax and we have 3 predictors cigarette tax, personal income tax, and unemployment rate. Their quantitative and graphical summary is shown in Table 1 and Fig 1, respectively. All of these time series are non-stationary.

Firstly, we want to identify trend and seasonal patterns of our response, monthly alcoholic beverage tax. Secondly, we intend to find out appropriate time series models to forecast the monthly alcoholic beverage tax. For forecasting process, we consider two forecasting scenarios: (1) drop last 24 observations as test set and the then forecast these values and verify our forecasts; (2) forecast a few steps into the future using all of the data.

(1) Trend and seasonal pattern analysis

Fig 2 shows the monthly New York State alcoholic beverage tax from April 1964 to August 2016. At first, we need to identify the primary pattern in the time history. It is difficult to decide whether a time series is stationary or not, so we conduct KPSS test to assess stationarity since we think it is a stationary time series as shown in Fig 2.

$$H_0: Stationary$$

$$H_1: There\ is\ a\ unit\ root.$$

However, the P-value of KPSS test for checking the stationary of the monthly New York State alcoholic beverage tax is 0.01, which indicates that we can reject the null hypothesis and conclude that it is non-stationary.

The ACF plot (Fig 3) shows a strong dependence structure at different lags and also clear indication of seasonal structure in the data. Compared to ACF plot, PACF plot (Fig 4) shows less significant autocorrelation, and there is little indication of seasonality.

Even after accounting for the intervening variables, the PACF shows that there are several significant lags of autocorrelation.

We apply simple linear regression to estimate the trend by fitting the model

$$x_t = \beta_0 + \beta_1 t + w_t \ ,$$

and obtained the estimated coefficient $\widehat{\beta_0} = -6901766.33$, and $\widehat{\beta_1} = 352.59$ (with $p - value = 2.74 * 10^{-13}$), which indicates that there exists a statistically significant increasing trend.

We set different bandwidth values for kernel smoother which uses a weight function or kernel to average the observations. With a wider bandwidth, it shows an increasing trend; with narrower bandwidth, it shows the seasonal pattern as shown in Fig 5. Moreover, after we decompose the response, Fig 6 shows that the trend accounts for relatively large variability of the response, compared to the seasonality and the noise. Although it is successfully detrended, it still has a seasonal pattern after deseasonalizing (Fig 7). We tried different smoothing techniques in order to figure out the systematic pattern of our response, the monthly alcoholic beverage tax.

(2) Model fitting and forecasting

Among all these smoothing techniques, the Holt-Winters method can not only show the pattern of a times series, but also forecast short-term alcoholic beverage tax. We adopt Holt-Winters and SARIMA on the training set and forecast monthly alcoholic beverage tax in two years without other predictors.

Moreover, we involve predictors to our models. The lagged scatterplot matrix is shown in Fig 8. We can see that there is a relative strong linear relationship between A_t and A_{t-12}. These results agree with peaks in ACF and PACF plots. A_t slightly increases with the value of A_{t-6} and then the slope tends to be larger when A_{t-6} exceeds 17000,

indicating a relative strong curvilinear relationship between A_t and A_{t-6}. Thus, we add a dummy variable in the model to account for the change of slope. The estimated function is:

$$A_t = 2.974 \times 10^3 + 7.401 \times 10^{-2} A_{t-6} - 4.150 \times 10^3 D_{t-6} + 7.448 \times 10^{-1} A_{t-12} + 2.394 \times 10^{-1} D_{t-6} A_{t-6} + w_t$$

where

$$D_{t-6} = \begin{cases} 0 & if\ A_{t-6} < 17,000 \\ 1 & if\ A_{t-6} \geq 17,000 \end{cases}$$

A scatterplot matrix, shown in Fig 9, indicates that a possible positive linear relationship between the response and unemployment rate, a possible positive curvilinear relationship between the response and cigarette tax, and a possible positive curvilinear relationship between the response and personal income tax. Firstly, we could form 8 models to deal with the curvilinear shape of response and predictors. A_t denotes alcoholic and beverage tax, U_t denotes unemployment rate, C_t denotes cigarette tax, P_t denotes personal income tax, and w_t denotes a random error term. They are

$$A_t = \beta_0 + \beta_1 t + w_t \quad (1)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + w_t \quad (2)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + \beta_3 (C_t - \overline{C_t}) + w_t \quad (3)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + \beta_3 (C_t - \overline{C_t}) + \beta_4 (P_t - \overline{P_t}) + w_t \quad (4)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + \beta_3 (C_t - \overline{C_t}) + \beta_4 (P_t - \overline{P_t}) + \beta_5 (C_t - \overline{C_t})^2 + w_t \quad (5)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + \beta_3 (C_t - \overline{C_t}) + \beta_4 (P_t - \overline{P_t}) + \beta_5 (C_t - \overline{C_t})^2 + \beta_6 (P_t - \overline{P_t})^2 + w_t \quad (6)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + \beta_3 (C_t - \overline{C_t}) + \beta_4 (P_t - \overline{P_t}) + \beta_5 (C_t - \overline{C_t})^2 + \beta_6 (P_t - \overline{P_t})^2 + \beta_7 (C_t - \overline{C_t})^3 + w_t \quad (7)$$
$$A_t = \beta_0 + \beta_1 t + \beta_2 U_t + \beta_3 (C_t - \overline{C_t}) + \beta_4 (P_t - \overline{P_t}) + \beta_5 (C_t - \overline{C_t})^2 + \beta_6 (P_t - \overline{P_t})^2 + \beta_7 (C_t - \overline{C_t})^3 + \beta_8 (P_t - \overline{P_t})^3 + w_t \quad (8)$$

where we adjust cigarette tax and personal income tax to avoid multicollinearity problems.

From Table 2, we can see that the p-values for $(C_t - \overline{C_t})^2$ and $(P_t - \overline{P_t})^2$ are not statistically significant. Thus, we conduct F test.

Its null hypothesis and its alternative:

$$H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$

$$H_1: o.w.$$

The test statistic:

$$\mathrm{F}^* = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q-1)} = \frac{(2.383 \times 10^9 - 2.256 \times 10^9)/6}{2.236 \times 10^9/(221-10-1)} = 1.99$$

which do not exceed $\mathrm{F}_{6,211} = 2.141736$.

Decision rules: If $\mathrm{F}^* \leq \mathrm{F}_{6,211}$, we conclude H_0. Otherwise, we conclude H_1. Since $\mathrm{F}^* = 2.31 \leq \mathrm{F}_{6,211} = 2.141736$, we conclude H_0, which indicates that given that we can drop these predictors.

Additionally, we also consider harmonic functions:

$$A_t = \gamma_0 + \gamma_1 t + \gamma_2 sin(t) + \gamma_3 cos(t) + \gamma_4 U_t + \gamma_5 (C_t - \overline{C_t}) + \gamma_6 (P_t - \overline{P_t}) \quad (10)$$

We apply forward stepwise selection to choose a subset of attributes related to the response and find the appropriate subset based on adjusted $R^2$and C_p criterion, as shown in the Fig 10. According to the plot of number of predictors vs adjusted $R^2$and C_p values, we choose five of them for our linear regression model. The estimated function is:

$A_t = 1.376 \times 10^4 - 9.501 \times 10^2 sin(t) - 1.582 \times 10^3 cos(t) + 5.049 \times 10^4 U_t +$

$2.985 \times 10^{-2}(C_t - \overline{C_t}) + 3.407 \times 10^4 (P_t - \overline{P_t})$

Result

(1) Models without predictors

The Holt-Winters method uses exponentially weight moving averages to update estimates of the seasonally adjusted mean(level), slope and seasonality. There are two types of Holt-Winters methods, additive and multiplicative.

The Holt-Winters methods with additive seasonal is:

$$\begin{cases} a_t = \alpha(z_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \\ b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \\ s_t = \gamma(z_t - a_t) + (1-\gamma)s_{t-p} \end{cases}$$

The Holt-Winters methods with multiplicative seasonal is:

$$\begin{cases} a_t = \alpha\left(\frac{z_t}{s_{t-p}}\right) + (1-\alpha)(a_{t-1} + b_{t-1}) \\ b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \\ s_t = \gamma\left(\frac{z_t}{a_t}\right) + (1-\gamma)s_{t-p} \end{cases}$$

where $a_t, b_t \; and \; s_t$ are the estimated level, slope and seasonal effect at time t, and $\alpha, \beta \; and \; \gamma$ are smoothing parameters.

Comparing SSE values of them, the Holt-Winters method with multiplicative seasonal has a relative smaller value. And we use it for our forecasting process, as shown in Fig11. For evaluation process, we first apply our fitted multiplicative Holt-Winters model on the test set, and the corresponding result shows that among 24 observations, there are 7 observations out of the prediction interval, with 95% confidence. After that, we use the whole data set to make prediction on future monthly New York State alcoholic beverage tax (Fig 12).

From previous analyses, we have found that the monthly New York State alcoholic beverage tax is a non-stationary time series with increase trend and a seasonal pattern. So, we use a SARIMA model to fit the data. Based on the AIC criterion, we

choose the ARIMA(2,1,1)$(1,1,0)_{12}$ as our best model. The specific result is shown in Fig 13.

In order to evaluate its performance, we first use our fitted model for our test set. From Fig 14, among 24 observations, there only two observations falling out of the prediction interval, with 95% confidence. This model has very high predictive ability. Moreover, we use all of the data to predict next two years' the monthly New York State alcoholic beverage tax as shown in Fig 15.

(2) Models with predictors

The estimated function of regression with lagged variables is:

$$A_t = 2.974 \times 10^3 + 7.401 \times 10^{-2} A_{t-6} - 4.150 \times 10^3 D_{t-6} + 7.448 \times 10^{-1} A_{t-12} + 2.394 \times 10^{-1} D_{t-6} A_{t-6} + w_t$$

where

$$D_{t-6} = \begin{cases} 0 & if\ A_{t-6} < 17,000 \\ 1 & if\ A_{t-6} \geq 17,000 \end{cases}$$

The adjusted R^2 is 76.67%. Then we evaluate it on test set. From Fig 16, all the predicted values are very close to their corresponding true values, indicating that this model performs well. Then we use total dataset to predict future alcoholic and beverage tax, as shown in Fig 17. The estimated function of relationship between predictors and the response is:

$$A_t = 1.376 \times 10^4 - 9.501 \times 10^2 sin(t) - 1.582 \times 10^3 cos(t) + 5.049 \times 10^4 U_t + 2.985 \times 10^{-2}(C_t - \bar{C_t}) + 3.407 \times 10^4 (P_t - \bar{P_t})$$

Among 24 observations, 4 true values fall out of the prediction interval, with 95% confidence (Fig 18).

Furthermore, we apply more complex model in the next step, compared to the previous analyses. We fit a ARIMA model with Cigarette Tax, Personal Income Tax and Unemployment rate. According to the AIC values, we choose the ARIMA (3,0,3) model with two predictors, which are Cigarette Tax and Personal Income Tax. Then we use this model to forecast the monthly New York State alcoholic beverage tax for the next 24 months. From Fig 19, 8 out of 24 observations fall out of the prediction interval, with 95% confidence. This prediction result is not desirable. The reason may be that we do not consider seasonal patterns. To improve this model, we apply a seasonal autoregressive integrated moving average model (SARIMA) on the training set. We choose the best model with AIC =3709.14, which is the SARIMA$(2,1,1)(1,1,0)_{12}$ model with Cigarette Tax and Personal Income Tax (model details in Fig 20).

Additionally, we predict the monthly New York State alcoholic beverage tax for the next 2 years and evaluate its performance. From Fig 21, only 2 out of 24 observations fall out of the prediction interval, with 95% confidence. This result is much better than the last one. To verify its reliability, we check assumptions of this model. From Fig 22, all of these lag coefficients are less than the significance level, indicating that the error terms are independent. We also plot the normal Q-Q plot, showing departure from normality at the tails due to the outliers (Fig 23). In order to further check the normality assumption, we conduct the Lilliefors normality test and the corresponding p-value is 0.05217, concluding that this model satisfies this assumption. Thus, it's a reasonable and exceptional model.

Discussion and Conclusions:

We apply several methods in our project (comparison details in Table 3). The regression model with lagged variables and SARIMA$(2,1,1)(1,1,0)_{12}$ are the best two

models. Tax fluctuates periodically, since the cycle of tax is one year. Thus, the autocorrelation between alcoholic and beverage tax at time t and alcoholic and beverage tax at time t-12 is the most significant one. Because of this tax season feature, the regression model with lagged variables reaches the best result. From previous analyses, we clearly discover seasonal patterns. That may be the reason why the SARIMA model performs better than the ARIMA model.

Bibliography

Taxes/Fees Collected Monthly - dataset by stateofny. (2016, October 13). Retrieved from https://data.world/stateofny/taxesfees-collected-monthly