Improved selection methods on external resources to get better performance on ARIMA prediction for GDP time series

Abstract:
Financial data usually contain time related trend and seasonality, which we prefer to analyze using time series. **Auto.arima()** in "**forecast**" package is widely accepted to be used in predicting ARIMA time series by minimizing the AIC, which also had parameters that will take **xreg** as the external resources in predicting future data. However, the function will not consider the relative performance of each external regressors, especially when there are huge amounts of data involved. Moreover, it will take the impact of the past value of external resources. This could take numerous of time in training and predicting. In this paper, we would show methods on selecting the best external regressors, the option of taking impact of their past values and evaluate their performances.
Table of 5 biggest industry.

Introduction.

The data set we used for training comes from Statistics Canada. We will introduce our methods performance on Agriculture, forestry, fishing and hunting (AFFH) Sector. We want to predict the GDP deflator, who has the following formula.

$$\text{GDP deflator} = \frac{\text{Nominal GDP}}{\text{Real GDP}} \times 100$$

We will further generalize our method and evaluate its performance in other industries.

1 Data Preprocessing

Before we built the model, the data preprocessing shows some high correlation of some features. It is reasonable since financial data for a specific sector are usually highly correlated. This problem is not a concern as long as for prediction, however for interpretation, it is a problem since it will shift the weights between related features and cause inconsistency. Another problem is for Hessian optimization, which is used in data validation section. A highly correlated would not guarantee the model consistency in cross validation. We will further explain in section three.

2 Model Improvement

We want to demonstrate two way for improving the impact of xreg. The original method in Auto.arima() for evaluating the weights of xreg is based on the following formula.

$y_t = \beta X_t + \epsilon_t$

$\epsilon_t = \varphi_1 \epsilon_{t-1} + ... + \varphi_p \epsilon_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... \theta_q \varepsilon_{t-q}$

Where ***y*** is the original data and ***x*** is xreg. There could be a potential problem, which has no selection process between these two steps; all xreg will be used. This could cause overfitting and further decrease the predicting abilities.

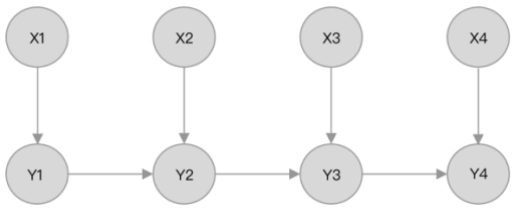
Approach A:

Our first approach uses ARIMA predict the original time series data at the beginning. We would expect the residual contain all the information linearly related to xreg. That information could be extracted using **StepAIC** linear fit. StepAIC is a generally accepted method for model selection, which outputs the model with minimum AIC in all combinations of xreg. The process could be formalized in the following notations.

$y_t = \varphi_1 y_{t-1} + ... + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... \theta_q \varepsilon_{t-q} + \epsilon_t$

$\epsilon_t = StepAIC(\beta X_t + \omega_t)$

This method will not take the impact of the past values of the xreg. The Bayesian Network representation is shown below.



We would introduce an alternative approach which will include the past value of xreg.

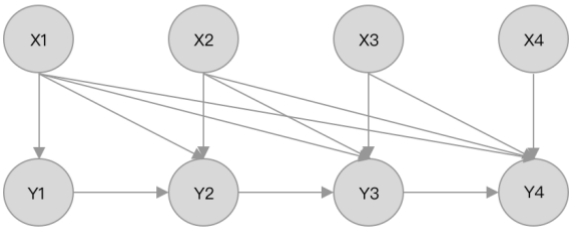
Approach B:

The second approach will first fit the original time series data StepAIC linear fit. Then use ARIMA to fit the residual, the math notation could show that this approach will take consider of the past value of the external regressors.

$$y_t = StepAIC(\beta X_t + \epsilon_t)$$

$$\epsilon_t = \varphi_1 \epsilon_{t-1} + ... + \varphi_p \epsilon_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... \theta_q \varepsilon_{t-q}$$

Our second approach is almost the same as the Auto.arima does, except we will do a StepAIC for parameters selection.



3 Model Validation

Traditionally AIC or BIC will be considered for model selection. However, intend to evaluate the absolute performance of our models. We would introduce cross validation for performance assessment. However, time series data is different from the usual data. As there exist high correlations between time lag, a traditional leave-one-out validation would have a poor performance. (Quote here, from announcement sent by Sam) We could use a specific designed Time Series Cross Validation in "**forecast**" package. The ideology is using leave-one-out based on time lag.

$$\epsilon = \sum_{i=1}^{n} (\widehat{y_{i-1}^{i}} - y_i)$$

However, model consistency now becomes a concern if we have highly correlated features. In another word, the model that predict value at time lag i based on information before i cannot guarantee to be consistent. If we set the parameters of ARIMA model to be consistent, some optimization errors will appear. The reasons could be various, the most two possible explanations are correlations and model generalization. We have mentioned the high correlation of some features at the beginning. In optimization, this could lead the xreg matrix to a singular matrix, which is not invertible and throw non optimizable error. Another possibility is the ARIMA model we set is for the whole datasets. It is totally possible it does not fit some partial data at all, which will cause this error. Intend to encounter this problem, we could trade the model consistency for a doable validation method; use Auto.arima instead of a set of defined parameters. However, this method could only be used to validate the efficiency of xreg, the model effect will be excluded. The comparison of model with/without xreg shows the xreg and the preselect significantly improve the performance of our model.

	AIC	BIC	Residual	tsCV MSPE	tsCV MAPE
ARIMA	417.36	427.3	328.4649	132.9285	0.1282
ARIMA With full Xreg	333.29	359.39	164.1785	96.3871	0.1176
ARIMA Approach A	338.09	358.16	188.1708	45.5178	0.0937
ARIMA Approach B	339.46	359.53	190.245	49.5747	0.0942

As we could see, Approach A has a better performance. We could agree more on that there is no time related impact of xreg to y.

4 Extended Dataset

This method could be applied in other financial data. Approach A could not guarantee to be better than Approach B, as this depend on the structure of xreg. However, we could see both Approach performs better on a relatively large external resource dataset than use xreg without selection. This could also help on interpretation if xreg has an independent structure.

5 Conclusion

External recourses would improve the model performance however the selection is critical, and we could improve this process.