Name: Anh Nhat Minh Nguyen

RUID: 228 007 570

Course: 22:544:608:60

PROJECT REPORT: A FIVE-YEAR FORECAST ON US ANNUAL BIRTH RATE

Introduction

The dataset used for analysis is the US Monthly Birth Data downloaded from Kaggle.com. The primary collecting agents are the National Bureau for Economic Research and the National Center for Health Statistics, and the data was collected from the National Center for Health Statistics (NCHS) and state vital records offices.

The initial dataset contains the monthly birth rate for each US county. To aggregate the data for analysis, Python was used to summarize all the county birth records during each month in every year. The resulting dataset contains 372 records, which are the monthly number of births from 1985 to 2015.

R programming language is used for this forecasting project.

Exploratory data analysis

Different statistical analysis was performed on the dataset in order to understand the data distribution. The analysis includes plotting the time series, the ACF plot, the data histogram, and the box plot, decomposing the data, and exploring key data summary metrics such as the mean, median, and interquartile range.

As per the Exploratory data analysis, a few patterns from the birth data were detected:

- The time series has a highly seasonal pattern, birth records are highest during July and August and lowest in February and November, suggested by the seasonal indices from the decomposition model.
- There is an underlying fluctuating trend in the dataset. The data changes its direction every 5 years on average.
- 75% of birth records in the US range from 310,00 to 370,000, with the median being nearly 330,000. There are no outliers detected in the dataset.

- The highest correlation is between 12 lags. That is, birth rates for different months in every year are highly correlated. The ACF plot also suggests that the birth record for each month can also be affected by its previous and following months.

Forecasting questions and evaluating criteria

The forecast results can be used by the US government to adjust their policies regarding family support, maternal care, education, etc accordingly over the following 5 years. The forecasting question stated is: What are the projected birth records for the following 5 years?

The evaluating criteria are the Mean Absolute Percentage Error (MAPE) and the residual validation. The divergence in percentage will help the governmental decision-makers estimate their resources such as grants, human resources, facilities, ect in the next five years.

Since there were no significant changes in the data pattern between 1985 and 2015, the full dataset was used as input for the forecasting models. The models include 8 models: Naïve, Mean, Simple Moving Average (MA), Simple Exponential Smoothing (SES), Holt-Winters, ARIMA, Decomposition, and Simple Linear Regression (SLR).

Forecasting summary

The fitting process and forecasting were executed on each of the stated models. Afterward, the residual analysis was performed to evaluate whether the model could capture all patterns of the data set. For the Moving Average (MA) model, the model with a window of 3 was selected among three models with the windows being 3,6,12. For the MA models, the residual analysis was performed only on the MA(3) model.

For the Simple Linear Regression model, the model did not pass the validation process since the predictor can explain 3% of the variation in the response variable. The suggestion is that the model should be elevated into a multivariable regression model, in which more socioeconomic predictors will be used.

The evaluation of the models are provided in the below table:

Model	MAPE	Forecasting plot	Residual analysis
Naive	4.30	The value for the last period: December 2015, will be used as future forecasts.	The naïve model cannot capture the seasonal component of the time series. Errors are higher for larger records
Mean	4.51	The mean birth records (327,413.5) will be used as future forecasts.	The mean model cannot capture most patterns of the time series. Both residuals and ACF plot are similar to those of the original time series
MA(3)	2.49	The data is smoothed out and deviates more from the actual values as the forecasting windows increase.	The residuals for MA(3) appear to be more random, with no seasonality and trend detected. The error range is the smallest among all models.
SES	3.80	The plot suggests the SES model tends to underestimate high values and overestimate low values	The SES model cannot fully capture the seasonality component, as indicated by the ACF plot. Residuals are scattered and are less likely to be close to 0.
HW	1.29	The HW model improves its performance after the data point 1990. Among alpha, beta, gamma, a	The residuals for HW appear to be random, with no seasonality and trend detected.

		higher alpha (0.2) was given to the most recent data point.	
Decomposition	1.23	The decomposition model performs better as the time period moves further.	The residuals for Decomposition model appear to be random, with no seasonality and trend detected.
ARIMA	1.16	Selected model: ARIMA(4,1,0)(0,1,1)[12] Forecastsing results fit the actual data well.	The residuals for ARIMA model appear to be random, with no seasonality and trend detected.

Summary

The ARIMA model provides the best predictions with the lowest MAPE and a valid residual analysis. ARIMA's performance is closely followed by that of the Decomposition and the Holt-Winters model.

The Holt-Winters should be the selected model due to its simplicity and less demanding processing requirements compared to the ARIMA and Decomposition. The MAPE provided by Holt-Winters is 1.29%, which is very slightly lower than that of the other two models. Also, the provided alpha, beta, and gamma parameters will help the decision maker better understand how the result was generated, or which aspects of the data set's pattern are prioritized in making the forecasts.

Lastly, there is a potential for a multivariable regression model because realistically, multiple factors may affect a human's decision to give birth. For this method, more data on different socioeconomic predictors such as employment rate, maternal care, etc, will be needed.