Math 320 Research Project: Forecasting Birth Rates by Race

Rachel Hong and Joshulyne Park

November 28, 2016

Contents

Summary

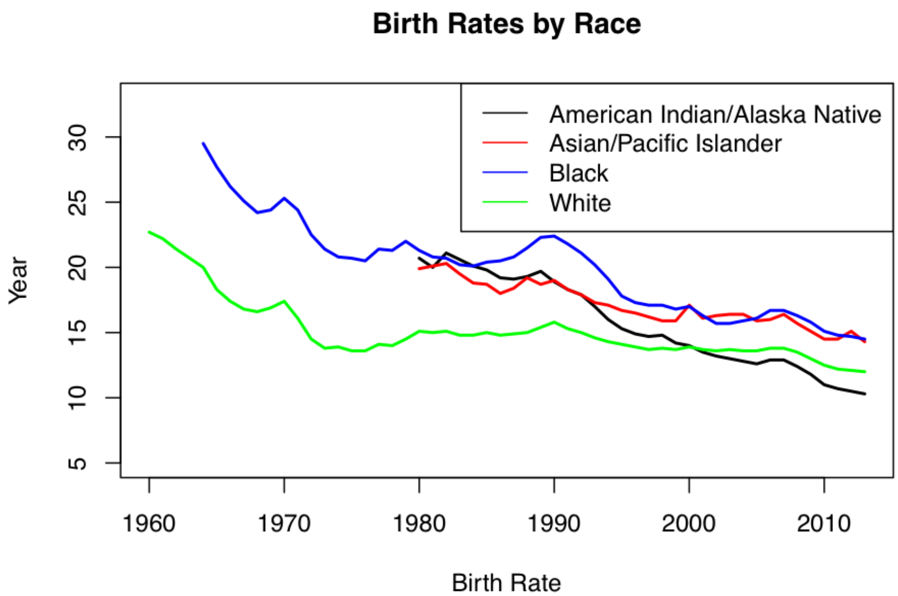
Our project focuses on a dataset from the National Center for Health Statistics, "NCHS - Births, Birth Rates, and Fertility Rates, by Race of Mother: United States, 1960-2013." It contains data regarding birth rates, categorized by race of the mother from the time period 1960-2013. The general perceived notion that we are exploring is that birth rates have been in a decline post the "baby boom" years. To confirm this idea, we will perform experiments to develop a more detailed understanding of this trend. Using R, we will build time series models that forecast birth rates for each race. The three main models we will examine and implement are: linear, exponential, and autoregressive integrated moving average (ARIMA). All three are commonly used time series models, starting from the most simple to a more sophisticated model that takes into consideration important time series elements such as trend and seasonality. While some of these models are already functions within R, we will break down the mathematical formulas used to generate these models and the forecasts. For each dataset, we will find the best model by analyzing the model fit and the residuals and ultimately select the best model with the smallest error, or the root mean squared error. Once our model is chosen, we will forecast birth rates for each subset of data for the next 10 years.

Models: In-Depth Analysis

Please see the attached document named "In-Depth Analysis on Linear, Exponential, and ARIMA models".

Data Prep

While the data was generally clean, there were a few missing values that needed to be taken out. After the data consisted of only relevant values, we organized the data into four subsets, one for each race in order to run our time series models. Below is a plot of birth rates by race.



Time Series Models

Linear Model

10 Year Horizon Forecast Values for Race: American Indian/Alaska Native

```
##      Point Forecast    Lo 95     Hi 95
## 2014       9.672193 8.326344 11.018041
## 2015       9.328755 7.976330 10.681179
## 2016       8.985317 7.625986 10.344648
## 2017       8.641879 7.275316 10.008443
## 2018       8.298442 6.924325  9.672558
## 2019       7.955004 6.573018  9.336989
## 2020       7.611566 6.221402  9.001730
## 2021       7.268128 5.869481  8.666775
## 2022       6.924691 5.517261  8.332120
## 2023       6.581253 5.164748  7.997758
```

10 Year Horizon Forecast Values for Race: Asian/Pacific Islander

```
##      Point Forecast    Lo 95    Hi 95
```

```
## 2014       14.29893 13.12239 15.47547
## 2015       14.13752 12.95524 15.31981
## 2016       13.97612 12.78780 15.16444
## 2017       13.81471 12.62007 15.00936
## 2018       13.65331 12.45206 14.85456
## 2019       13.49190 12.28377 14.70003
## 2020       13.33050 12.11522 14.54577
## 2021       13.16909 11.94640 14.39178
## 2022       13.00769 11.77731 14.23806
## 2023       12.84628 11.60797 14.08458
```

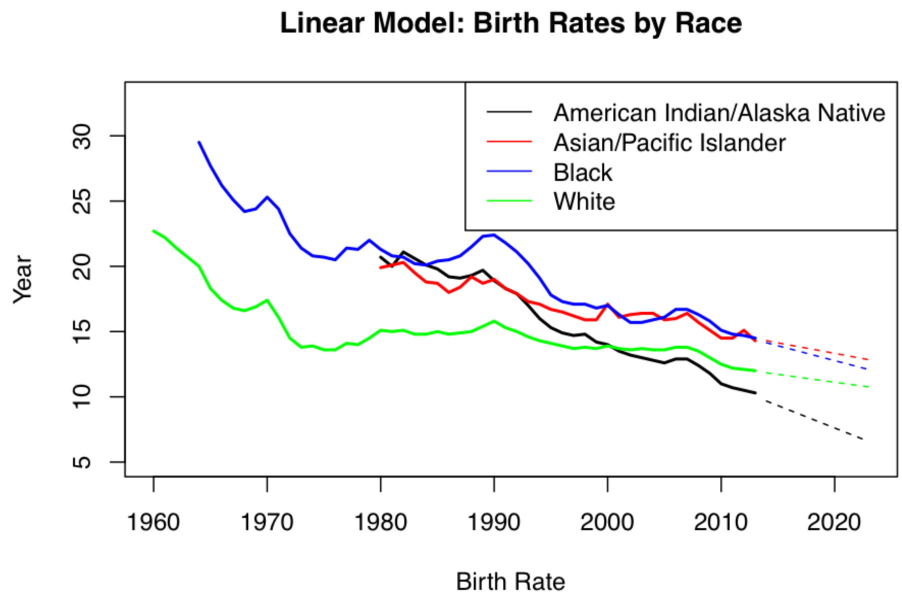
10 Year Horizon Forecast Values for Race: Black

```
##      Point Forecast     Lo 95    Hi 95
## 2014       14.14220 11.285594 16.99881
## 2015       13.91406 11.050863 16.77725
## 2016       13.68591 10.815895 16.55592
## 2017       13.45776 10.580691 16.33482
## 2018       13.22961 10.345254 16.11396
## 2019       13.00146 10.109584 15.89334
## 2020       12.77331  9.873685 15.67294
## 2021       12.54516  9.637557 15.45277
## 2022       12.31701  9.401203 15.23282
## 2023       12.08886  9.164624 15.01311
```

10 Year Horizon Forecast Values for Race: White

```
##      Point Forecast    Lo 95    Hi 95
## 2014       11.83990 8.663654 15.01615
## 2015       11.71987 8.537330 14.90241
## 2016       11.59984 8.410794 14.78889
## 2017       11.47981 8.284048 14.67557
## 2018       11.35978 8.157092 14.56247
## 2019       11.23975 8.029929 14.44957
## 2020       11.11972 7.902560 14.33688
## 2021       10.99969 7.774986 14.22439
## 2022       10.87966 7.647208 14.11211
## 2023       10.75963 7.519228 14.00003
```


3



Exponential Model

10 Year Horizon Forecast Values for Race: American Indian/Alaska Native

```
##      Point Forecast    Lo.95     Hi.95
## 2014      10.349156 9.512041 11.259943
## 2015      10.120373 9.297930 11.015563
## 2016       9.896646 9.088451 10.776711
## 2017       9.677866 8.883509 10.543254
## 2018       9.463922 8.683014 10.315061
## 2019       9.254708 8.486877 10.092006
## 2020       9.050118 8.295009  9.873967
## 2021       8.850052 8.107324  9.660823
## 2022       8.654408 7.923737  9.452457
## 2023       8.463089 7.744164  9.248755
```

10 Year Horizon Forecast Values for Race: Asian/Pacific Islander

```
##      Point Forecast    Lo.95    Hi.95
## 2014       14.45608 13.51625 15.46126
## 2015       14.32073 13.38531 15.32153
## 2016       14.18666 13.25541 15.18332
## 2017       14.05384 13.12657 15.04660
## 2018       13.92226 12.99877 14.91136
## 2019       13.79191 12.87201 14.77756
## 2020       13.66279 12.74629 14.64519
## 2021       13.53487 12.62160 14.51422
```

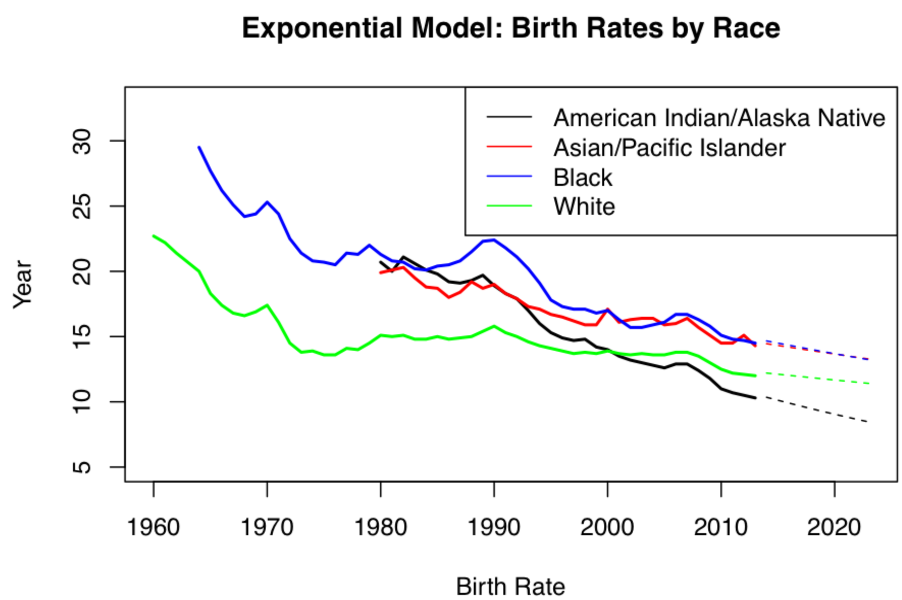
```
## 2022       13.40815 12.49795 14.38464
## 2023       13.28262 12.37533 14.25643
```

10 Year Horizon Forecast Values for Race: Black

```
##      Point Forecast    Lo.95    Hi.95
## 2014       14.66684 12.87004 16.71450
## 2015       14.49956 12.71942 16.52883
## 2016       14.33418 12.57042 16.34541
## 2017       14.17069 12.42304 16.16420
## 2018       14.00906 12.27725 15.98516
## 2019       13.84928 12.13304 15.80828
## 2020       13.69132 11.99040 15.63352
## 2021       13.53516 11.84932 15.46085
## 2022       13.38078 11.70977 15.29024
## 2023       13.22816 11.57175 15.12168
```

10 Year Horizon Forecast Values for Race: White

```
##      Point Forecast     Lo.95    Hi.95
## 2014       12.20612 10.191368 14.61917
## 2015       12.11570 10.112260 14.51606
## 2016       12.02595 10.033646 14.41386
## 2017       11.93687  9.955523 14.31255
## 2018       11.84845  9.877892 14.21211
## 2019       11.76068  9.800750 14.11255
## 2020       11.67356  9.724097 14.01386
## 2021       11.58709  9.647931 13.91601
## 2022       11.50126  9.572251 13.81900
## 2023       11.41606  9.497056 13.72283
```



5

Autoregressive Integrated Moving Average Model

10 Year Horizon Forecast Values for Race: American Indian/Alaska Native

```
##      Point Forecast    Lo 95     Hi 95
## 2014       9.984848 9.163265 10.806432
## 2015       9.669697 8.507802 10.831592
## 2016       9.354545 7.931520 10.777571
## 2017       9.039394 7.396226 10.682562
## 2018       8.724242 6.887125 10.561360
## 2019       8.409091 6.396629 10.421552
## 2020       8.093939 5.920232 10.267646
## 2021       7.778788 5.454997 10.102578
## 2022       7.463636 4.998884  9.928388
## 2023       7.148485 4.550408  9.746562
```

10 Year Horizon Forecast Values for Race: Asian/Pacific Islander

```
##      Point Forecast     Lo 95    Hi 95
## 2014       14.13030 13.123130 15.13748
## 2015       13.96061 12.536249 15.38496
## 2016       13.79091 12.046435 15.53538
## 2017       13.62121 11.606866 15.63556
## 2018       13.45152 11.199408 15.70362
## 2019       13.28182 10.814759 15.74888
## 2020       13.11212 10.447392 15.77685
## 2021       12.94242 10.093709 15.79114
## 2022       12.77273  9.751209 15.79425
## 2023       12.60303  9.418070 15.78799
```

10 Year Horizon Forecast Values for Race: Black

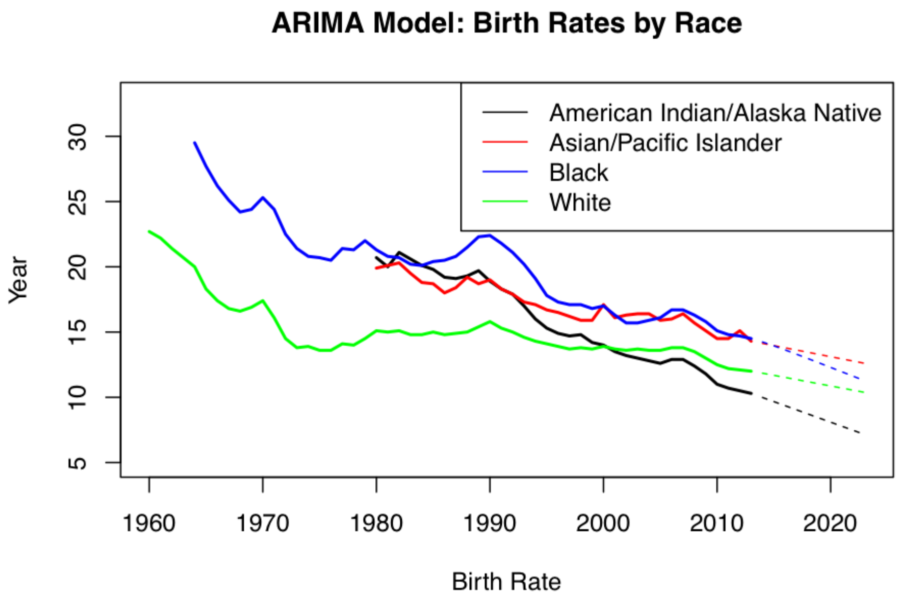
```
##      Point Forecast     Lo 95    Hi 95
## 2014       14.23952 13.156519 15.32251
## 2015       13.94421 11.923109 15.96532
## 2016       13.62886 10.739130 16.51860
## 2017       13.30198  9.626053 16.97790
## 2018       12.96845  8.583371 17.35353
## 2019       12.63109  7.603422 17.65877
## 2020       12.29154  6.677137 17.90594
## 2021       11.95072  5.796088 18.10535
## 2022       11.60916  4.953081 18.26525
## 2023       11.26719  4.142192 18.39219
```

10 Year Horizon Forecast Values for Race: White

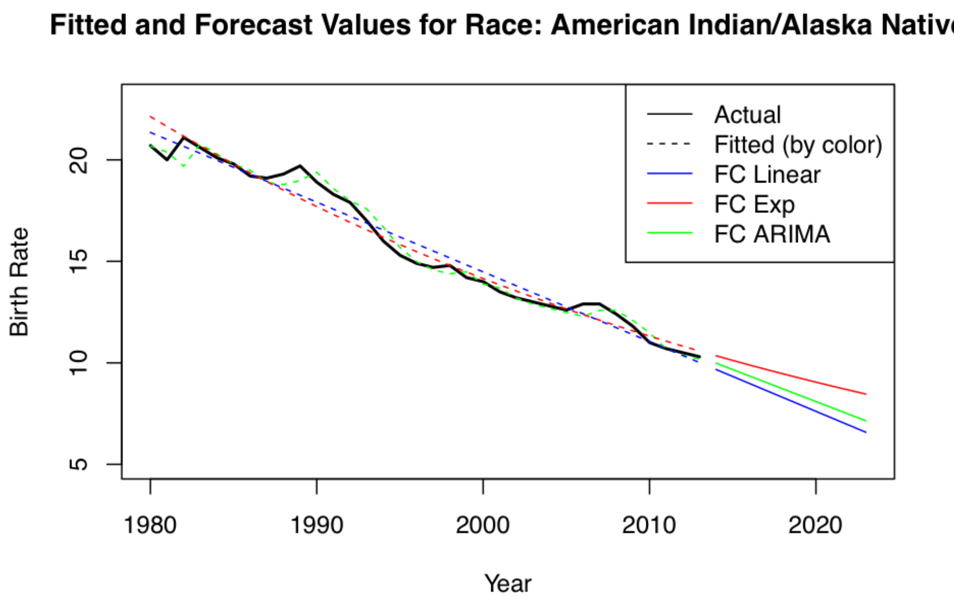
```
##      Point Forecast     Lo 95    Hi 95
## 2014       11.84784 11.043678 12.65201
## 2015       11.68438 10.126049 13.24271
## 2016       11.52092  9.357491 13.68434
## 2017       11.35745  8.626565 14.08834
## 2018       11.19399  7.905307 14.48267
```


6

```
## 2019       11.03053  7.182348 14.87871
## 2020       10.86706  6.452099 15.28203
## 2021       10.70360  5.711531 15.69567
## 2022       10.54014  4.958916 16.12136
## 2023       10.37667  4.193249 16.56010
```

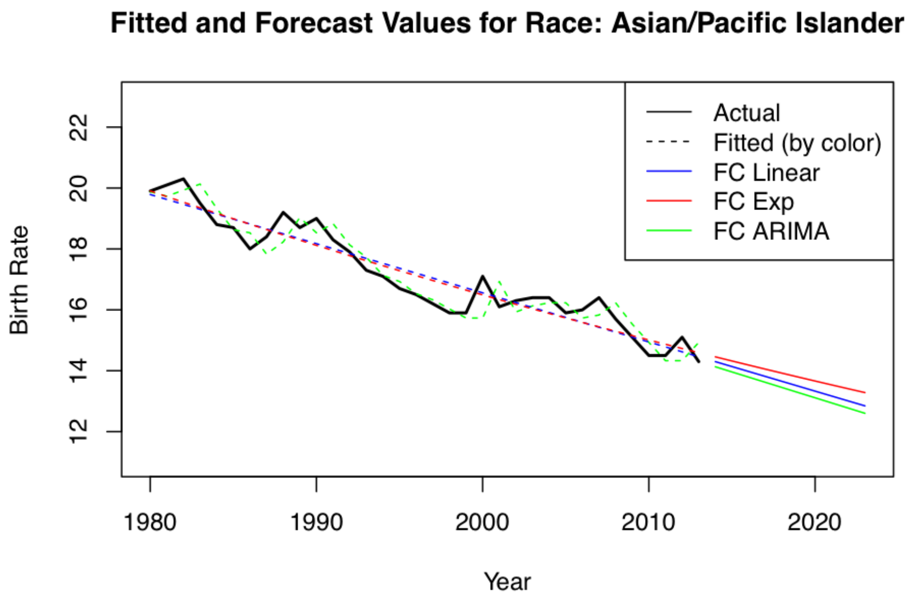


Forecast Comparison / Model Selection



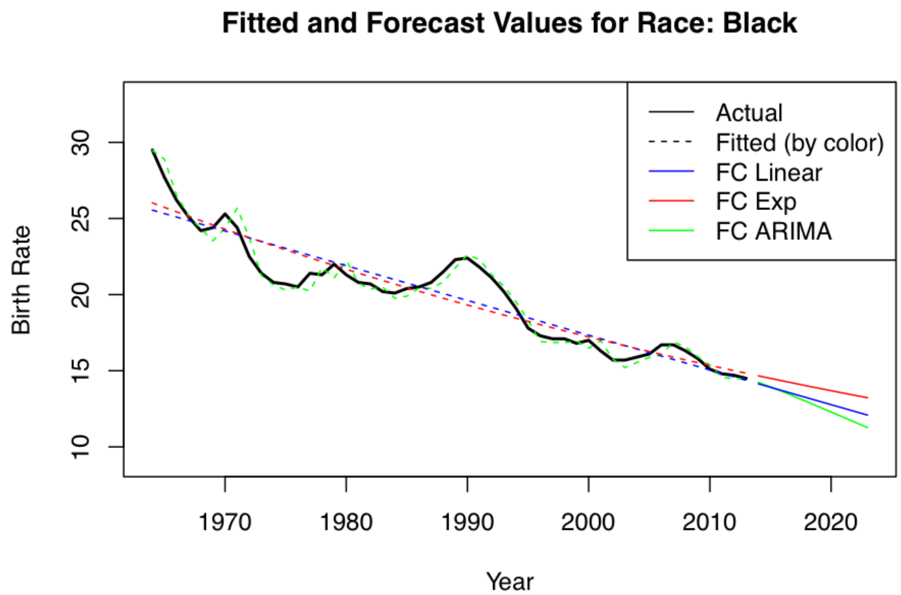
Model	RMSE Value
Linear	0.6048757
Exponetial	0.6541382
ARIMA	0.4066674

8

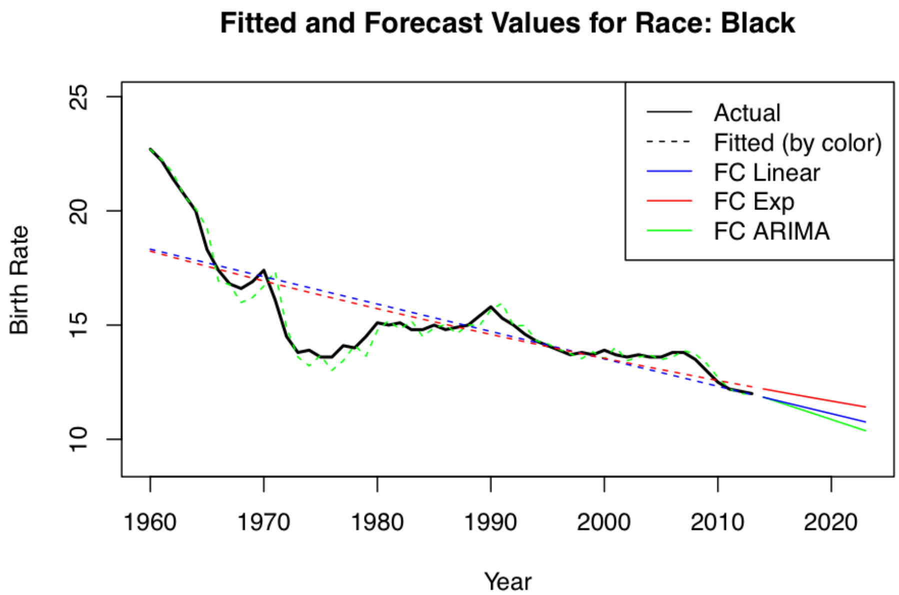


Model	RMSE Value
Linear	0.5287808
Exponetial	0.5139744
ARIMA	0.4985302

9



Model	RMSE Value
Linear	1.33798
Exponetial	1.309761
ARIMA	0.5357265



Model	RMSE Value
Linear	1.497296
Exponetial	1.455345
ARIMA	0.3948074

Conclusion

For all of our models, the RMSE values revealed that the ARIMA model was the best fit for each of subset of data. This is unsurprising as the ARIMA model really takes into consideration the historical data in creating its forecasts unlike the linear and exponential model. In conclusion, our assumption that birth rates are in a decline were confirmed to be true. Although some datasets had periods of increase, overall the general declining trend overwhelmed such periods.

In-Depth Analysis on Linear, Exponential, and ARIMA Models

Linear Model

Linear least squared regression is a basic method for finding the relationship between two variables. It is used in time series modeling to find a trend line that represents the observed values (the dependent variable) as a linear function of time (the independent variable) so that we can see how it changes over the years.

The result is a linear equation in the form $y_t = \beta_t x_t + \epsilon$, where y_t is the dependent variable and x_t is the vector of independent variables, which are also called the regressors or the predictor variables. For our model, x_t represents time t. β_i is the vector of regression coefficients, which represents the change in y for a one-unit change in x_t – in other words, it is the slope; β_0 is the intercept ($x_0 = 1$); ϵ is the error term. The residual is the difference between the approximate value of y_t and the true value of y, and with the least-squared method, the linear equation is determined by finding the values for β_t that minimize the sum of the squared residuals.

For our data set, we conduct an analysis of birth rates over time for each race. Since our data is time series, our independent variable is time in units of years and our dependent variable is the birth rate by rate. β_1 represents the amount that we expect the dependent variable (birth rate) to change each year, and β_0 represents the birth rate during the first year in our analysis.

To find the coefficients that minimize the sum of the squared residuals, we can write the equation $y_t = \beta_t x_t + \epsilon$ in matrix form:

$$\{y\} = \{\beta\}[X] + \{\epsilon\}$$

For our model, we have:

$$[X] = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_2 \end{bmatrix}$$

$\{\beta\}$ contains the coefficients, $\{\epsilon\}$ contains the residuals, and $\{y\}$ contains the observed values of the dependent variables. They are all column variables. To solve for the coefficients, we solve the system of equations:

$$\{\beta\} = ([Z]^T[Z]) \backslash ([Z]^T * \{y\})$$

To solve for the residuals between our fitted linear equation and the actual values of the data, we use the equation:

$$\{\epsilon\} = \{y\} - \{y_i\} = \{y\} - [X]\{\beta\}$$

Exponential Trend Model

Exponential models are useful and common in time series because it allows us to look at growth rates over time. Often, the observed data does not follow a linear trend, but the change in the observed data does. Exponential modeling is especially applicable when looking at the growth of a population, which is often non-linear when we look at it in intervals but linear once we take the logarithm.

The exponential equation can be written as $y_t = \beta_0 e^{\beta_1 x_t}$, where the birth rate has a constant rate of growth at rate β. If we take the logarithm of this, then our equation becomes:

$$\ln(y_t) = \ln(\beta_0) + \beta_1 x_t$$

Thus, $\ln(y_t)$ is a linear function of x_t. For our analysis, x_t is time, so the interpretation is that $\ln(y_t)$ is a linear function of time, and thus we can solve for it with linear least squares regression and then use the exponentiation of our result to find β.

To find the coefficients, we write the logarithm of our original exponential equation $y_t = \beta_0 e^{\beta_1 x_t}$ and put it in matrix form:

$$\{\ln(y_t)\} = \{\beta\}[X] + \{\epsilon\}$$

Thus we can solve the linearized equation of the exponential function in the same way as the linear function.

Autoregressive Integrated Moving Average Model

The autoregressive integrated moving average (ARIMA) model is a combination of three different time series components designed to create the best fit model for a time series data set. The autoregressive (AR) component can be simplified to a stochastic difference equation. It is a model in which the current value of the series is linearly related to its past values with an additive shock value. The moving average (MA) component takes another approach to time series modeling. Utilizing the fact that variation in time series data is driven by past shocks, the approach takes distributed lags of current and past shocks to model the current value of the series. Lastly, the integrated aspect of the ARIMA model takes into consideration a very important concept in time series modeling, stationarity. A stationary time series is a data set whose properties do not depend on time. Therefore, a dataset that exhibits trend or seasonality is not stationary, however, a white noise series is stationary – random and independent of when it is observed.

To stabilize this variance dependent on time, the integrated component calculates the differences between consecutive values in the time series, a process known as differencing:

$$y'_t = y_t - y_{t-1}$$

If the data set does not appear stationary after differencing, it may be necessary to difference the data a second time, or second-order differencing.

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

In lag operator form or backshift notation, differencing of the time series data set x_t is represented by:

$$y_t = (1 - L)^d x_t$$

The AR(p) model is represented by:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + e_t$$

where c is a constant, $\phi_1, \ldots, \phi_p$ are the parameters of the model, and e_t is white noise.
In lag operator form or backshift notation:

$$\Phi(L) y_t = (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) y_t = e_t$$

The MA(q) model is represented by:

$$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} = c + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t$$

where c is a constant, $\theta_1, \ldots, \theta_q$ are the parameters of the model, and e_t is white noise.
In lag operator form or backshift notation:

$$y_t = c + \left(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q\right) e_t = \Theta(L) e_t$$

Combining all three models results in the following autoregressive integrated moving average ARIMA(p,d,q) model and is represented by:

$$y_t = (1 - L)^d x_t \qquad (I)$$

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \qquad (AR \,\&\, MA)$$

where p is the order of the autoregressive model and q is the order of the moving average model.
In lag operator form or backshift notation:

$$y_t = (1 - L)^d x_t \qquad (I)$$

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) y_t = \left(1 + \sum_{i=1}^{p} \theta_i \, L^i\right) e_t \qquad (AR \,\&\, MA)$$

To determine the p, d, and q parameters of the ARIMA model, the Box-Jenkins approach utilizes the shape of the autocorrelation and the partial autocorrelation functions for model identification. The autocorrelation function measures autocorrelations within the dataset, or the relationship between y_t and y_{t-k} for various values of k. The partial autocorrelation function also measures autocorrelations within the dataset, but takes into consideration the effects of other time lags and consequently removes them.

For the AR(p) process, the autocorrelation function exhibits either an exponential decay to zero or an alternating positive and negative decay to zero shape. The partial autocorrelation function of the AR(p) process becomes zero at lag $p + 1$ and greater, allowing one to identify the p paramenter.

For the MA(q) process, the autocorrelation function exhibits one or more spikes and becomes zero at the lag $q + 1$ and greater, allowing one to identify the q parameter without the need to observe the partial autocorrelation function. To select the optimal combination of these p, d, and q parameters for the ARIMA model, the Akaike information criterion (AIC) is utilized. The AIC is calculated by the following:

$$AIC = -2 \log(L) + 2(p + q + k + 1)$$

where L is the likelihood of the data, p is the order of the autoregressive process, q is the order of the moving average process, and k is number of parameters in the model. The objective is to minimize the AIC values in order to obtain a good model fit. The AIC value, however, can only be used to compare ARIMA models with the same order of differencing.

The `auto.arima` function in R chooses the optimal p, d, and q parameters by selecting the combination of parameters that result in the lowest AIC value.