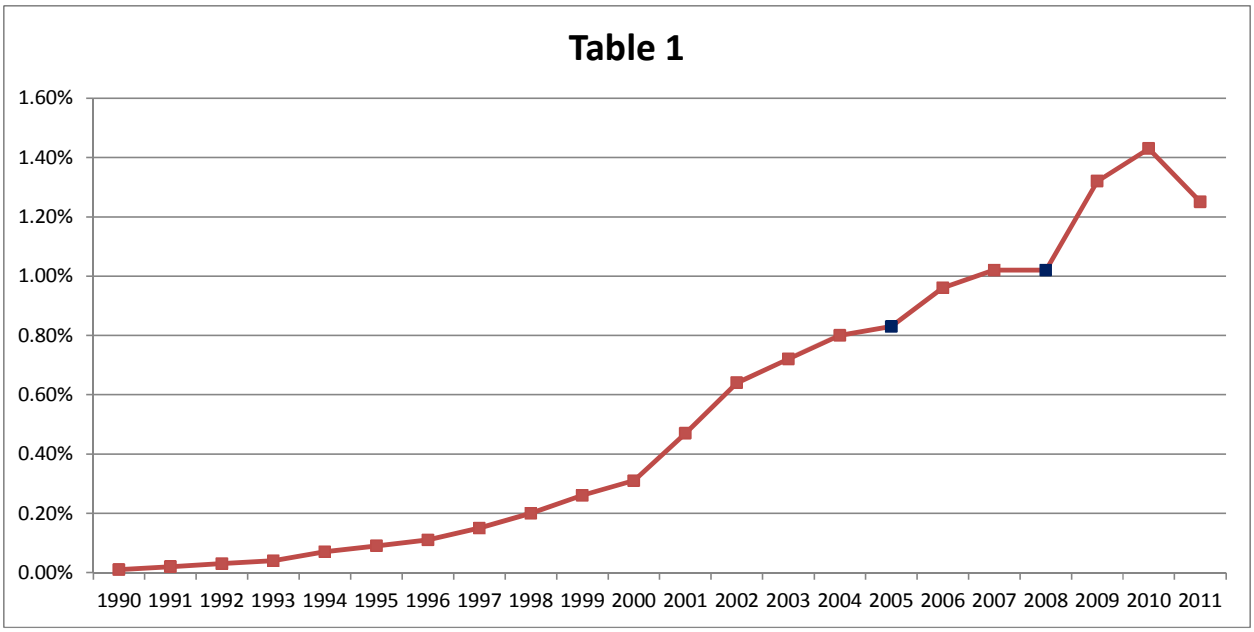
Popularity of Jacob

By Kimberly Walker

Introduction:

There are many considerations when selecting a name for your baby, including meaning, family tradition, past associations, and popularity. The media and pop culture can have a large impact on the popularity of names. Stochastic events, such as a popular television show or athlete can increase the popularity of a name. Like many females across the country, I read and watched all the Twilight books/movies. In 2005, the first Twilight book was released. Since the release of the book, the name of the main character – Isabella – has jumped to the top as the number one name for girls. Figure 1 shows the percentage of girls named Isabella since 1990. As shown, there steady increase in popularity until 2008, when there is a large spike in popularity. I believe this spike in popularity is a result of the release of the first Twilight movie in theaters across the United States. Originally, I had wanted to analyze the popularity the Twilight epidemic had on the popularity of the name Isabella. There is a lack of credible data, however, as there is no data for the name prior to 1990. This is either do to the lack of popularity of the name, and thus Isabella was not in the 1,000 most popular names during those years, or because the Social Security Administration simply did not have any data for those years. In addition, there is only 5-6 years of data after the release of the first book. As a result of the lack of credible data, I chose a different name, Jacob, which is also featured in the twilight series. In this project, I use different modeling techniques to determine a model that accurately predicts the popularity of the name Jacob.

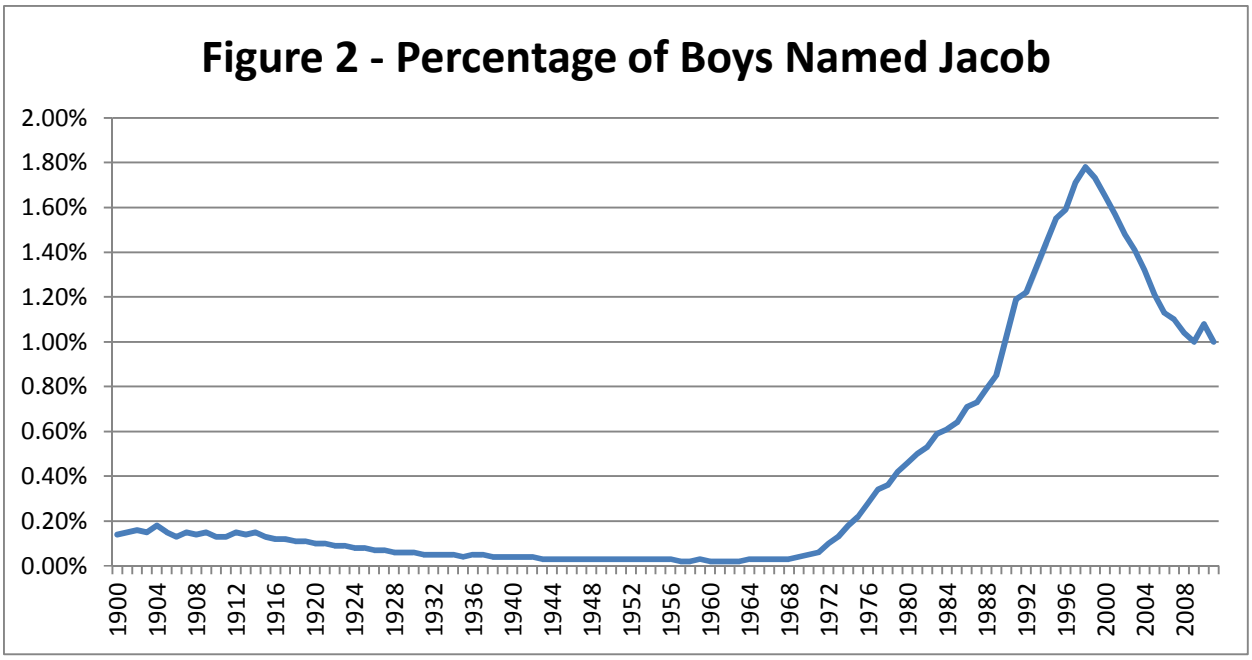


Data:

I obtained historical data on the percentage of baby boys named Jacob from the website http://babynamesworld.parentsconnect.com/popularity_of_Jacob.html. From the website, I was easily able to obtain a listing of the percentage of baby boys born each year named Jacob using data from the Social Security Administration database. The data is only available for years during which the name Jacob was in the top 1,000 ranked names and excludes years prior to the formation of the Social Security Administration. The data also excludes anyone who did not apply for a Social Security number and illegal immigrants.

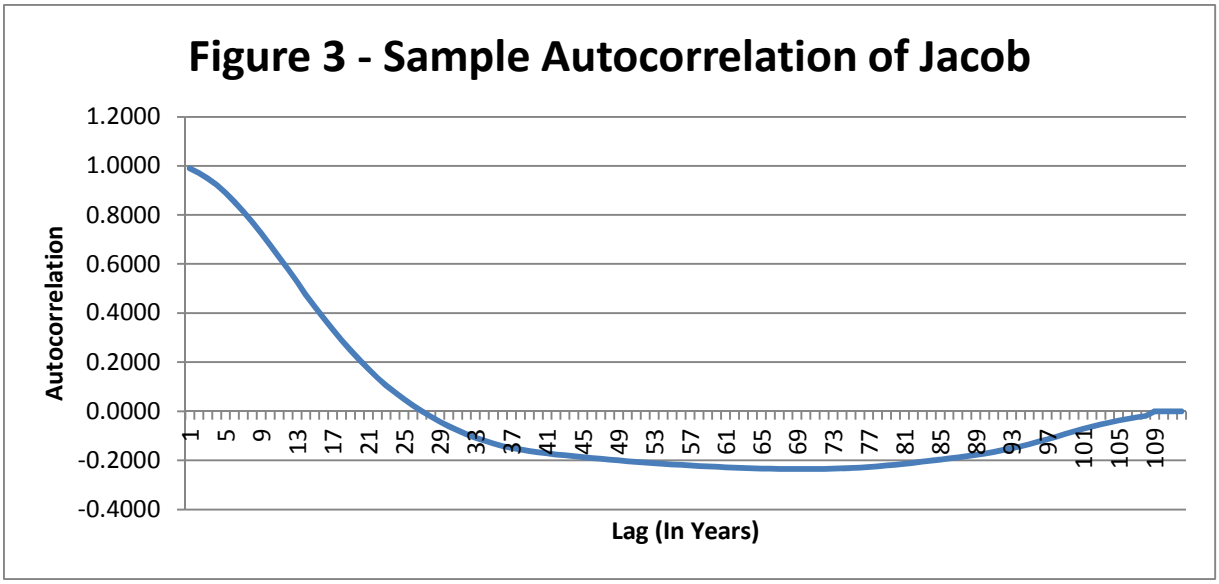
In Figure 2 below, 112 years of yearly data, from 1900-2011, are displayed for the male name Jacob. There are a few trends in the data:

- A steady, slow decline in popularity from 1914-1968
- A rapid increase in popularity from 1968-1998, when it hits a peak
- Decline from 1998-2011

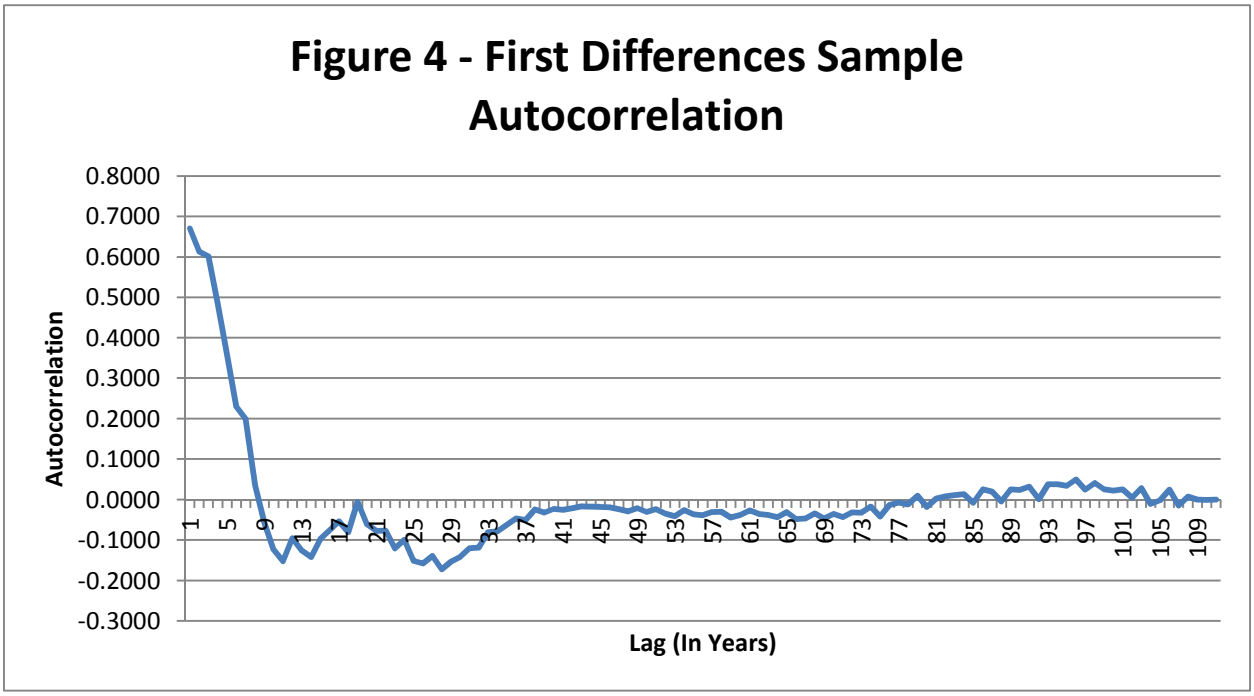


Test of Stationarity:

Figure 3 below shows the graph of the autocorrelation function for the name Jacob. The graph indicates that the time series is not stationary because it is u-shaped about zero. There are no definite patterns or seasonality related to this time series, so I conclude that no seasonal adjustment is necessary.



To create a stationary time series and remove the trend of the original time series, I took the first differences of the time series data. The 1st difference is calculated by taking the percentage of boys named Jacob in year (X) minus the percentage of boys named Jacob in year (X-1). Figure 4 below shows the graph of the autocorrelation function of the 1st difference of the time series. The transformed is a stationary series which declines more rapidly to zero and then remains closer to zero.



Process for Modeling the Data

The sample autocorrelation function for the times series data for the popularity of the name Jacob as a percent of total boy births shows that there are correlations for up to eleven years. There also appears to be a gradual drop in the popularity of the name Jacob which makes selecting an autoregressive model a better fit than selecting a moving average time series model. It is possible that there may be a moving average component to the data, but I do not think it will have a significant impact.

Process for Modeling the Data

I used the regression add-in function in Excel to fit autoregressive models to the time series. The below models were fit to the time series of the percentage of baby boys named Jacob:

AR(1): $= 0.00007149 + 1.0016$

AR(2): $= 0.00004104 + 1.6966 \quad - 0.7028$

AR(3): $= 0.00004367 + 1.4729 \quad - .1701 \quad - 0.3112$

AR(4): $= 0.00005116 + 1.4230 \quad - 0.3171 \quad + 0.1934 \quad - 0.3109$

AR(5): $= 0.00004643 + 1.4383 \quad - 0.3305 \quad + 0.2152 \quad - 0.3811 \quad = 0.04731$

For each of the above five models, I tested the fit using the model diagnostics summarized below:

	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Durbin-Watson Statistic	0.6276	2.2106	2.0522	1.8921	1.8906
Box Pierce Statistic (Q)	195.08	26.72	22.22	19.25	17.33
Chi-Square (10% confidence)	24.77	24.77	23.54	23.54	22.31
Adjusted R-squared	0.9923	0.9959	0.9961	0.9964	0.9964

The Durbin-Watson statistic is used to check the residuals for serial correlation and lies in the range of 0-4, with 2 indicating no serial correlation. A Durbin-Watson statistic value less than 2 indicates positive correlation and a value of greater than 2 indicates negative serial correlation. Except for AR(1), the results of the Durbin-Watson test indicate that there is no serial correlation. The test statistic for AR(1) indicates that there is positive correlation.

By itself, the Durbin-Watson statistic is not a valid measure, so I also calculated the Box-Pierce statistic. This statistic is used to determine whether or not to reject the null hypothesis that the residuals are a white noise process. If the Box-Pierce Q statistics is lower than the corresponding critical chi squared values (at the 10% significance level), then we do not reject the null hypothesis. For the Box-Pierce statistic, I chose K=20. As shown above, AR(1) and AR(2) have Q statistics less than the associated Chi-Square p-value, whereas, AR(3), AR(4), and AR(5) do

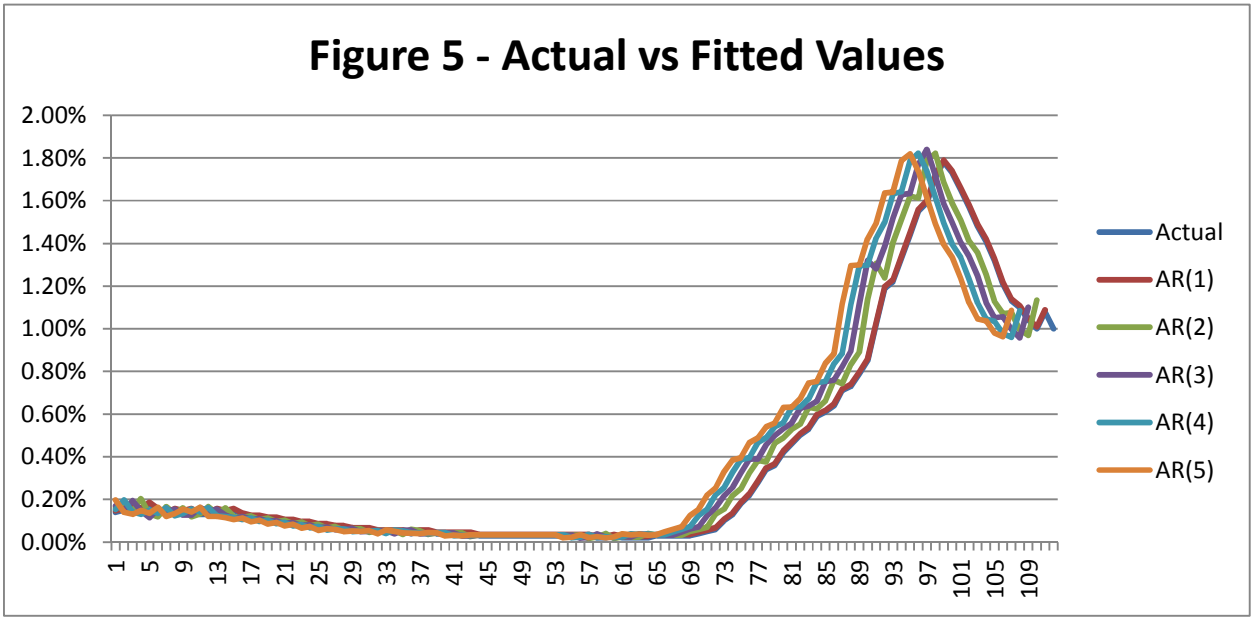
not. Therefore, we reject the null hypothesis that the residuals form a white noise process for AR(1) and AR(2) only.

I also reviewed the Adjusted R-squared value as a measure of goodness of fit. This statistic provides a measure of how accurate the model is at predicting future outcomes. The Adjusted R squared values for each model is approximately 1.0, indicating each model is very accurate at predicting future outcomes.

AR(2) appears to be the best model as the Durbin-Watson statistic is close to 2, we are able to reject the null hypothesis that the model is only detecting noise, and the Adjusted R-squared value is close to 1.0.

Model Evaluation:

In order to confirm the outcome that AR(2) is the best fit for modeling the popularity of the name Jacob, I compared the graph actual percentage of boys named Jacob series to the forecasted values for each of the models. All of the models appear to be close, but the AR(2) line in green appears to be closest to the actual values.



Conclusion:

The purpose of this project was to determine one model, not an exhaustive list of models or the absolute best model. In making my determination, I considered test statistics such as Durbin-Watson, Box-Pierce, Chi-Squared, and the Adjusted R-squared. I also considered the effects that seasonality and stationarity have on the models. After running five different models, I have determined that AR(2) is the best fit for modeling the popularity of the name Jacob for boys in the United States. While this may not be the best method that exists, this is the best model in this exercise.