Business Forecasting Project

5-year forecast for US monthly birth data

Fall 2024

Anh Nhat Minh Nguyen

Index

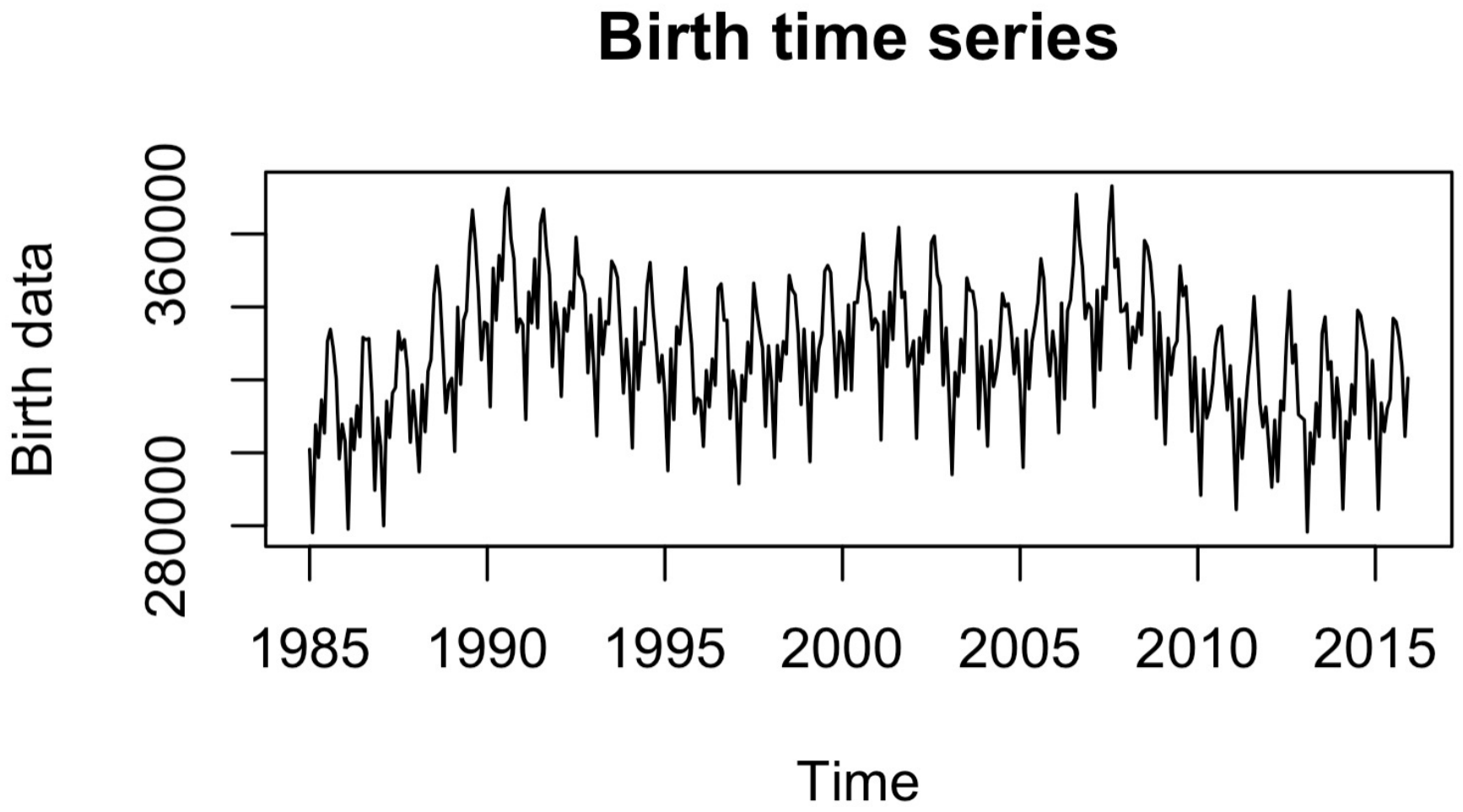
1. Introduction: US Birth data
2. Exploratory Data Analysis
3. Forecasting question & evaluation criteria
4. Forecasting models
5. Summary

1. Introduction: US Birth data

- **Data source:** Kaggle.com
- **Collecting agent:** National Bureau for Economic Research and the National Center for Health Statistics. Birth data is provided by the National Center for Health Statistics (NCHS) and state vital records offices
- **Raw data:** the number of monthly births in each US county in each month from 1985 - 2015
- **Aggregated data:** the number of : the number of monthly births in the US in each month from 1985 – 2015
- **Data size:** 372 records
- **Programming language:** R

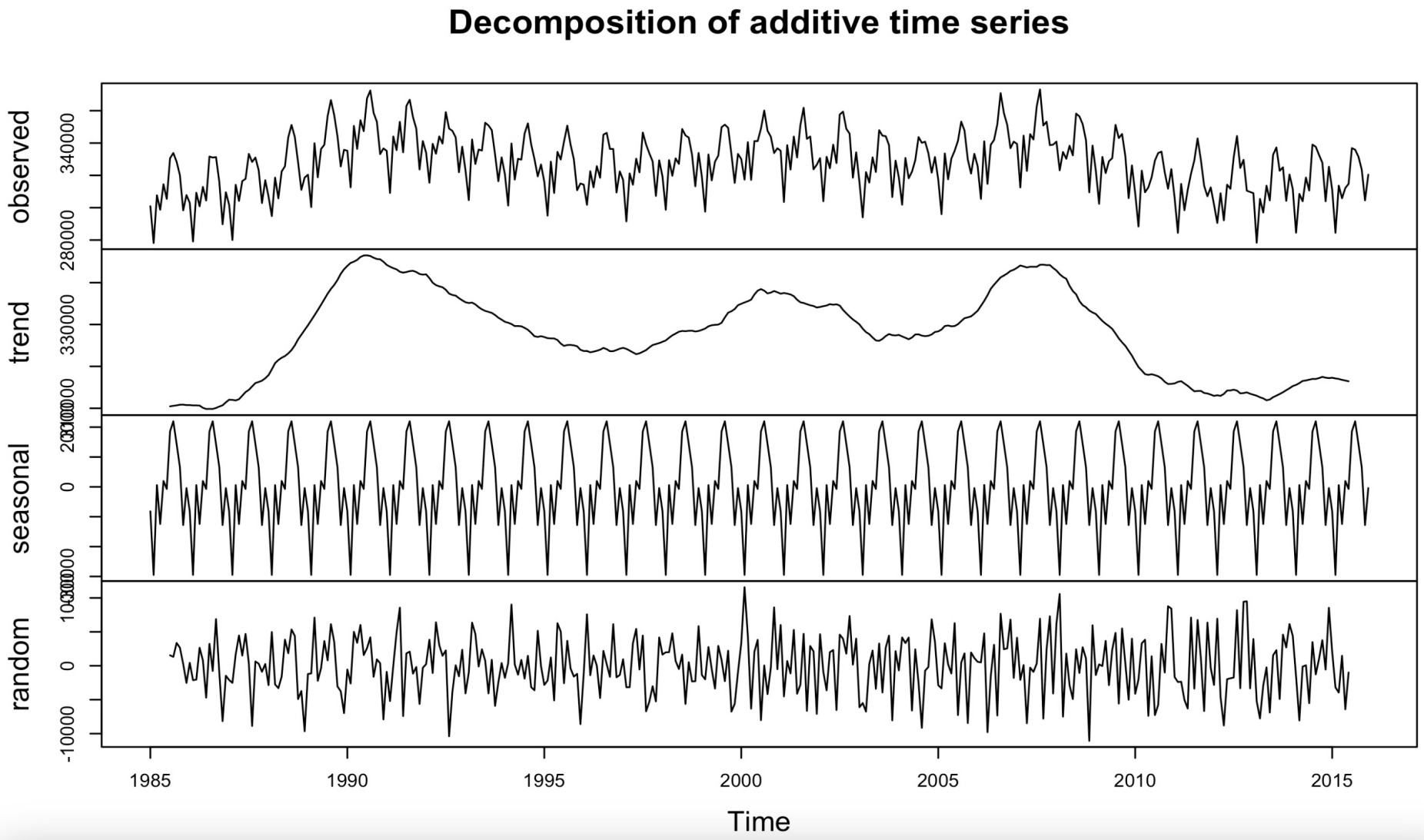
State	Month	Year	countyBirths	stateBirths	County
1	1	1985	36.0	5027	1001.0
1	2	1985	36.0	4627	1001.0
1	3	1985	43.0	4738	1001.0
1	4	1985	40.0	4626	1001.0
1	5	1985	34.0	4834	1001.0
1	6	1985	42.0	4755	1001.0
1	7	1985	40.0	5255	1001.0
1	8	1985	55.0	5364	1001.0
1	9	1985	37.0	5356	1001.0
1	10	1985	38.0	5144	1001.0
1	11	1985	30.0	4883	1001.0
1	12	1985	48.0	5127	1001.0
1	1	1985	127.0	5027	1003.0
1	2	1985	96.0	4627	1003.0
1	3	1985	107.0	4738	1003.0
1	4	1985	100.0	4626	1003.0
1	5	1985	107.0	4834	1003.0

2. Exploratory Data Analysis



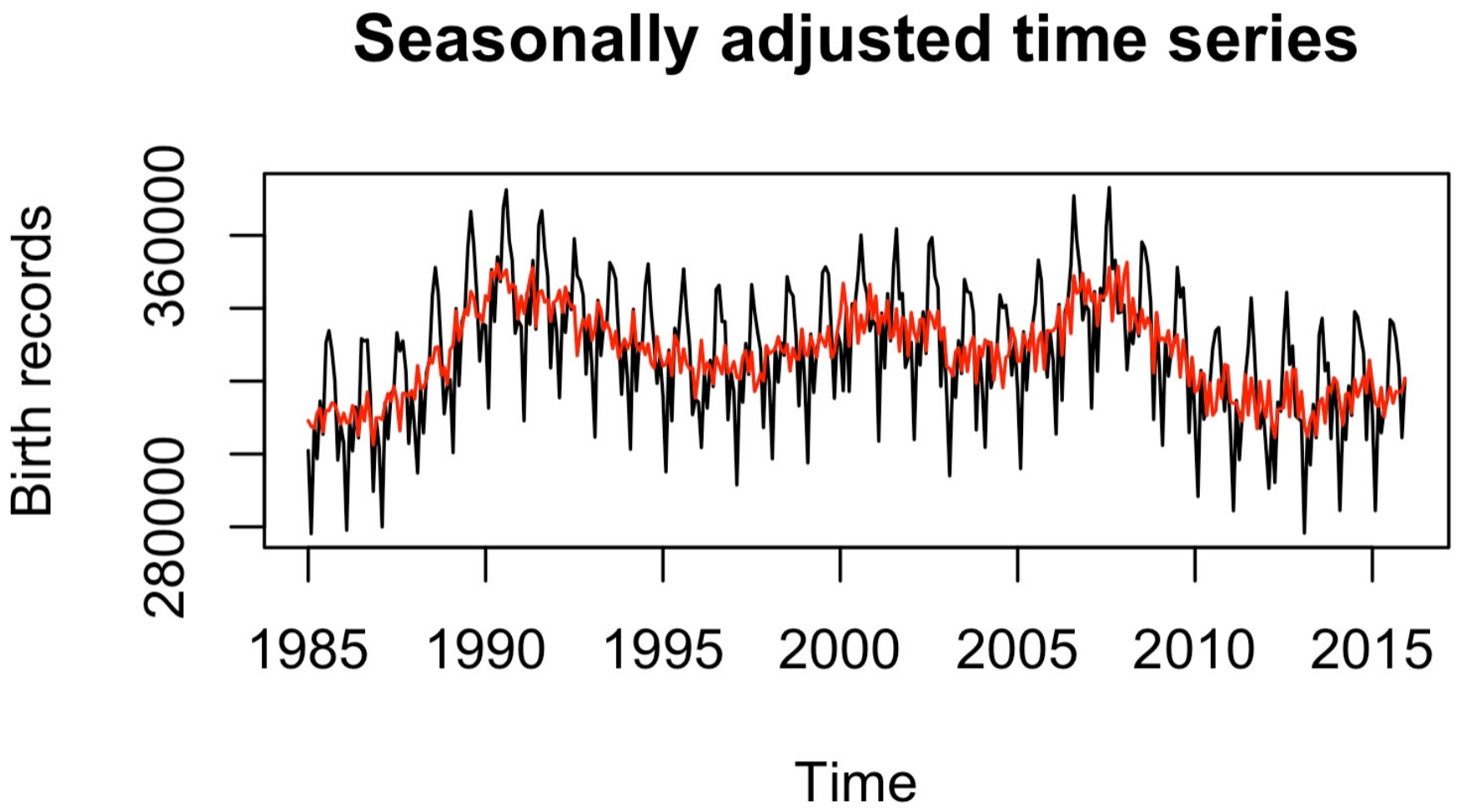
- The time series shows that the birth data has a **seasonal pattern** and there is an **underlying fluctuating trend.**

2. Exploratory Data Analysis



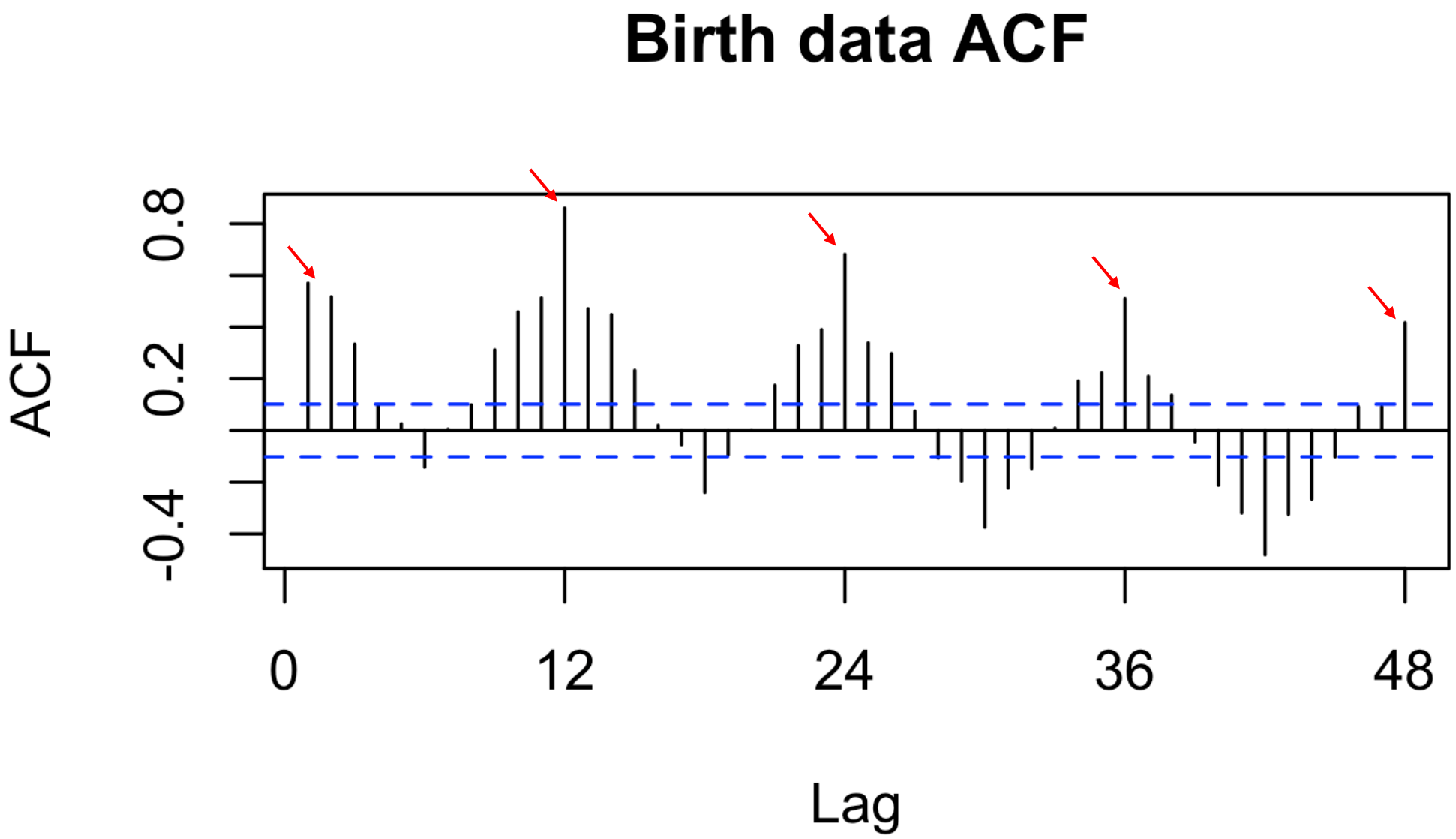
- **Type:** additive
- **Trend:** the time series has an underlying fluctuating trend
- **Seasonality:** Months with the most births: July & August, months with the least births: February & November

2. Exploratory Data Analysis



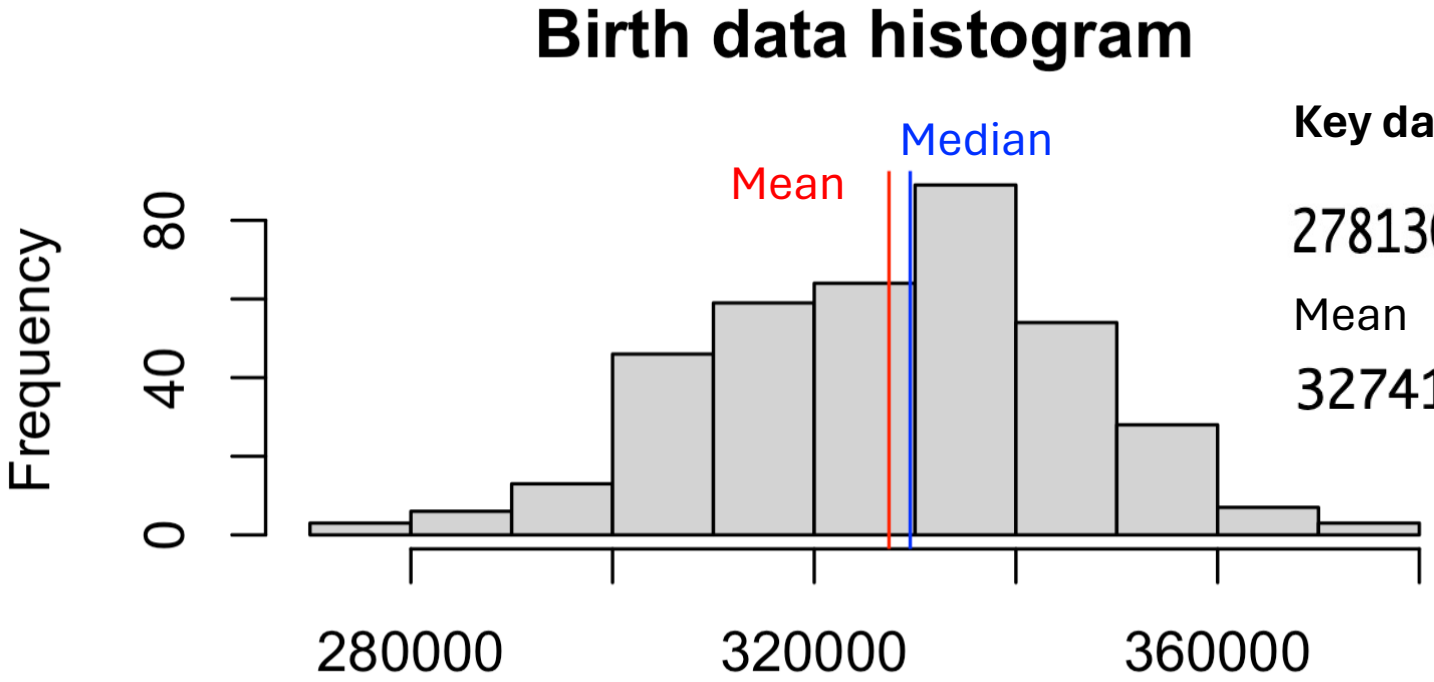
- The time series is heavily affected by its seasonal component.

2. Exploratory Data Analysis



- A high correlation is observed every **12 lags**.
- Clustered and repetitive ACF values: birth records in a month can be affected by adjacent previous and following months

2. Exploratory Data Analysis

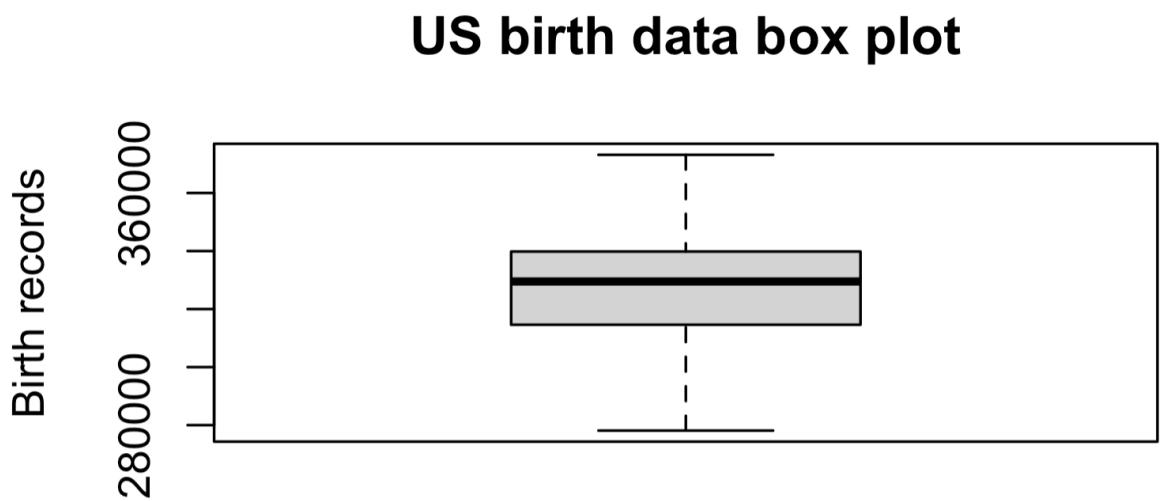


Key data summary: Min-Q1-Median-Q3-Max

278130.0 314605.5 329517.0 339823.0 373161.0

Mean

327413.5



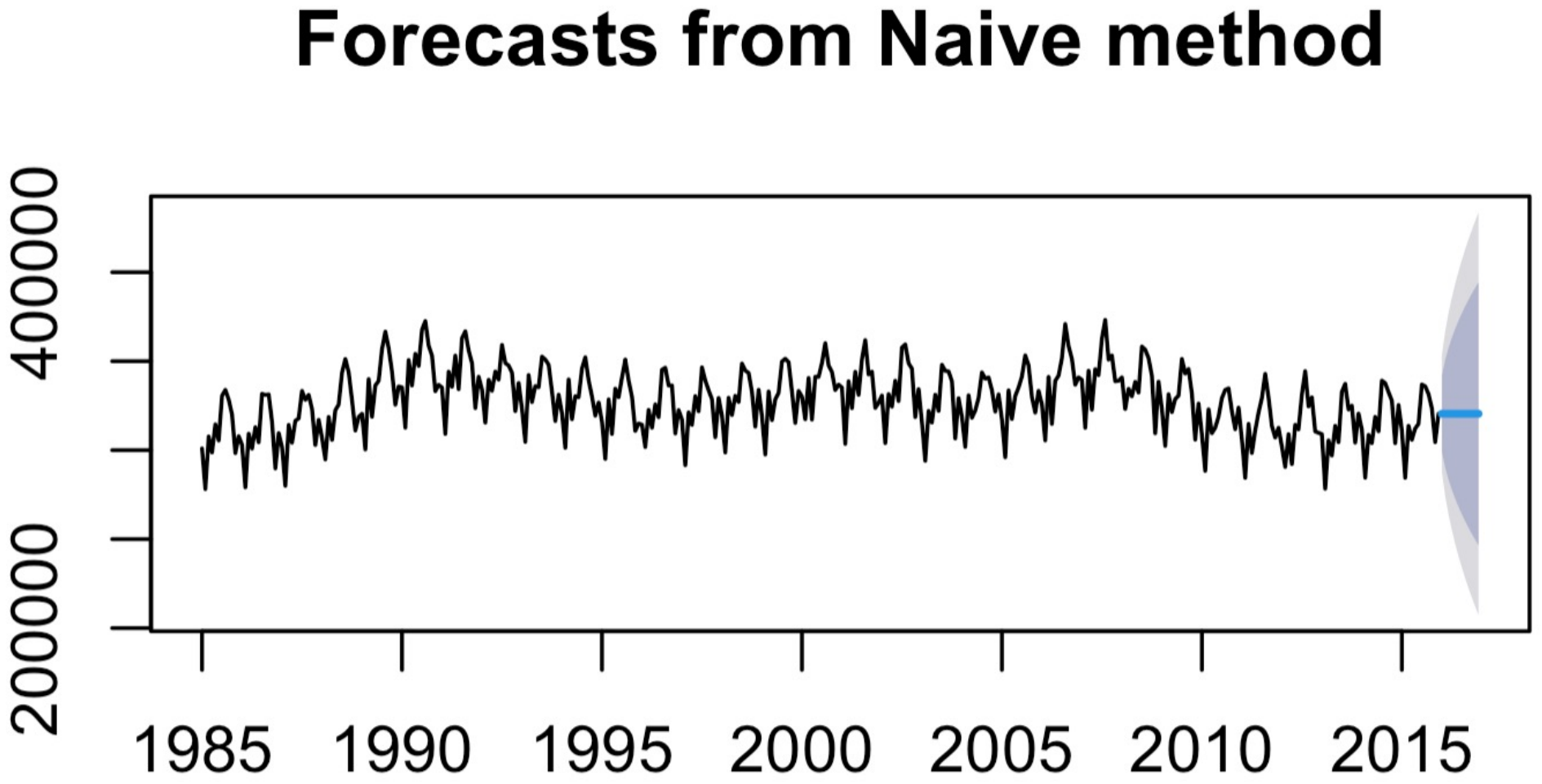
- On average, the US has around **330,000** babies each year. The annual birth rate mostly ranges from **310,000 to 370,000**.
- There are no outliers detected.

3. Forecasting question & evaluation criteria

- The exploratory analysis suggests the time series has a seasonal and an underlying fluctuating trend.
- The level shifts every 5 years.
- **Forecast question:** ***What are the projected birth records for the following 5 years?***
- **Application:** The forecast can be used by the US government to rectify their policies in family support, maternal healthcare, education, etc.
- **Forecast models:** *(1) Naïve (2) Mean (3) Simple Moving Average (MA) (4) Simple Exponential Smoothing (SES) (5) Holt-Winters (6) ARIMA (7) Decomposition (8) Simple Linear Regression (SLR)*
- **Evaluation criteria:** ***(1) Mean Absolute Percentage Error (MAPE) (2) Residual validation***. The estimated percentage deviation can help the US government plan their budget, resources, and facilities.
- The entire dataset will be used for forecasting since the time series does not change its patterns between 1985 and 2015.

4. Forecasting models

Naïve model: forecasts

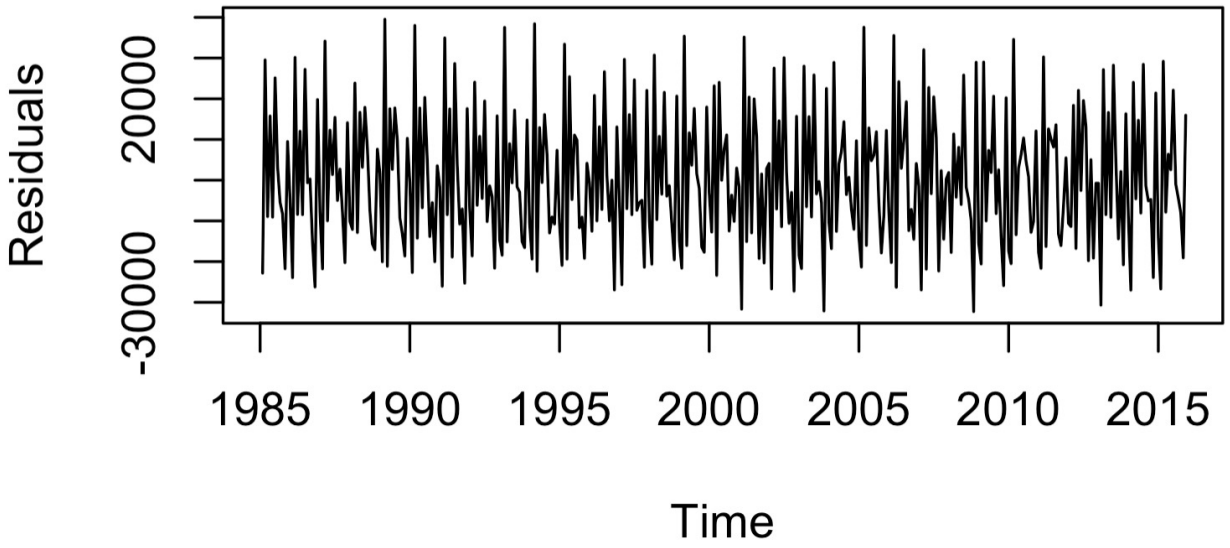


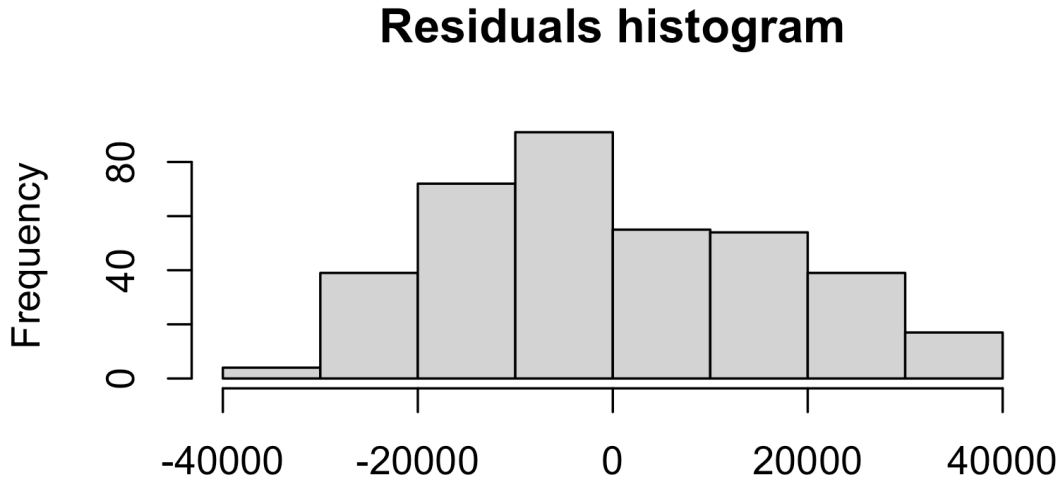
- The value for the last period: December 2015, will be used as future forecasts.

4. Forecasting models

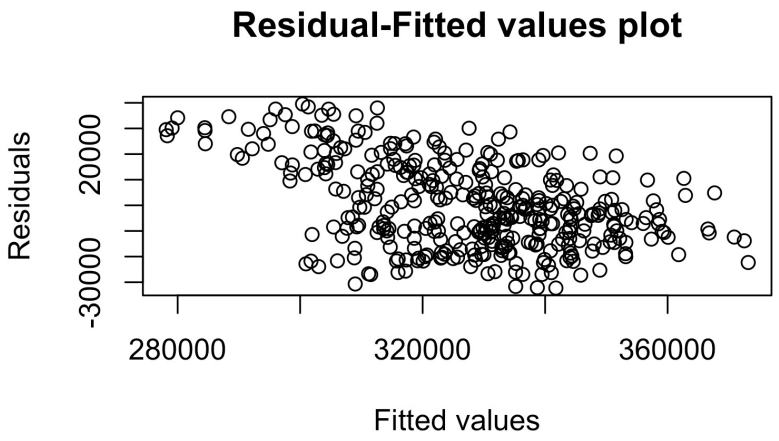
The naïve model cannot capture the seasonal component of the time series.

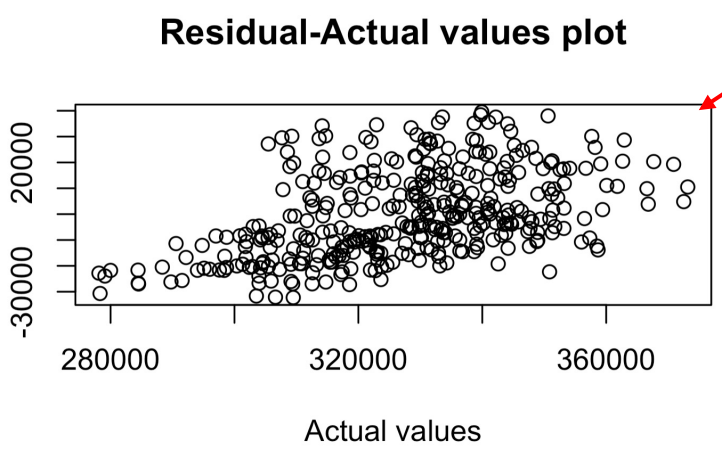
Naïve model: residual analysis

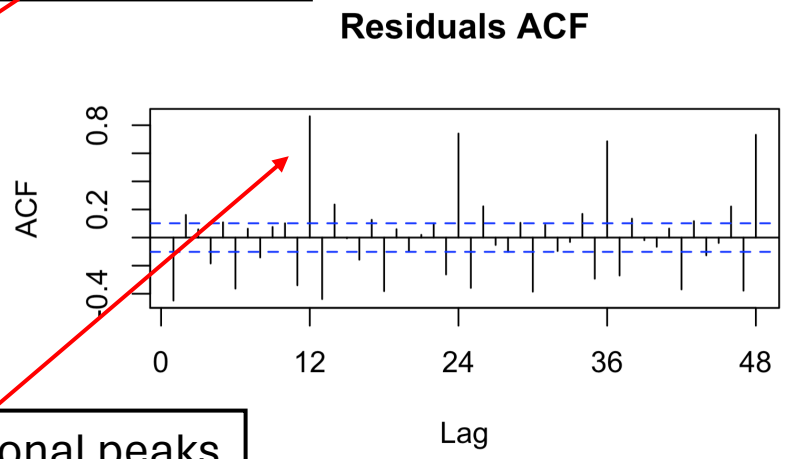




Errors are higher for larger records



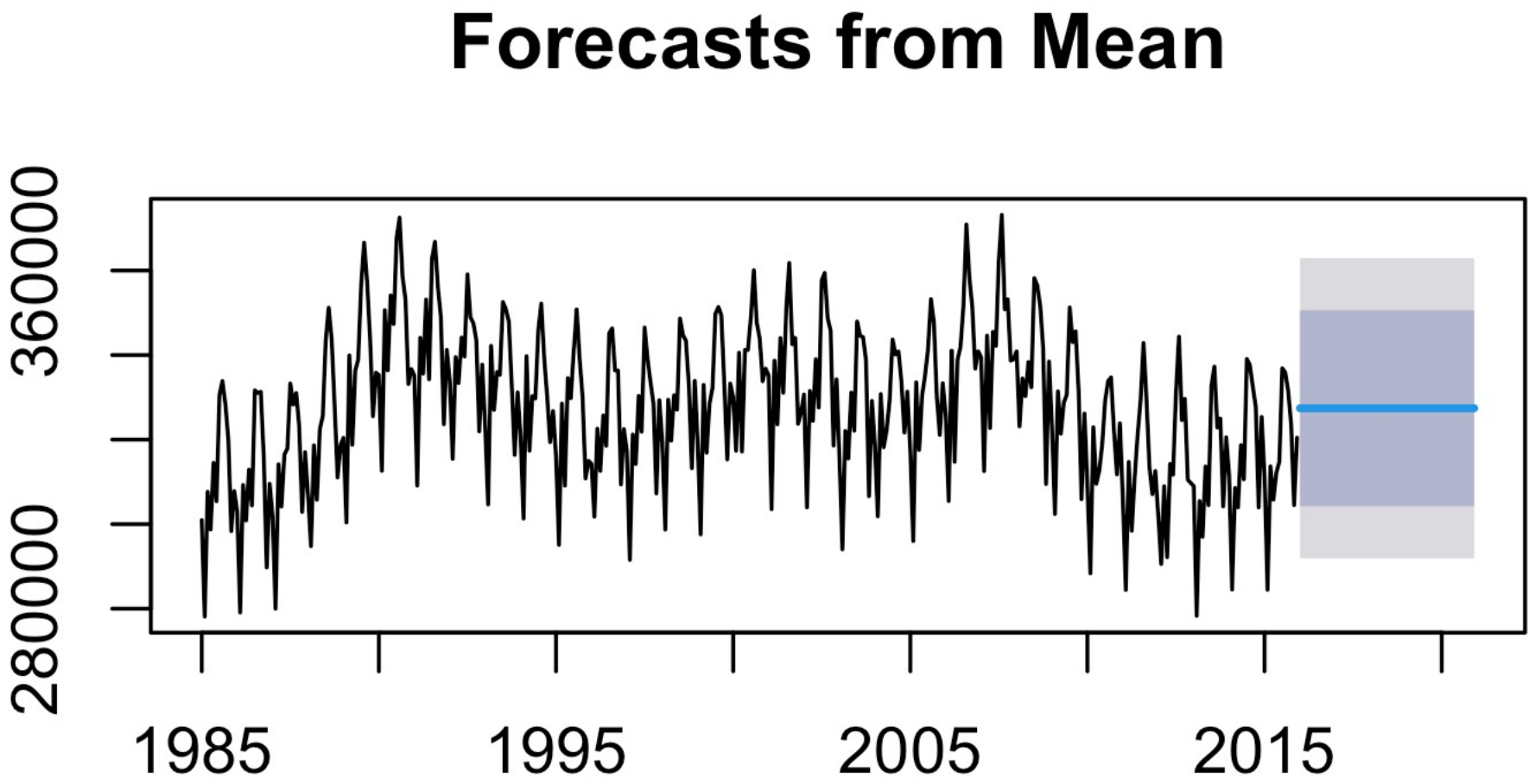




Seasonal peaks

4. Forecasting models

Mean model: forecasts

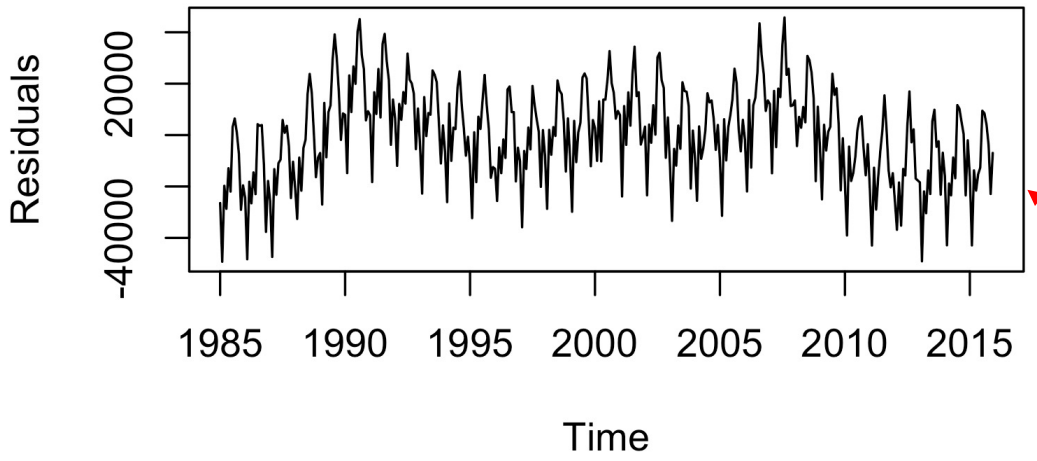


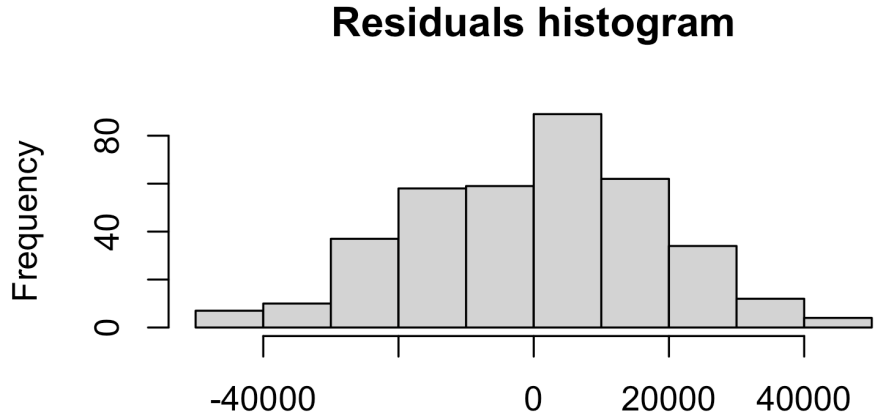
- The mean birth records (327,413.5) will be used as future forecasts.

4. Forecasting models

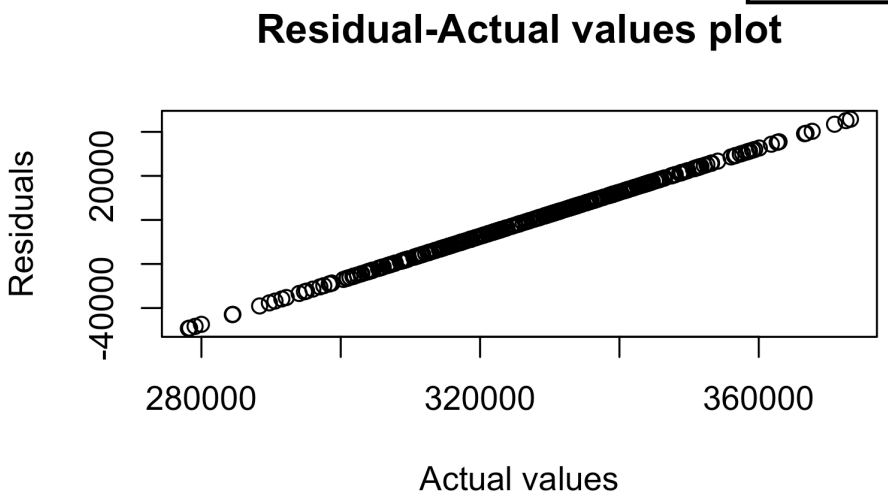
The mean model, similarly, cannot capture most patterns of the time series.

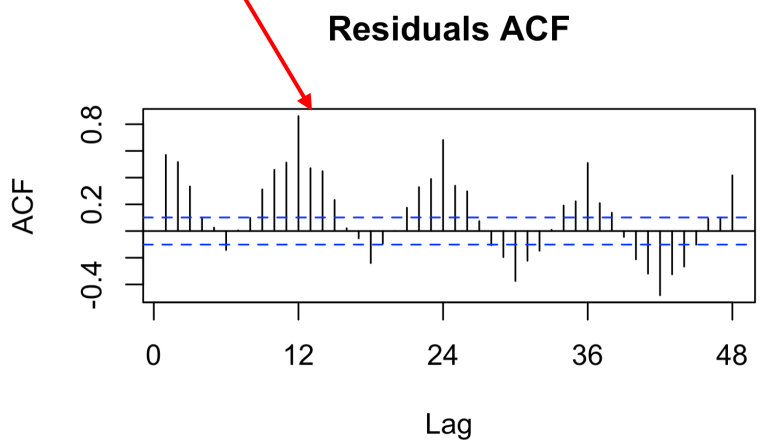
Mean model: residual analysis





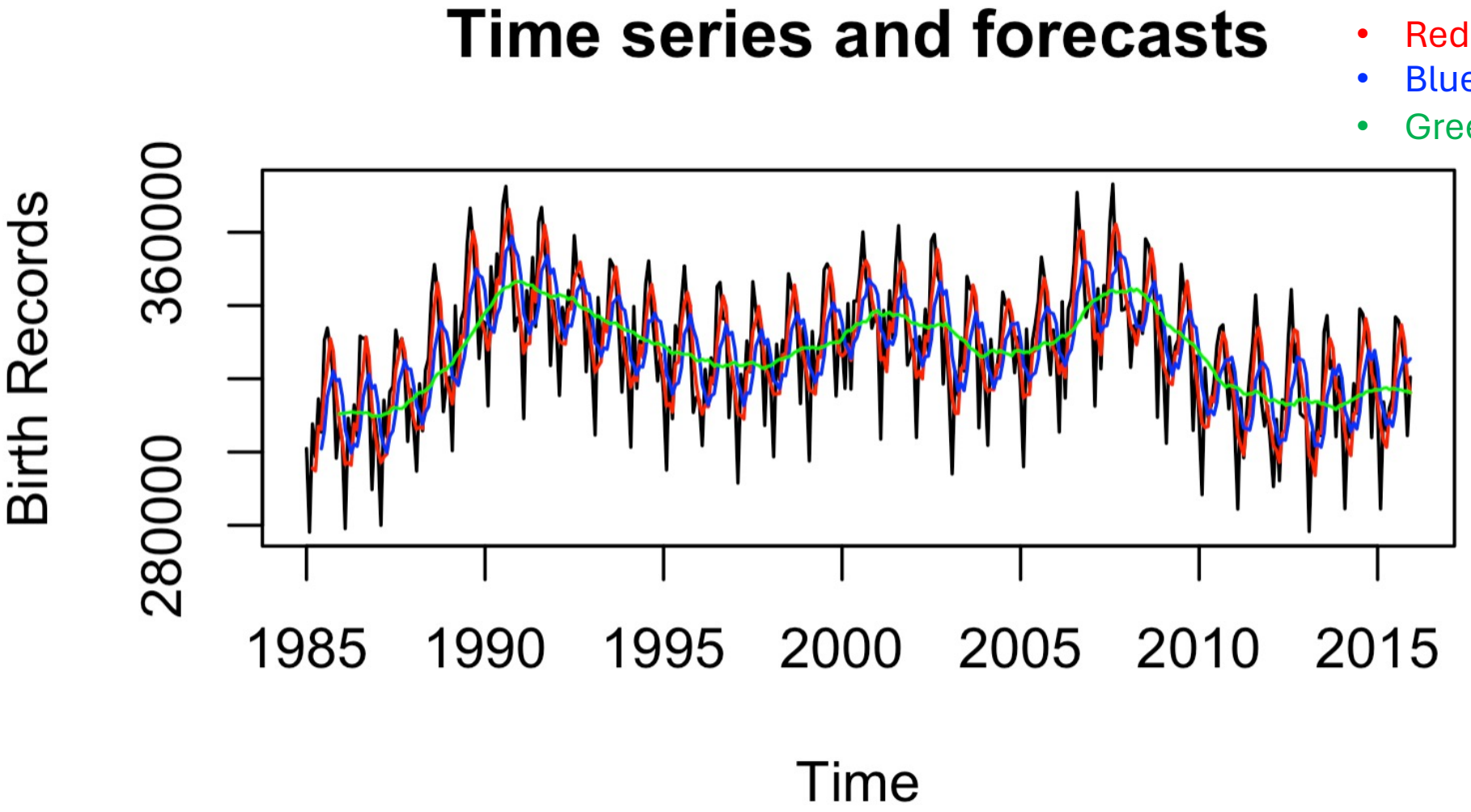
Both residuals and ACF plot are similar to those of the original time series





4. Forecasting models

MA(3), **MA(6)**, and **MA(12)** models



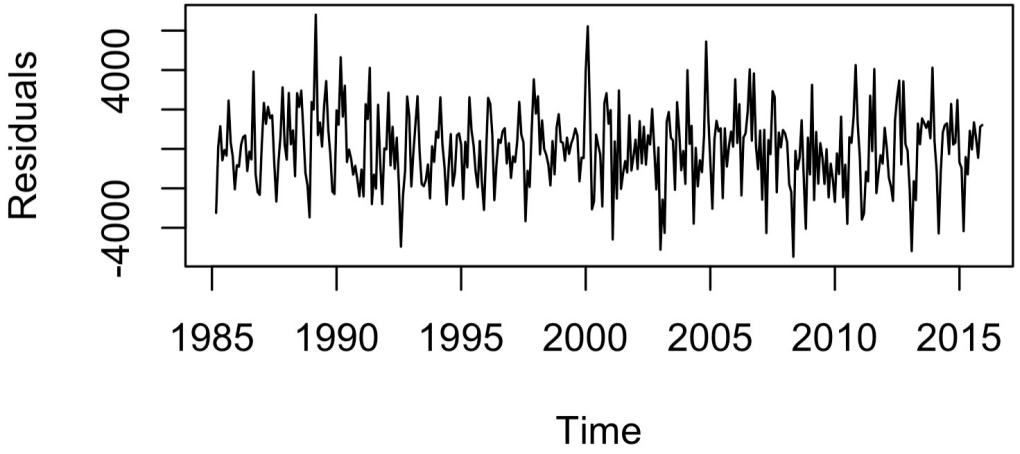
- Red: MA(3)
- Blue: MA(6)
- Green: MA(12)

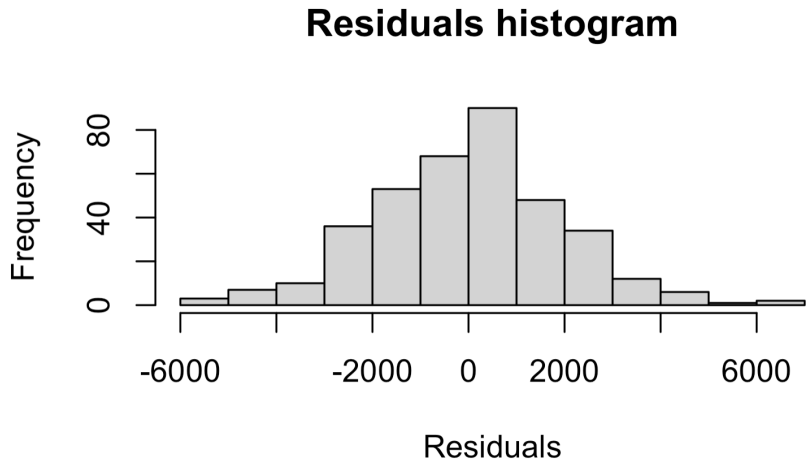
- The data is smoothed out and deviates more from the actual values as the forecasting windows increase.

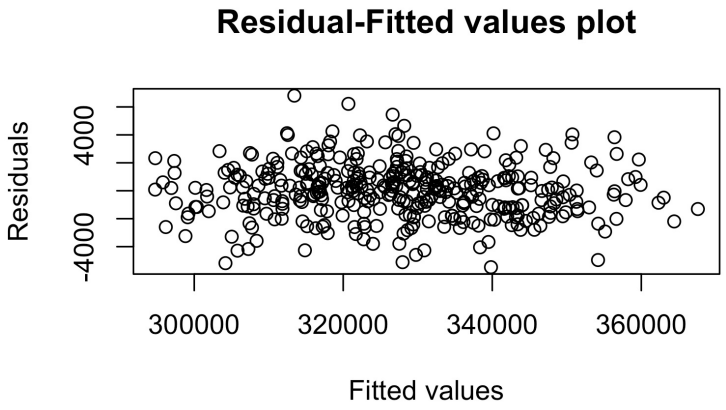
4. Forecasting models

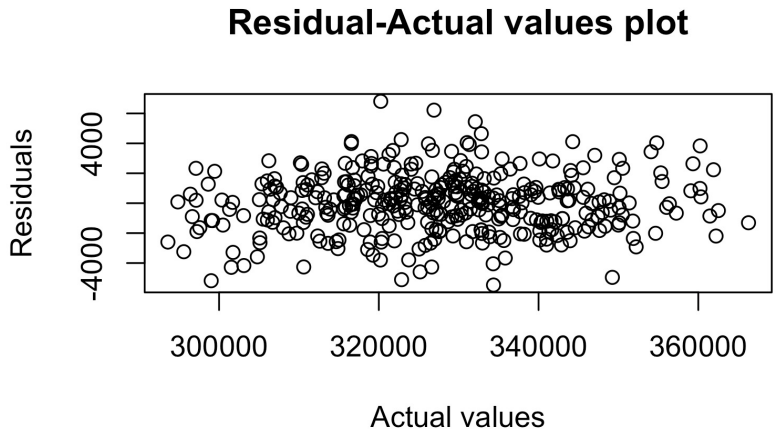
The residuals for MA(3) appear to be more random, with no seasonality and trend detected. The error range is much smaller.

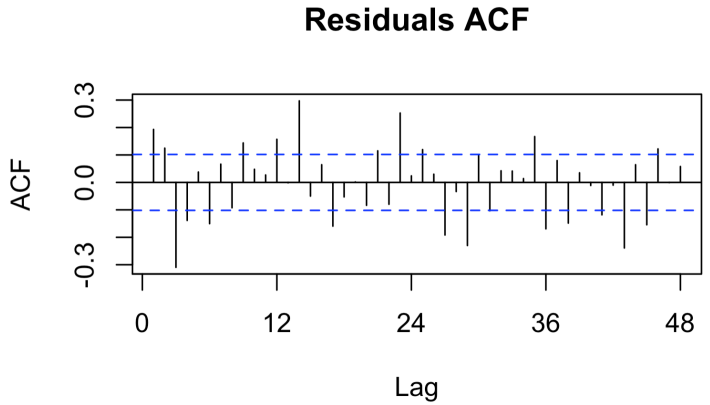
MA(3) model: residual analysis





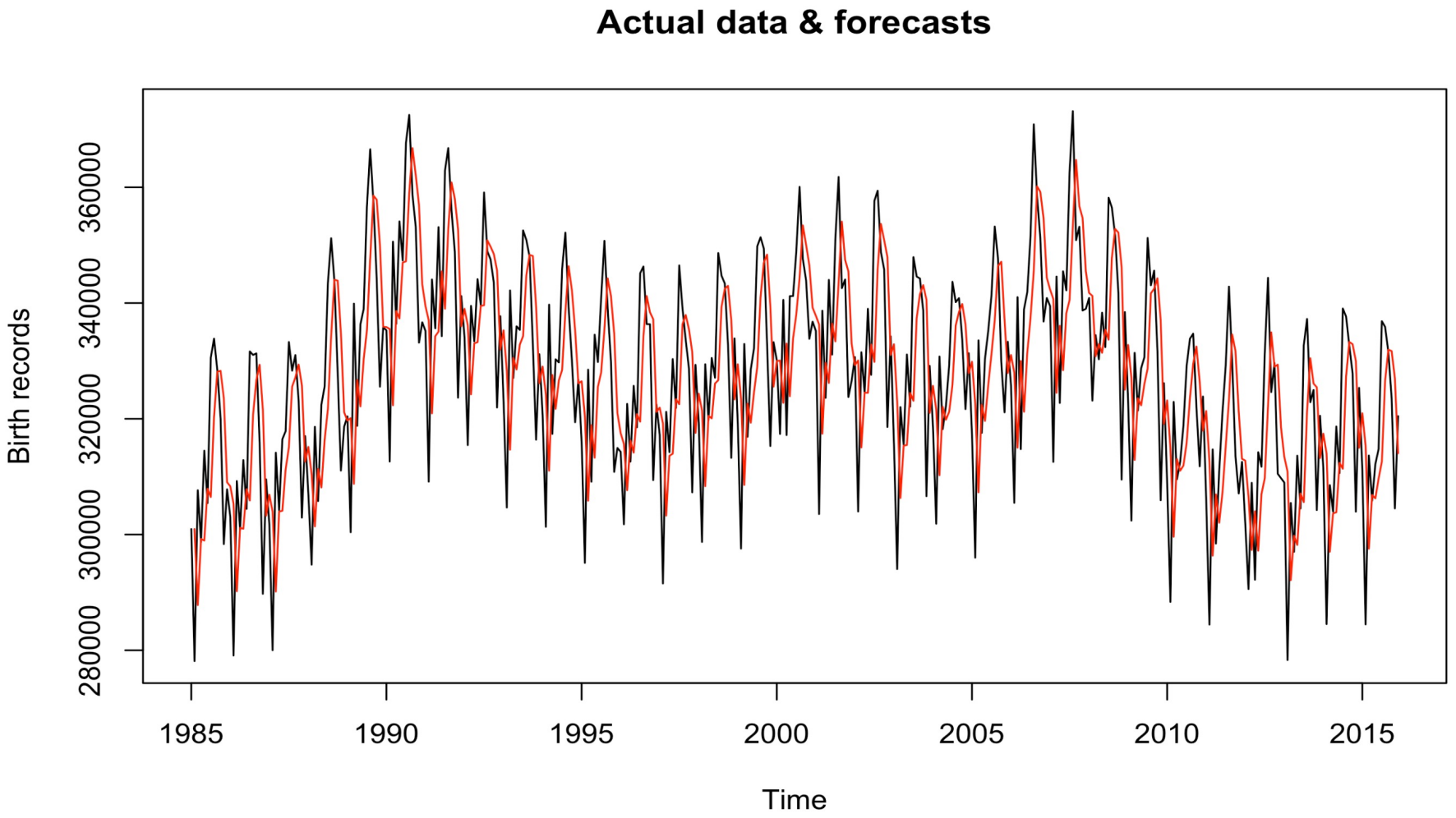






4. Forecasting models

Simple exponential smoothing (SES) model: forecasts

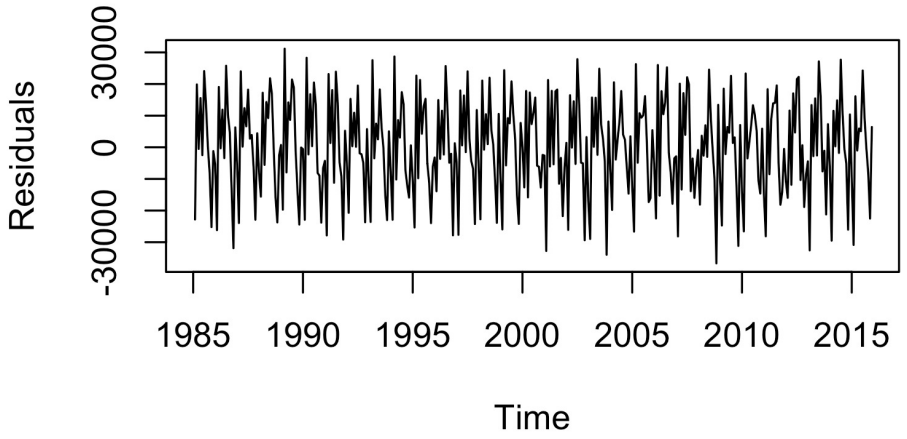


- The plot suggests the SES model tends to underestimate high values and overestimate low values.

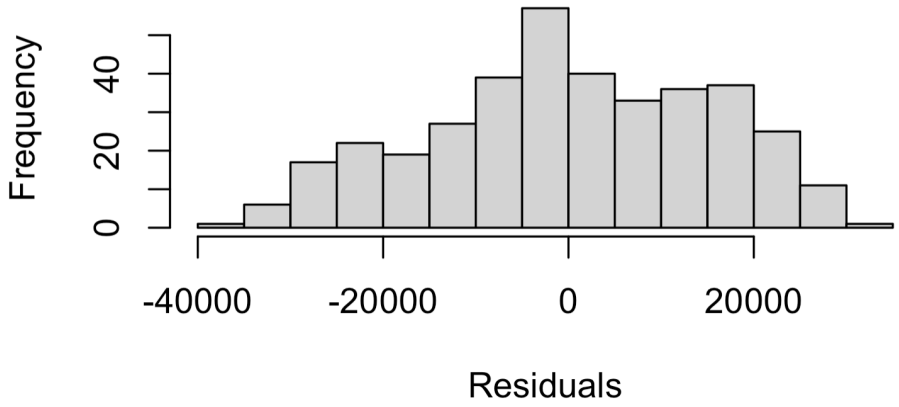
4. Forecasting models

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.

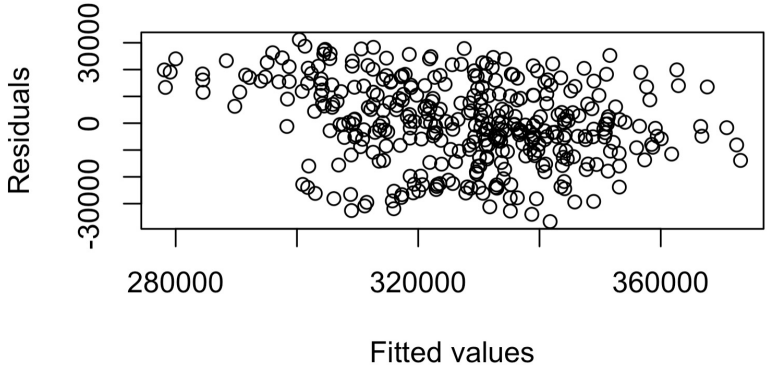
SES model: residual analysis



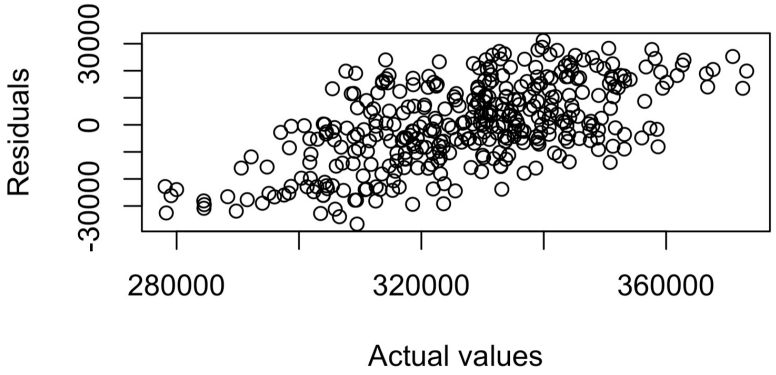
Residuals histogram



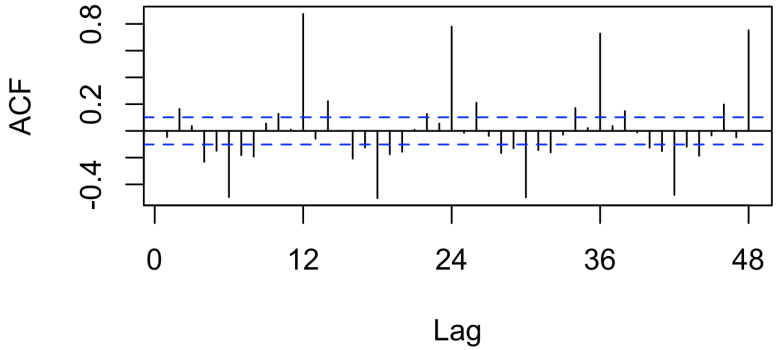
Residual-Fitted values plot



Residual-Actual values plot



Residuals ACF

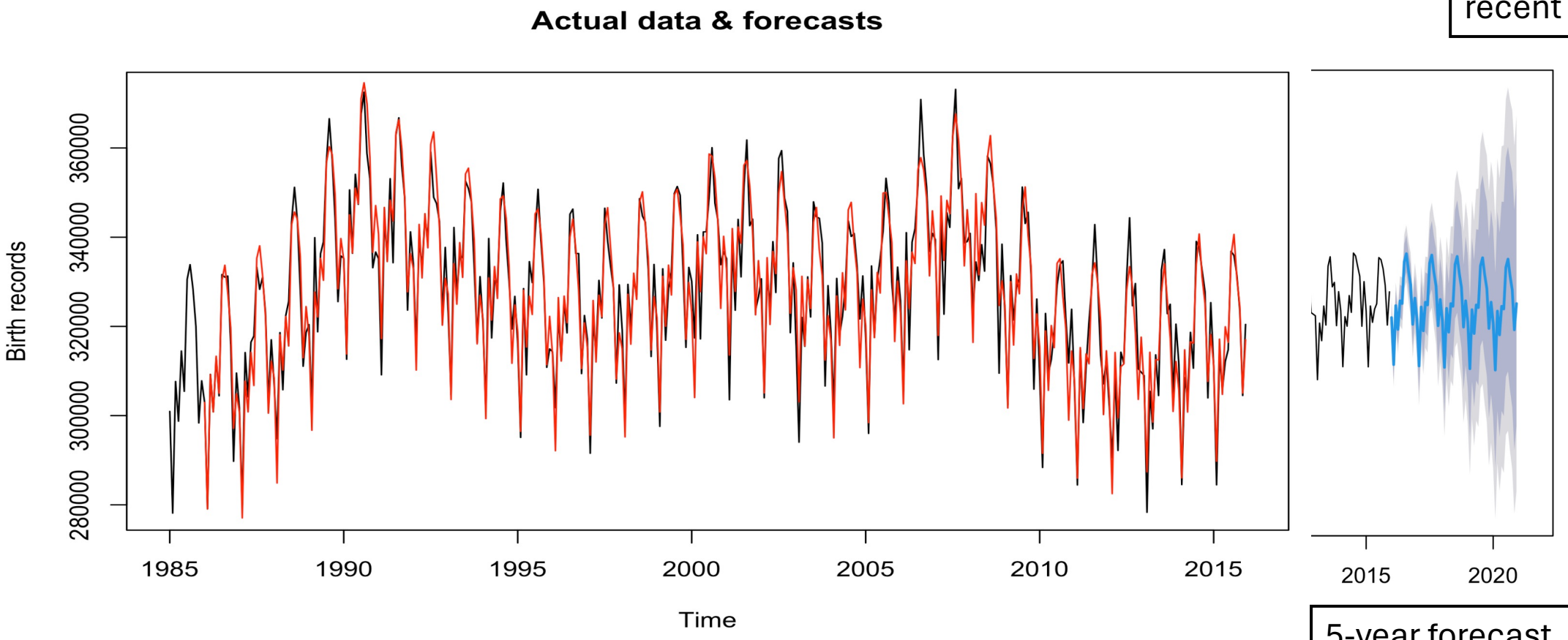


4. Forecasting models

Holt-Winters (HW) model: forecasts

- The HW model improves its performance after the data point 1990.

A higher weight is given to the most recent point.



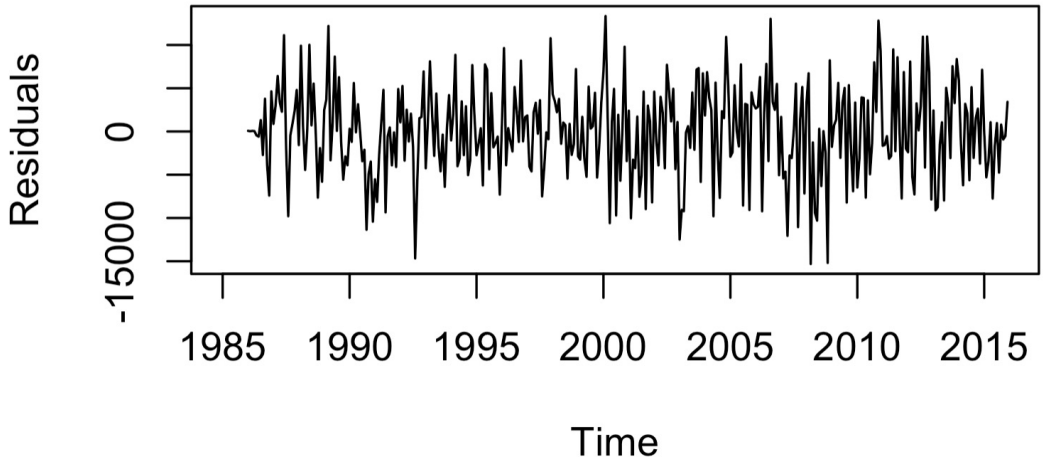

```
alpha
0.2066015
beta
0.1299084
gamma
0.1355967
```

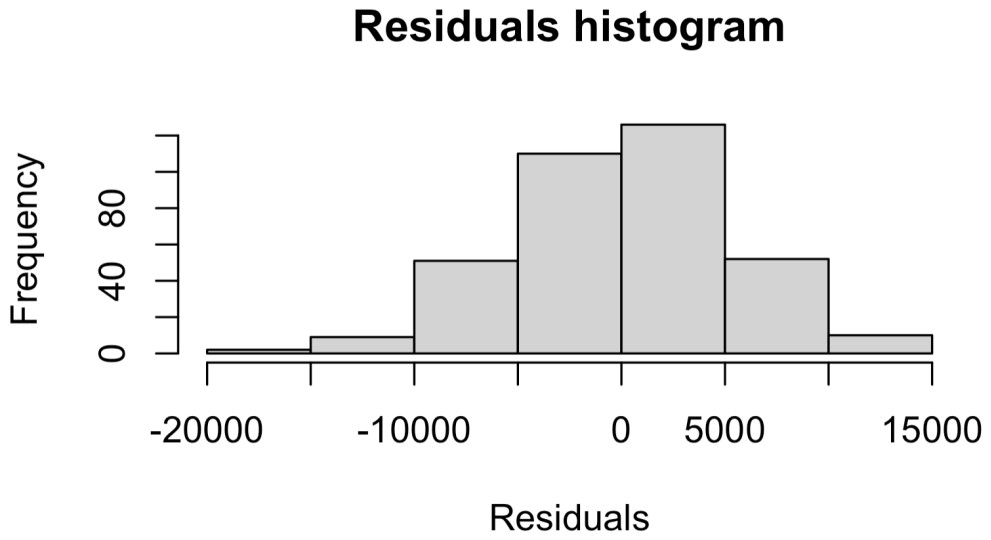
5-year forecast

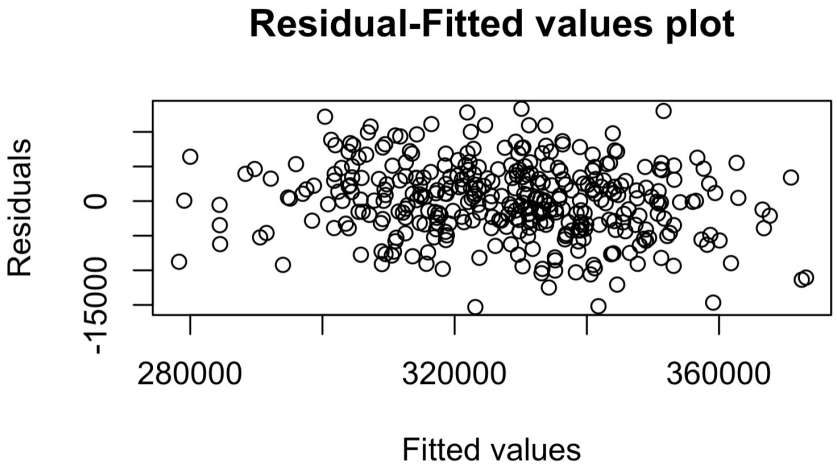
4. Forecasting models

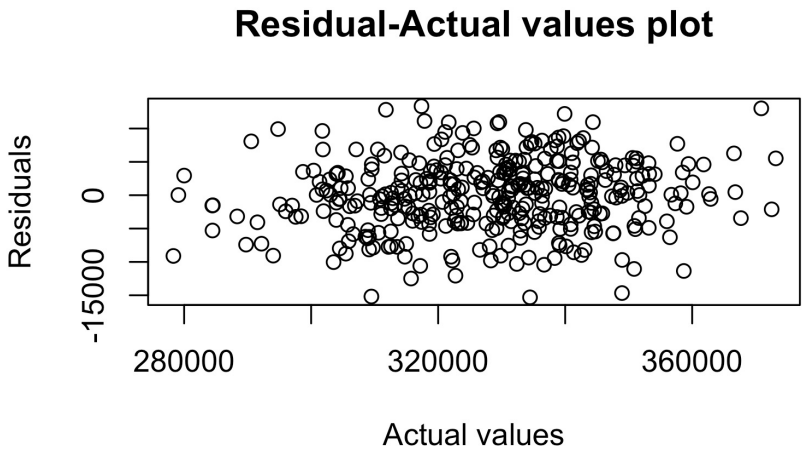
The residuals for HW appear to be random, with no seasonality and trend detected.

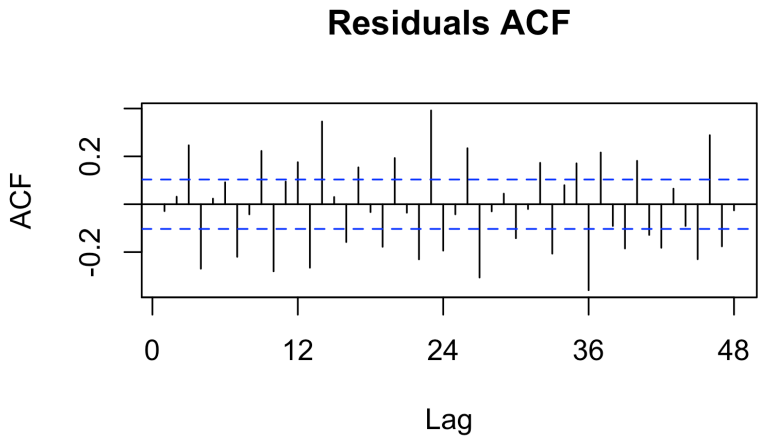
HW model: residual analysis







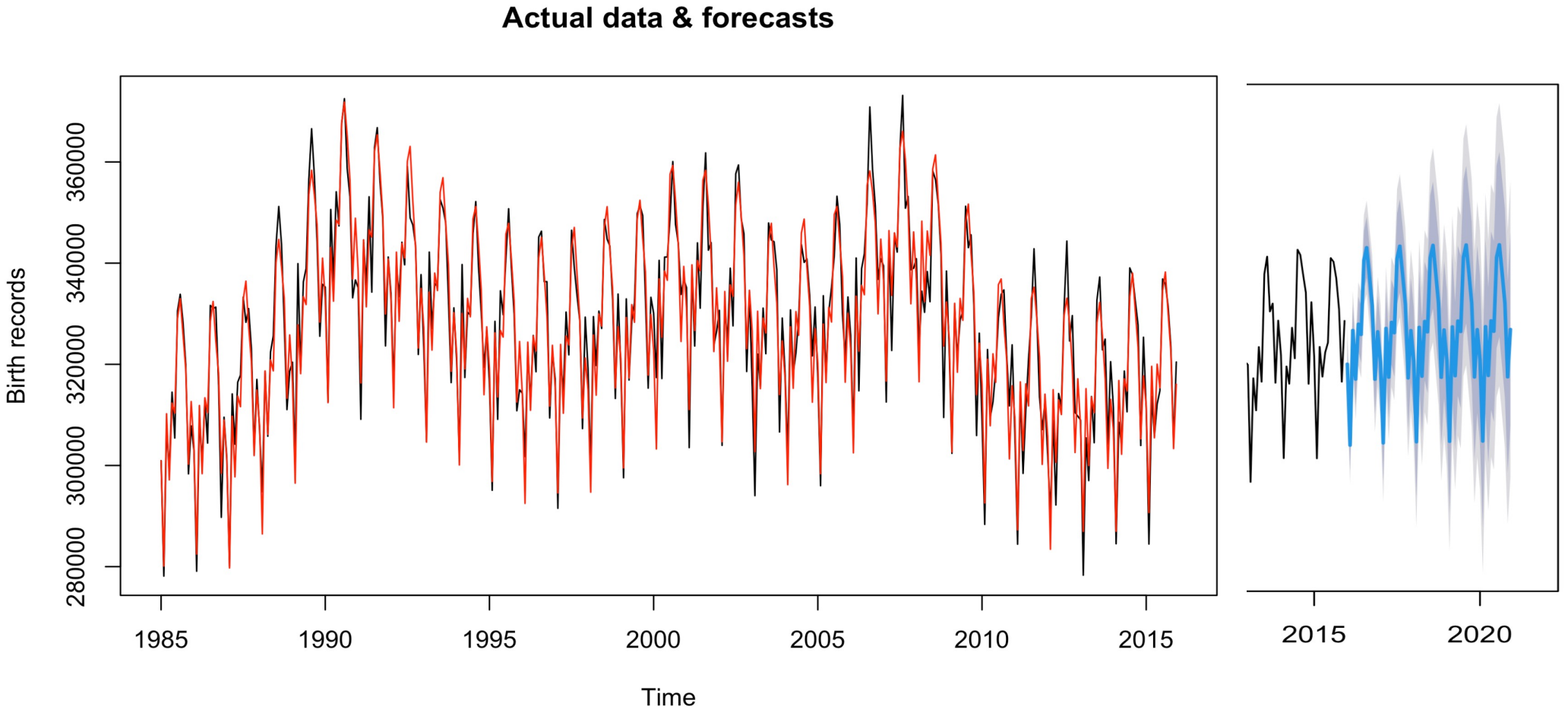




4. Forecasting models

Decomposition model: forecasts

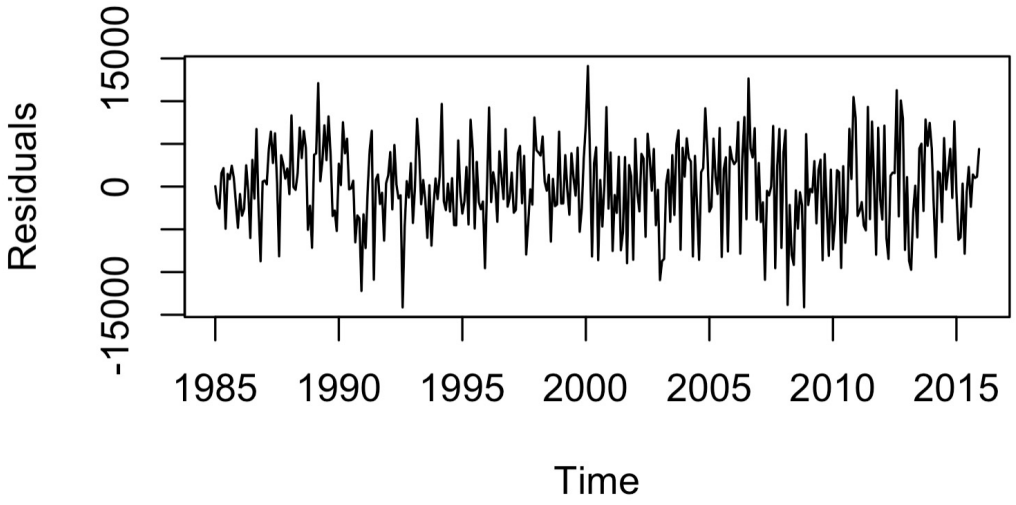
- Similar to HW, the decomposition model performs better as the time period moves further.

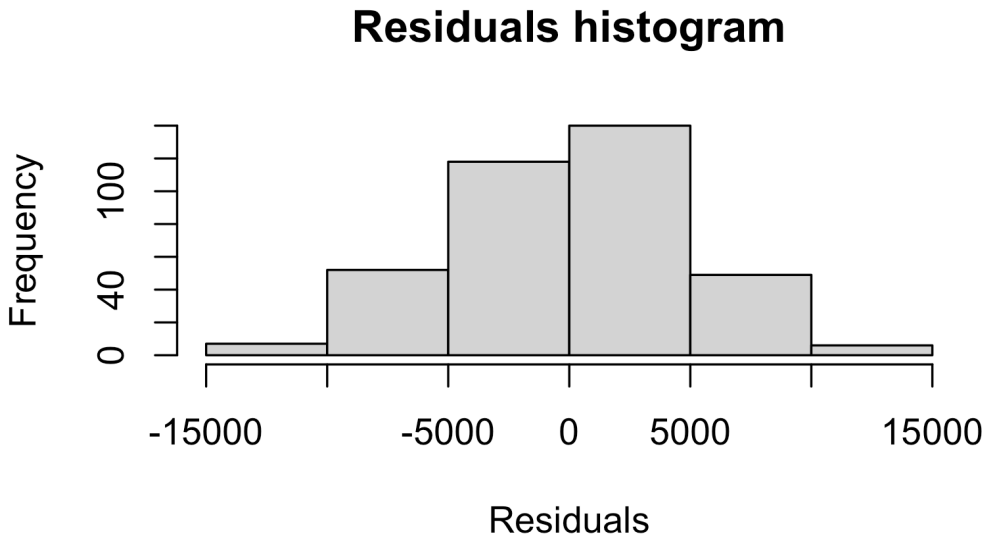


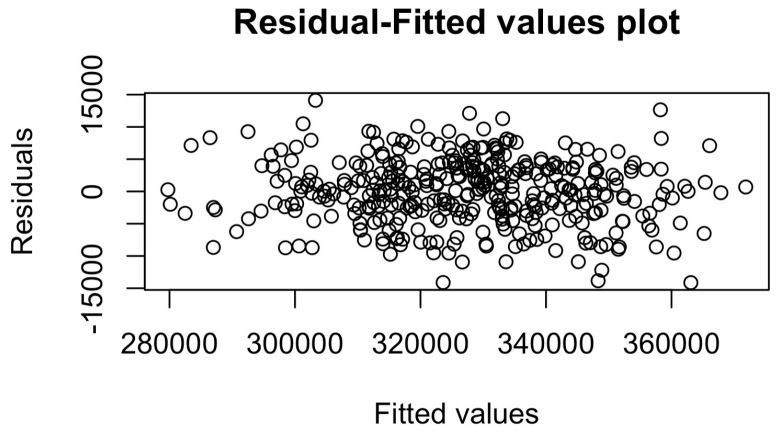
4. Forecasting models

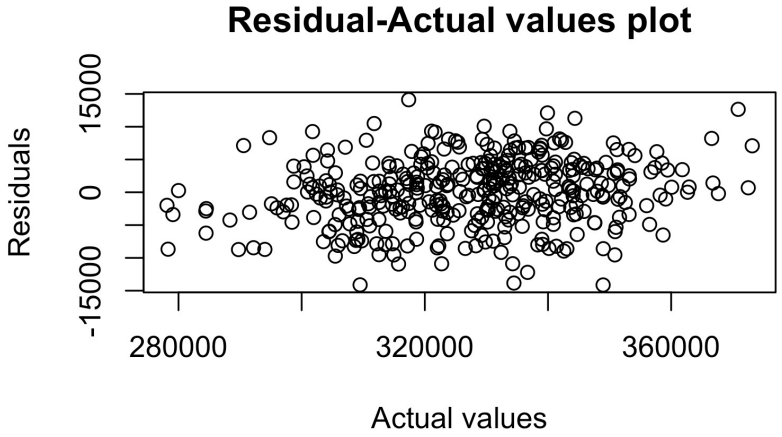
The residuals for Decomposition model appear to be random, with no seasonality and trend detected.

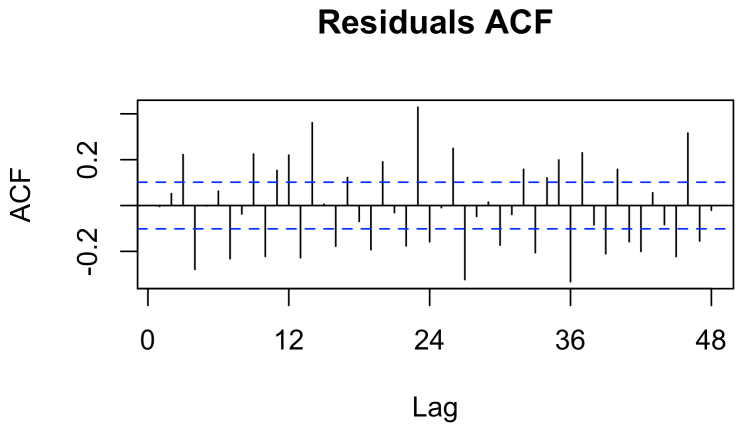
Decomposition model: residual analysis







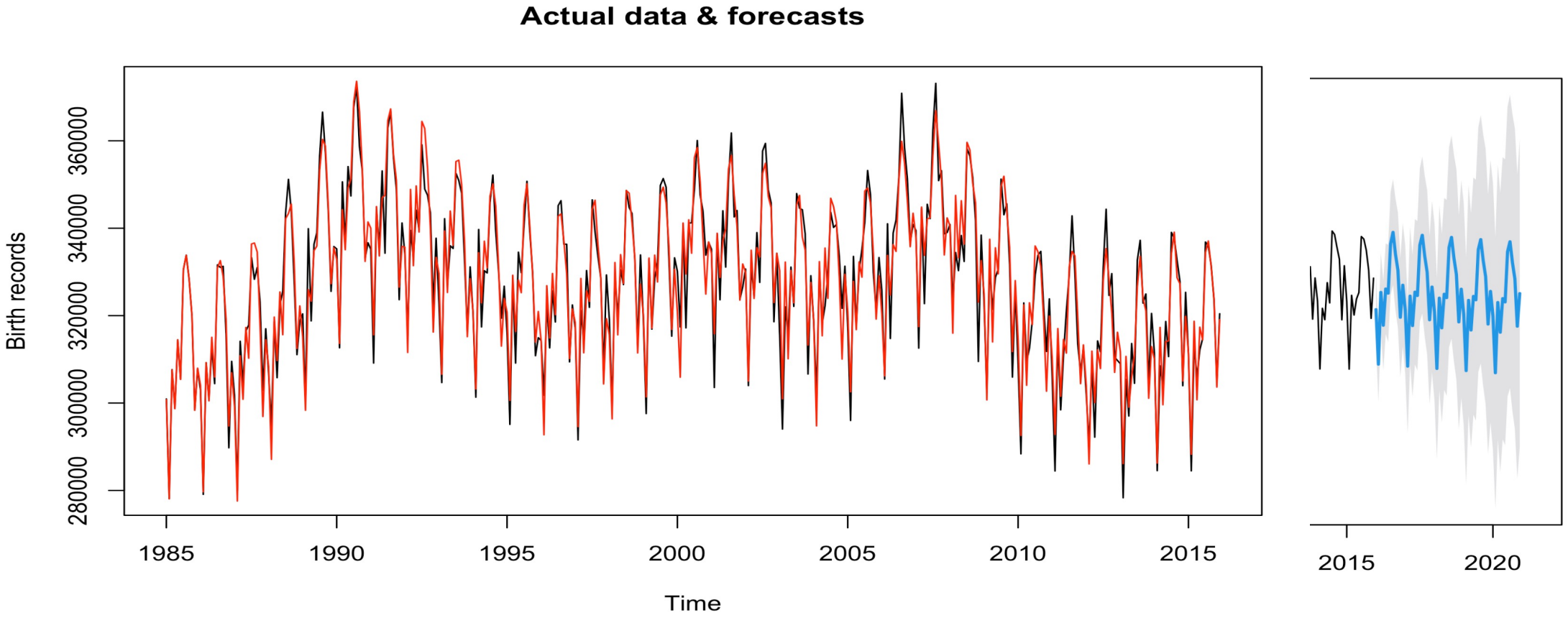




4. Forecasting models

ARIMA model: forecasts

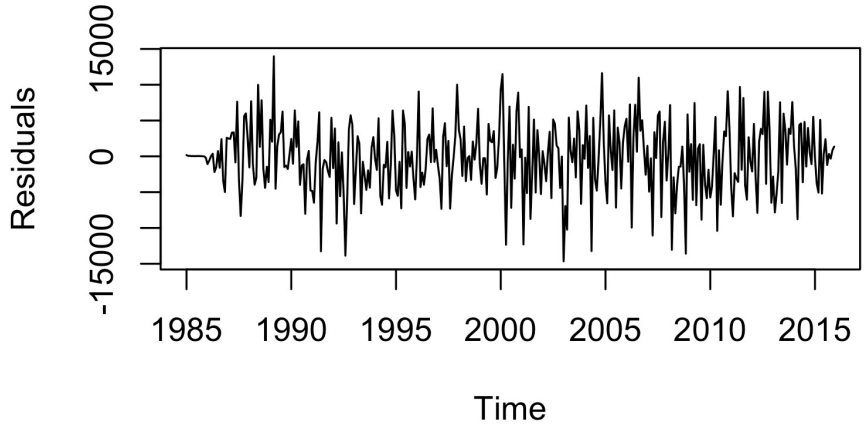
- **Selected model:** ARIMA(4,1,0)(0,1,1)[12]
- Forecastsing results fit the actual data well.

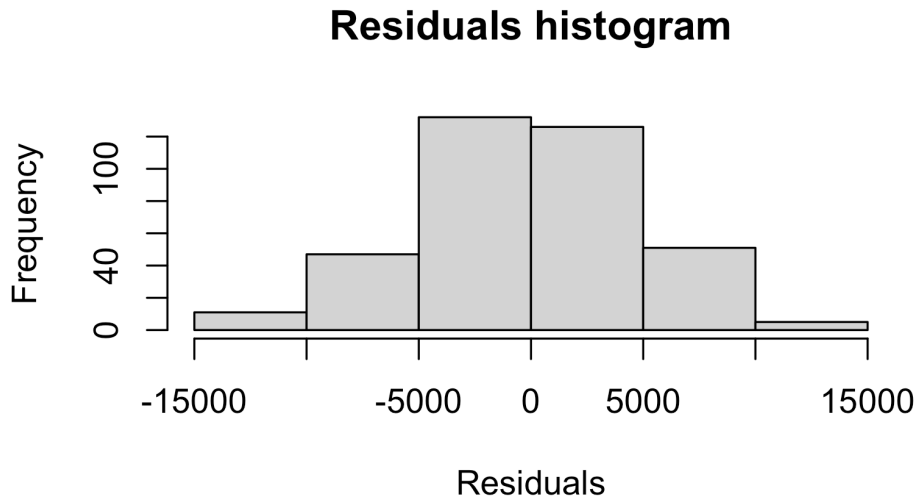


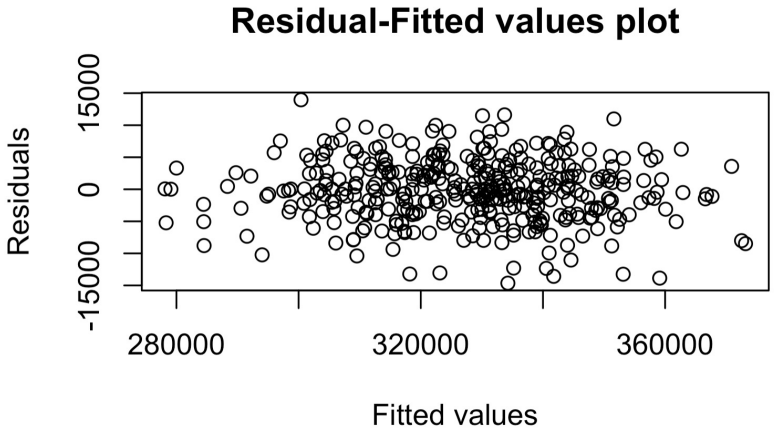
4. Forecasting models

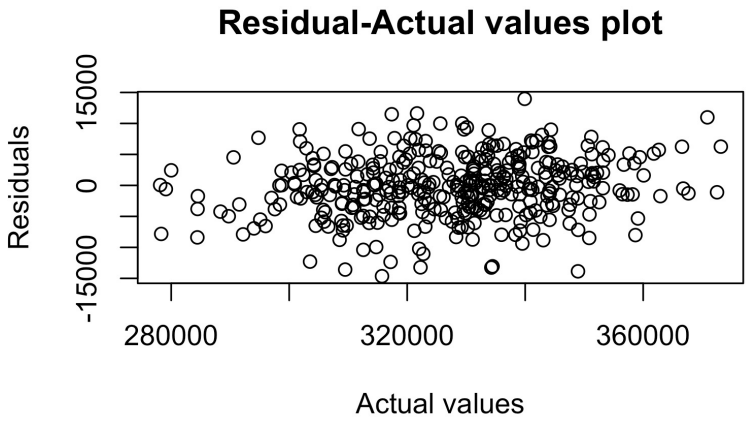
The residuals for ARIMA model appear to be random, with no seasonality and trend detected.

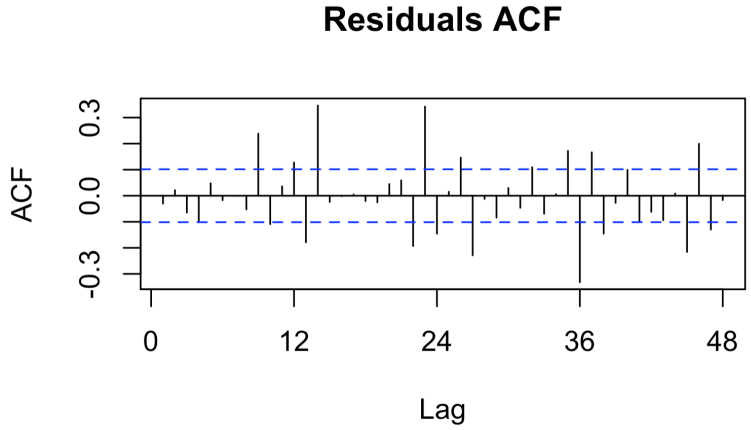
ARIMA model: residual analysis









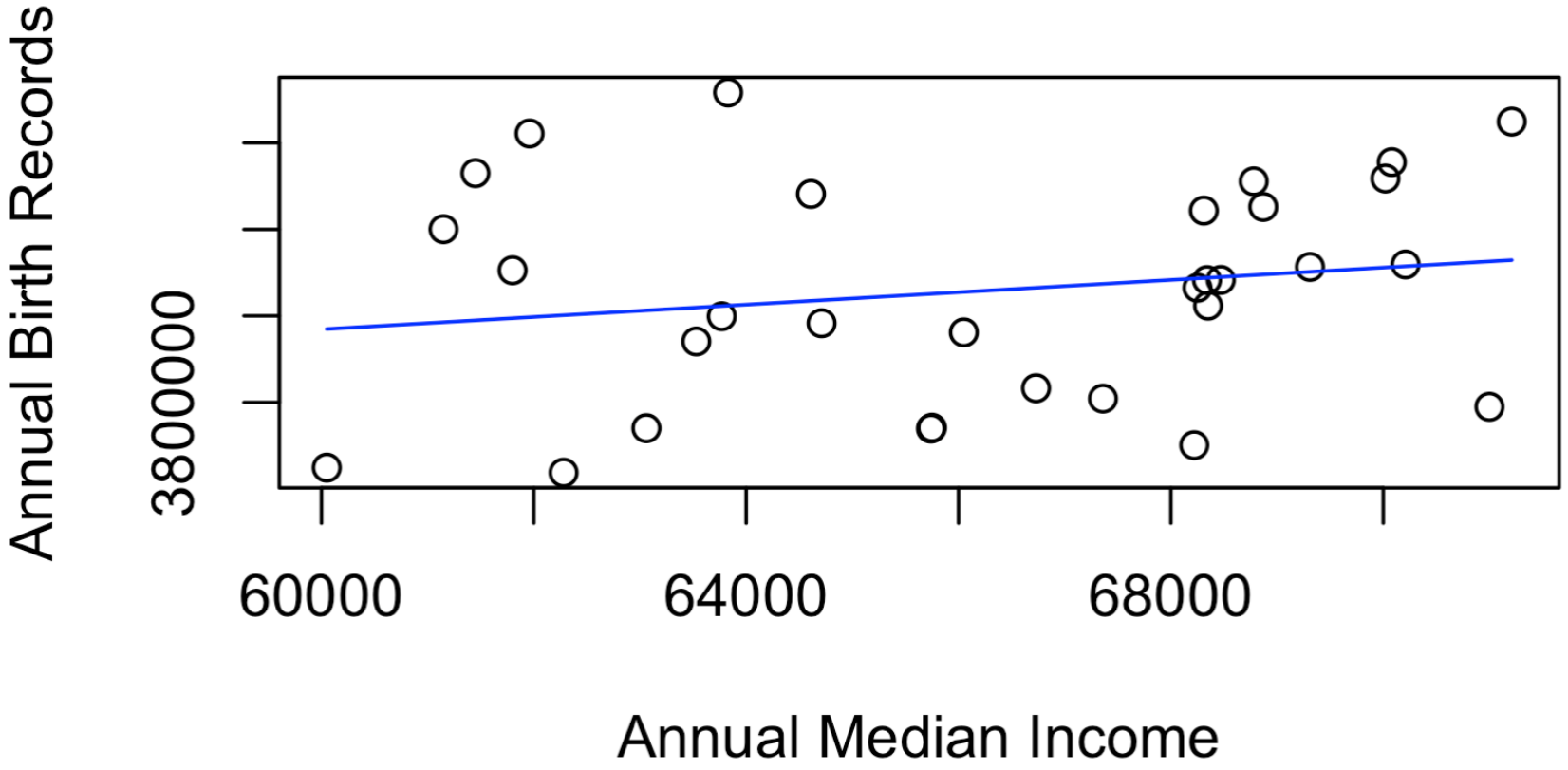


4. Forecasting models

Simple Linear Regression (SLR)

- **Predictor:** Median annual household income from 1985 – 2015.
- **Regression result:**

```
Coefficients:
(Intercept)        inc$MI
  3.456e+06     7.136e+00
```



4. Forecasting models

SLR: Model validation

```
Residuals:
    Min      1Q  Median      3Q     Max
-192650 -124417   -3367  105196  246405

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.456e+06  4.719e+05   7.324 4.56e-08 ***
inc$MI      7.136e+00  7.116e+00   1.003    0.324
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127000 on 29 degrees of freedom
Multiple R-squared:  0.03352,   Adjusted R-squared:  0.0001895
F-statistic: 1.006 on 1 and 29 DF,  p-value: 0.3242
```

- The validation process suggests that using the income can explain only 3% of the variations in birth rates.
- More social indices such as healthcare, education, etc are needed for a multivariable regression model

4. Summary

Forecasting model evaluation

Model	MAPE (%)	Residual analysis
Naive	4.30	Cannot capture seasonality
Mean	4.51	Cannot capture most features
MA(3)	2.49	No seasonality/ trend detected
SES	3.80	Cannot capture seasonality
HW	1.29	No seasonality/ trend detected
Decomposition	1.23	No seasonality/ trend detected
ARIMA	1.16	No seasonality/ trend detected
SLR	Invalid	

- As per the evaluation, the ARIMA model by far provides the best forecast for the birth data, with a low MAPE and a valid residual analysis.
- The ARIMA's performance is closed ly followed the Decomposition and Holt-Winters. These can be great alternative options since they are less computationally demanding compared to ARIMA

4. Summary

- I would pick the **Holt-Winters** since its MAPE differs from ARIMA and Decomposition very little, but it is easier to comprehend and demands less processing power. The alpha, beta, and gamma from HW are useful in understanding how the forecasts are generated.
- A multivariable regression model is also a potential method. However, more social economic metrics are needed as input since human birth rates are more likely to be affected by multiple factors.

THANK YOU