Stat 153 Project

Phoebe Abramowitz and Jack Moorer

Introduction

Elaborate on the background of the problem/dataset. Where does the dataset come from? How are the data collected?

The dataset describes the number of female births by day in California in 1959, January first through December 31st. The units are a count and there are 365 observations. The data is provided by the Time Series Data Library, found at https://datamarket.com/en/data/list/?q=provider:tsdl. The data comes from analysis of birth certificates.

Explain the motivation for studying the particular dataset of interest. Why is the dataset interesting?

This dataset is both interesting for the sake of general curiosity and for the impact that date of birth can have on individuals' lives. For example, in his popular book *Outliers*, Malcolm Gladwell demonstrates that being born on a date such that you're the oldest of your cohort in school and childhood activities leads to increased rates of success in academia, sports, ect. Regardless of it's veracity, astrology is an immensely popular way for people to make sense of their lives. Timing of birth-down to the specific day- can affect people's perceptions of themselves and their circumstances. These perceptions can then impact people's decision making.
"The birthday problem" is often discussed in the study of probability. The problem, which asks how likeley it is that at least two people in a group share a birthday, is often prefaced with the assumption that "all birthdays are equally likeley". However, the reasoning for and extent of the disclaimer that this is not necessarily true is rarely expounded upon in these discussions.

EDA

Before we begin our analysis of the Female Birthrates time series, we need to do some exploratory data analysis. First, let's plot the original time series.

Before we continued, we checked for outliers using the method suggested to us. Using the tso() function, we saw we had one additive outlier at time 266. Interesting, we found that this outlier could correlate with conception on New Years Eve, or just during the holidays in general, so the outlier does make sense. We decided to remove this outlier. Here is a plot showing the original time series vs the time series without the additive outlier:

Now that we have removed any outliers we can check for stationarity. The first thing we did was to see if there is an underlying trend in the data. In order to do that we fit a simple linear model, regressing birthrates on time. Below you can see a plot of the time series with the regression line. Clearly the time series has a trend, and the mean is based on time, meaning the time series is not stationary.

We have two options to try to make this time series stationary. The first is we can de-trend the time series by subtracting the linear trend. In order to do that, we just examine the residuals of regression birthrates on time. Below is the time series de-trended via OLS.

Clearly the mean is not dependent on time, but in order to check if this time series is stationary we need to check the ACF and PACF of the de-trended time series. Below is the ACF and PACF of the de-trended time series.

For both the PACF and ACF the values are not tailing off, and actually reach a spike at lag-21 before tailing off. This could mean the time series is not stationary, or it could mean the time series is seasonal. In order to see if the time series is seasonal lets look at the original ACF and PACF of the time series.

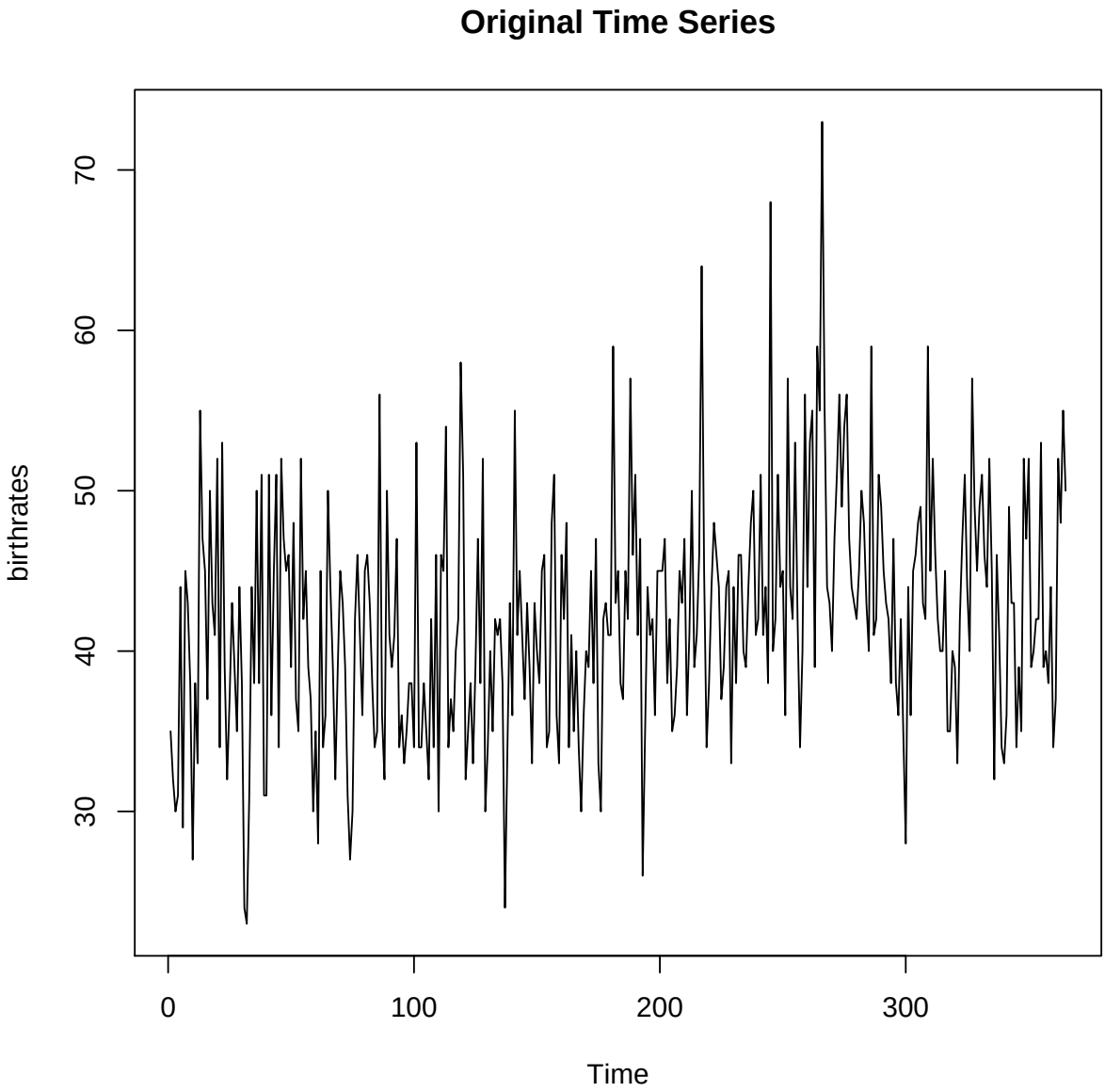


Figure 1: original_time_series

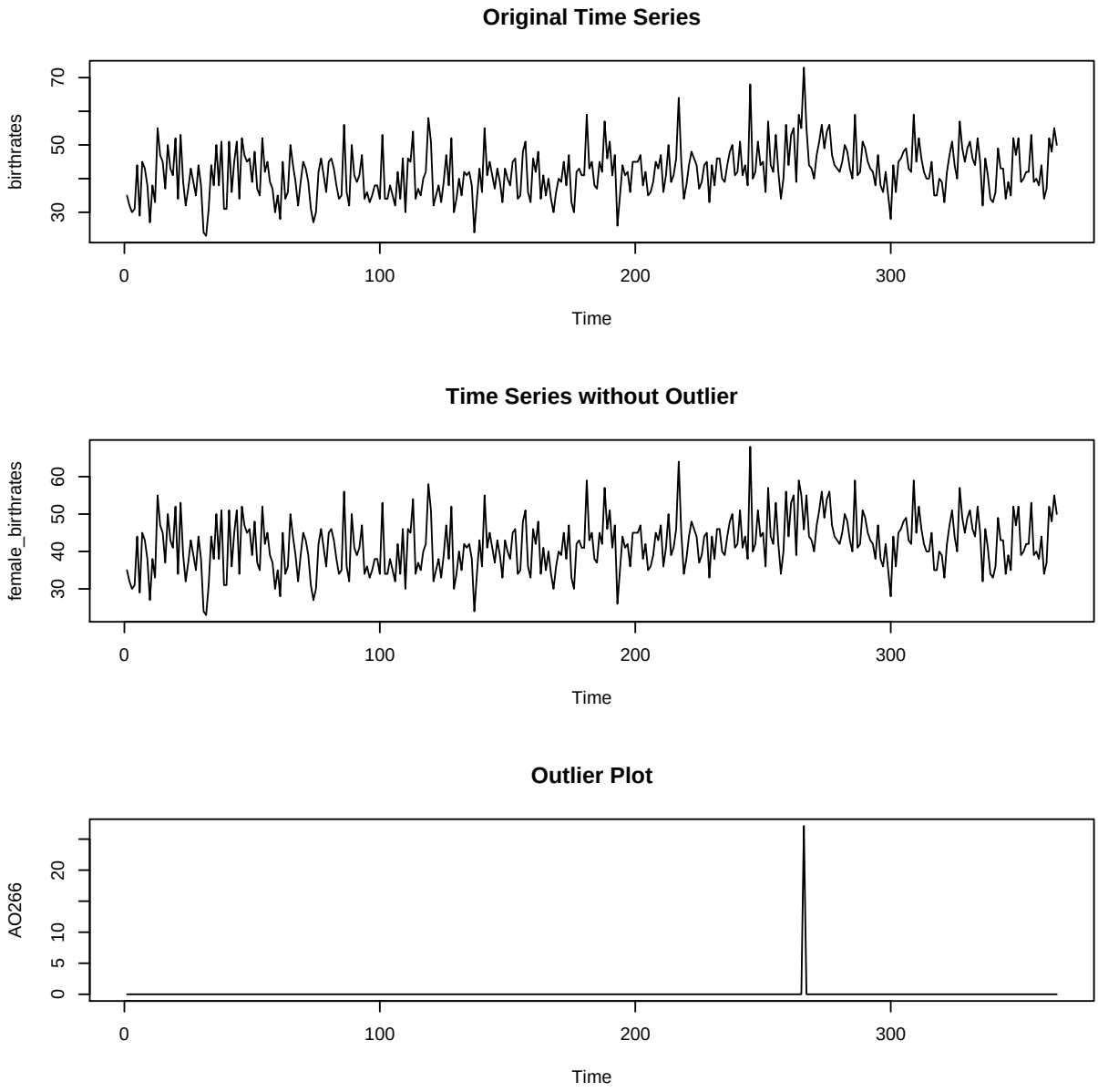


Figure 2: outlier_affect

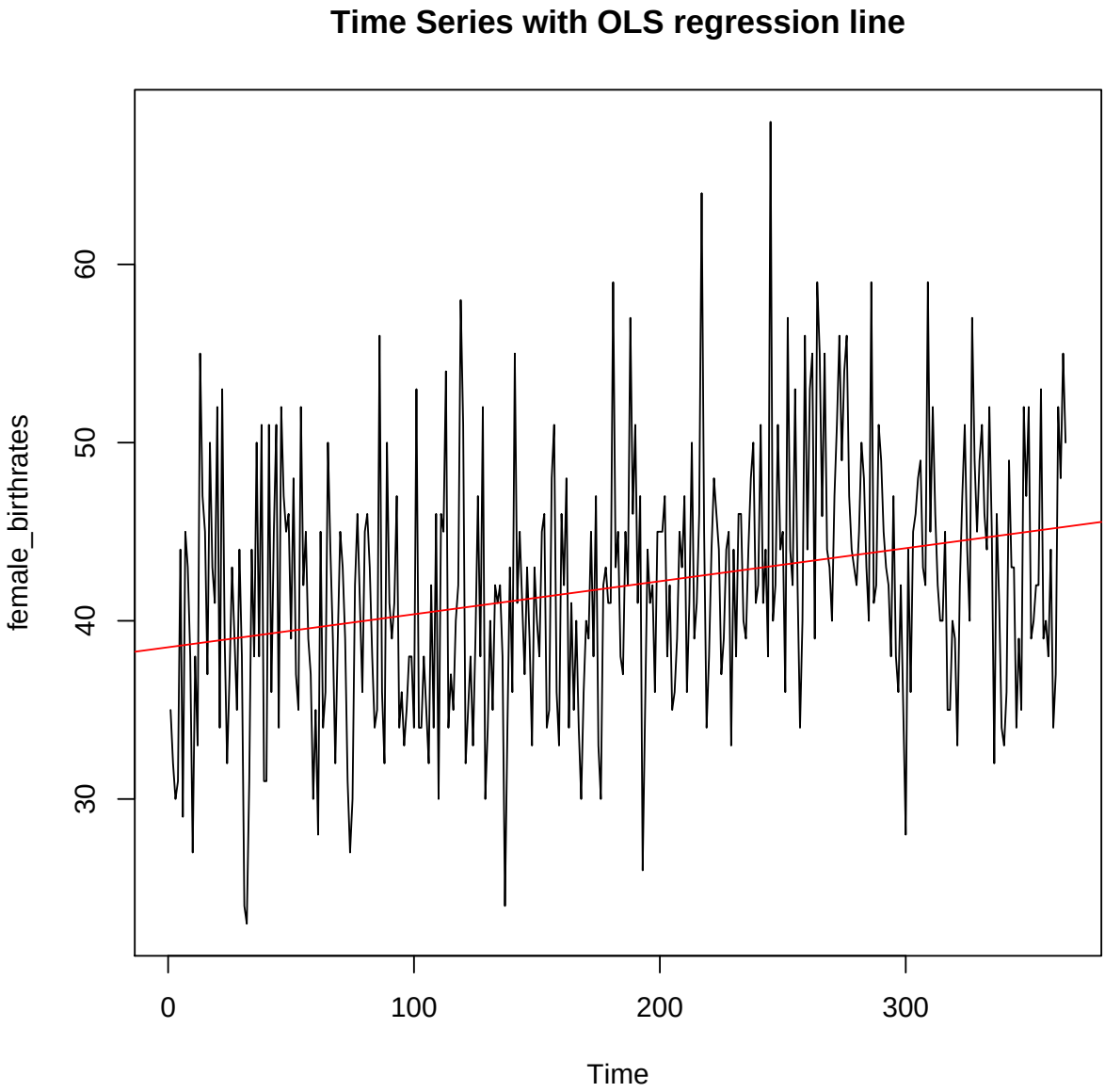


Figure 3: trend

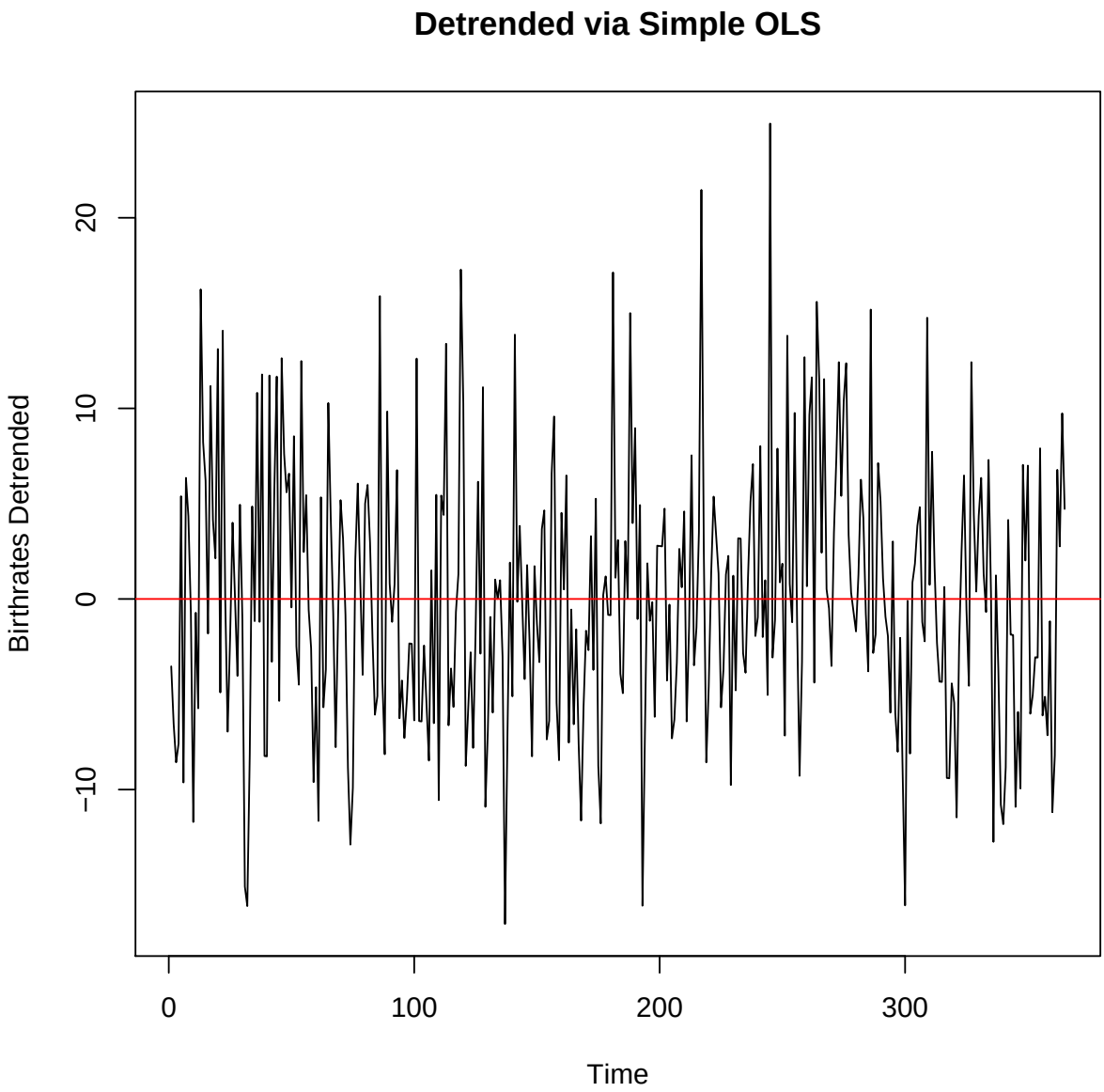


Figure 4: detrend

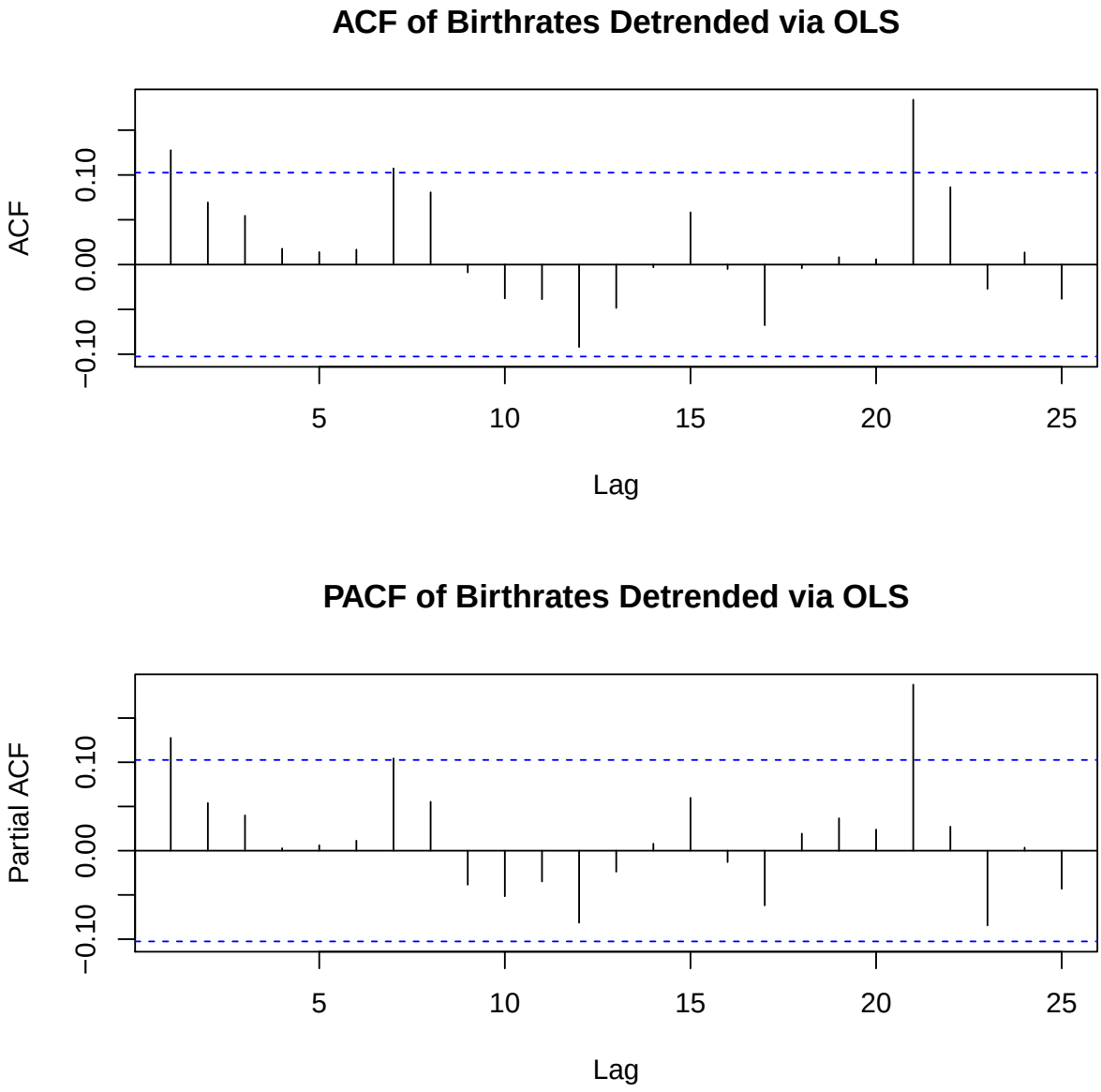
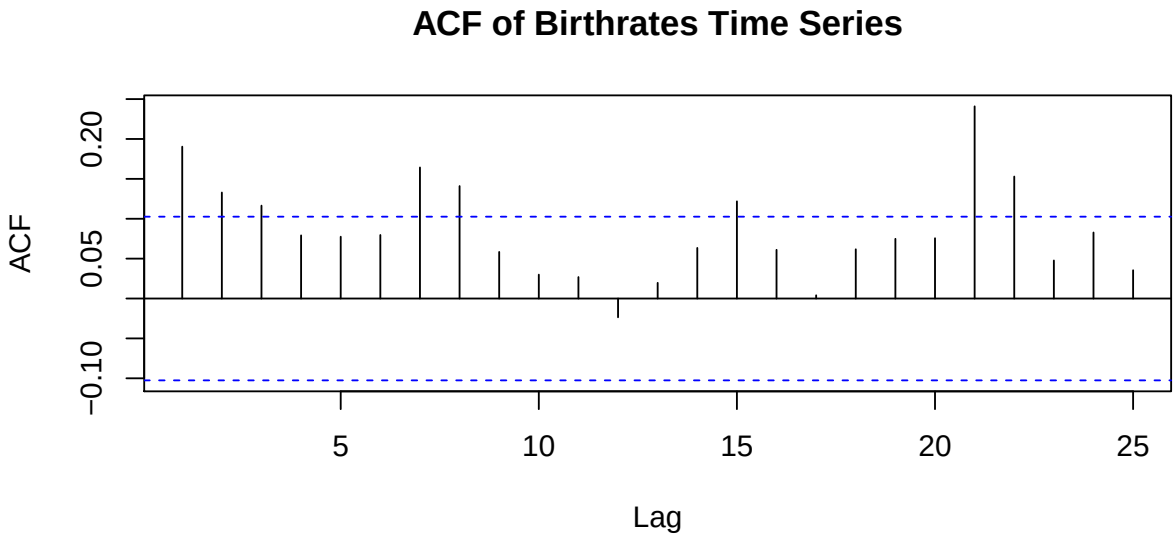


Figure 5: detrend_acf_pacf



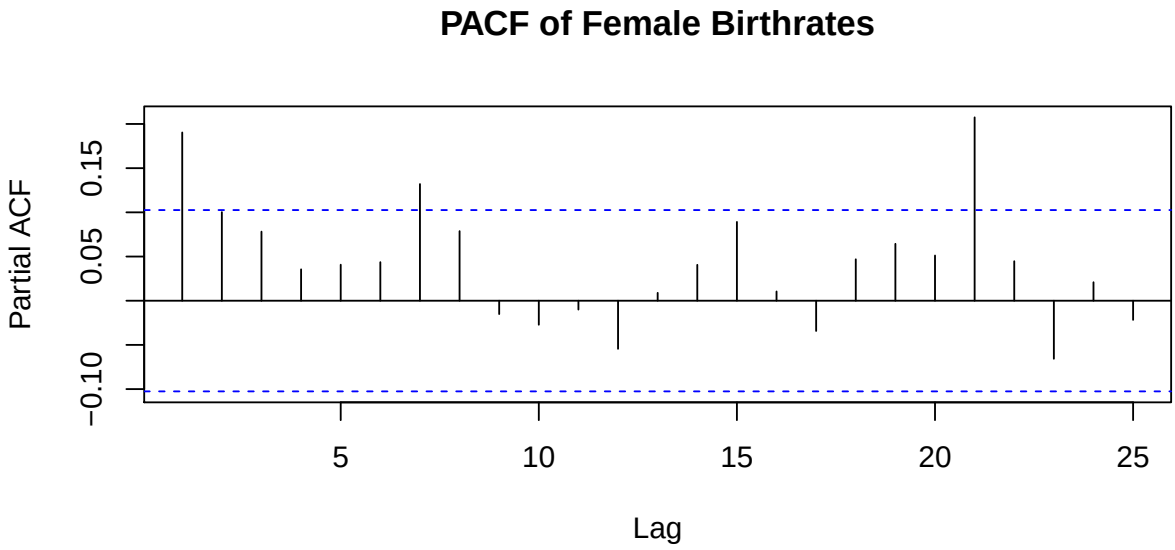


Figure 6: orginal_acf_pacf

There doesn't seem to be a seasonal trend in the original time series or the de-trended time series. We will come back to this when we are fitting our ARIMA model, but we have strong evidence throughout our analysis that this data is not seasonal. The next step would be to look at the differenced time series. Below is the differenced time series with a line noting the average of the time series. Clearly the mean of differenced time series does not depend on time.

Now we can look at the ACF and PACF of the time series. These will also be used to try to fit our ARIMA model later. Below is the ACF and PACF of the differenced time series.

Once again we don't have clear evidence this is stationary. Clearly the model is not just a AR or MA model, since both the PACF and ACF don't cut off sharply. For the ACF the significance seems to cut off after lag-1, but there is a spike at lag-21. For the PACF the significance cuts off after lag-7, but there is a spike at lag-20. Neither the PACF nor ACF tails off.

Because we don't have strong evidence we have a stationary time series yet we are going to look at the second difference of the time series. Because we want to avoid over-differencing for our ARIMA models, we will only use the second difference if it clearly makes the model look stationary. You can see the results of these plots in the appendix, however there isn't any strong evidence that differencing again made an impact on making the time series more stationary, so to avoid over-differencing we will only look at the first difference for out models.

ARIMA

Based on the first difference, we don't have evidence that we have a purely MA or AR model. However, the last significant lag for the PACF is lag 20 and the last significant lag for the ACF is lag-21. Because of this one of the models we will try to fit will be an ARIMA(20, 1, 21) model. However, this model is very complicated, and we want to avoid over-fitting our data. Because of this, we will look at two simpler ARIMA models as well. Note that for the PACF of the first difference of the time series, the significance cuts off at lag-7, before reaching a spike at lag-20. In addition, for the ACF the significance cuts off at lag-1 before reaching a spike at lag-21. Because of this we will also fit an ARIMA(7, 1, 1) model. We also wanted to look at an overly-simple model, and fit an ARIMA(1, 1, 1) model.

Using sarima() here are the results of the ARIMA(20, 1, 21) model:

There are several insights the plots from sarima() provide for us here. For one, the Normal Q-Q plot of standardized residuals tells us that the residuals are normally distributed, so the assumption of a Gaussian distribution is valid. We also see from the ACF of the residuals that the ACF at any lag is within the innovations significance bar, which gives evidence that this is a good model. Before looking at the model statistics, lets use sarima() on the other two models. Here are the results of running sarima() on the ARIMA(7, 1, 1) model:

Once again the assumption of normality is supported, but there is a spike at lag-21 for the ACF of the residuals, implying this may not be a great model. Let's look at the ARIMA(1, 1, 1) model, and then compare the model statistics for each:

The plots for this model look similar to ARIMA(7, 1, 1), however, the Ljung-Box statistics is showing higher p-values, implying this is a worse model. You can see the model statistics of each ARIMA model in the appendix.

The ARIMA(20, 1, 21) model has the best model statistics. However, as I mentioned before, I don't want to only look at a very complicated model, so I will continue to analyze ARIMA(7, 1, 1), since it has the second best model statistics. In order to find which of these two models is best, I am going to see how they perform in forecasting unseen data. I am going to train both models using a train data set with the last 10 observations from female birthrates missing, then see how well the models can forecast these 10 observations.You can see both plots for the forecasted training data in the appendix. Both forecasts don't look like they are doing great, but let's compare the test MSE to see which one performed better.

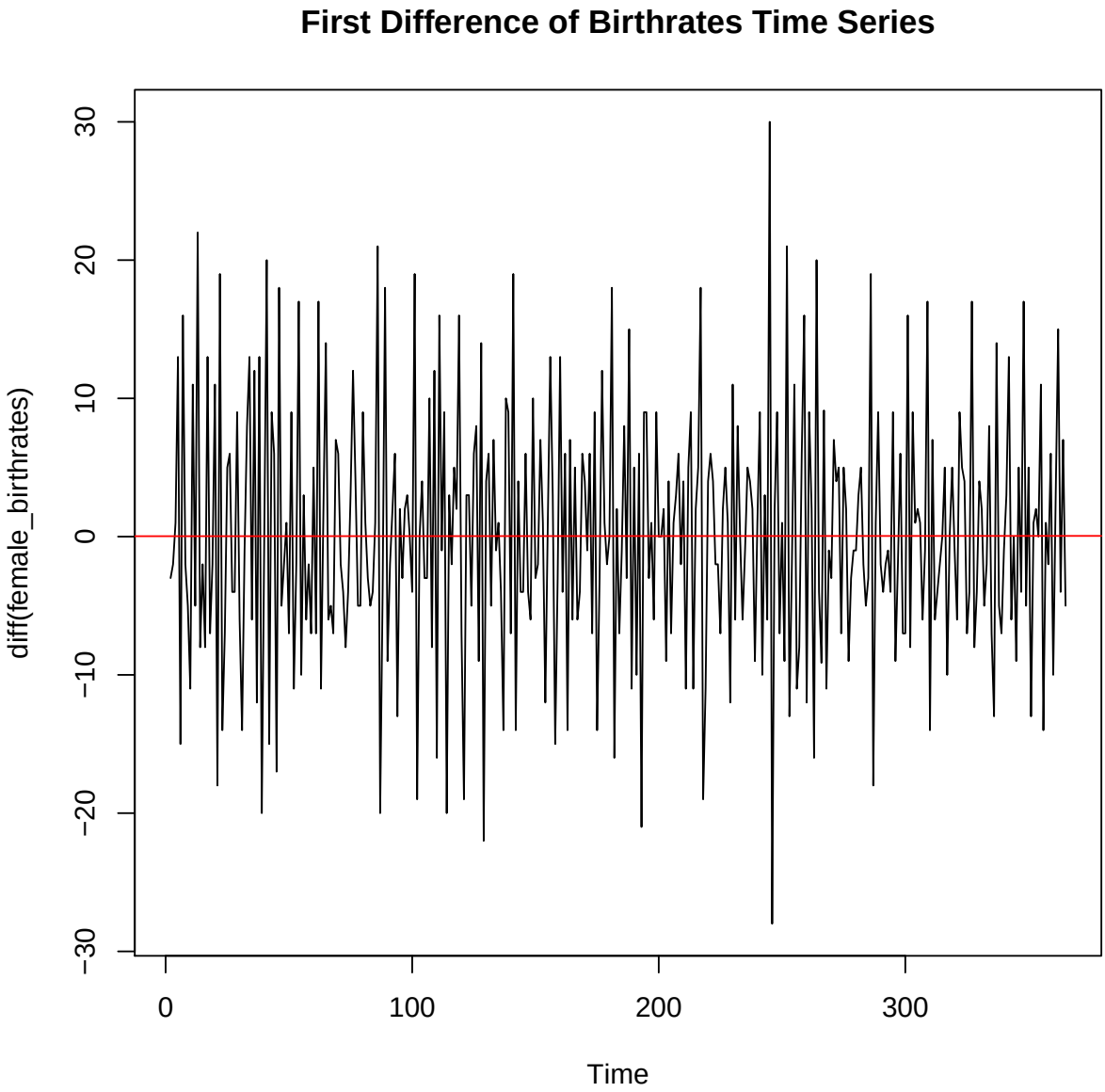
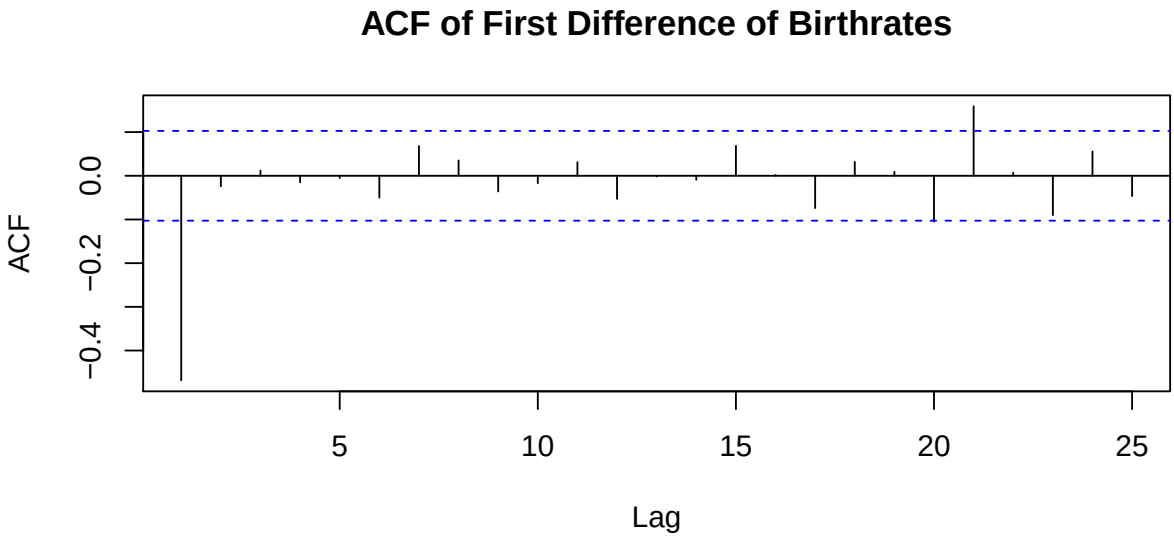


Figure 7: differnced_ts



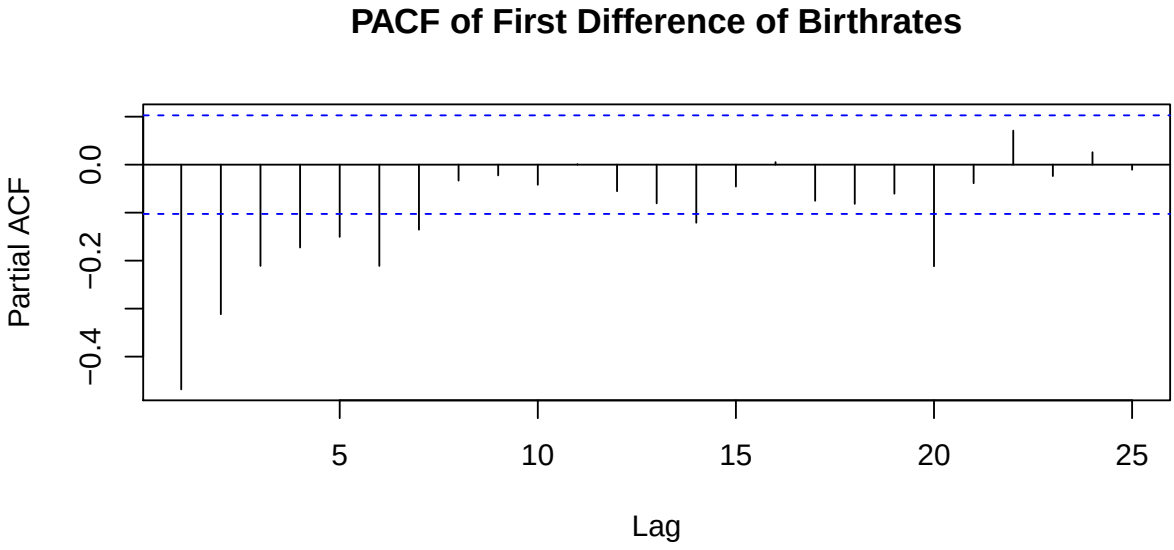


Figure 8: differnced_ts_acf_pacf

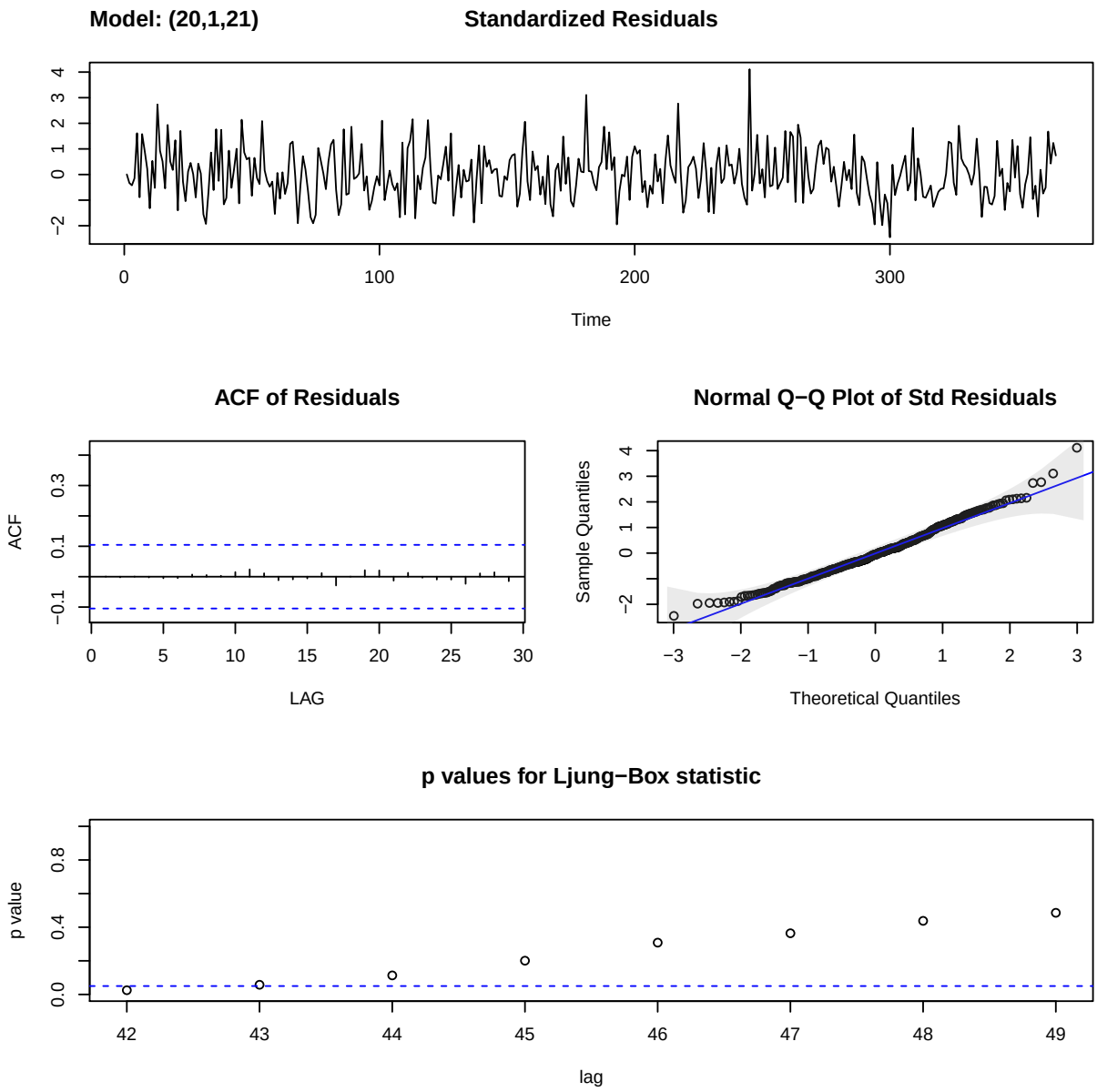


Figure 9: second_model

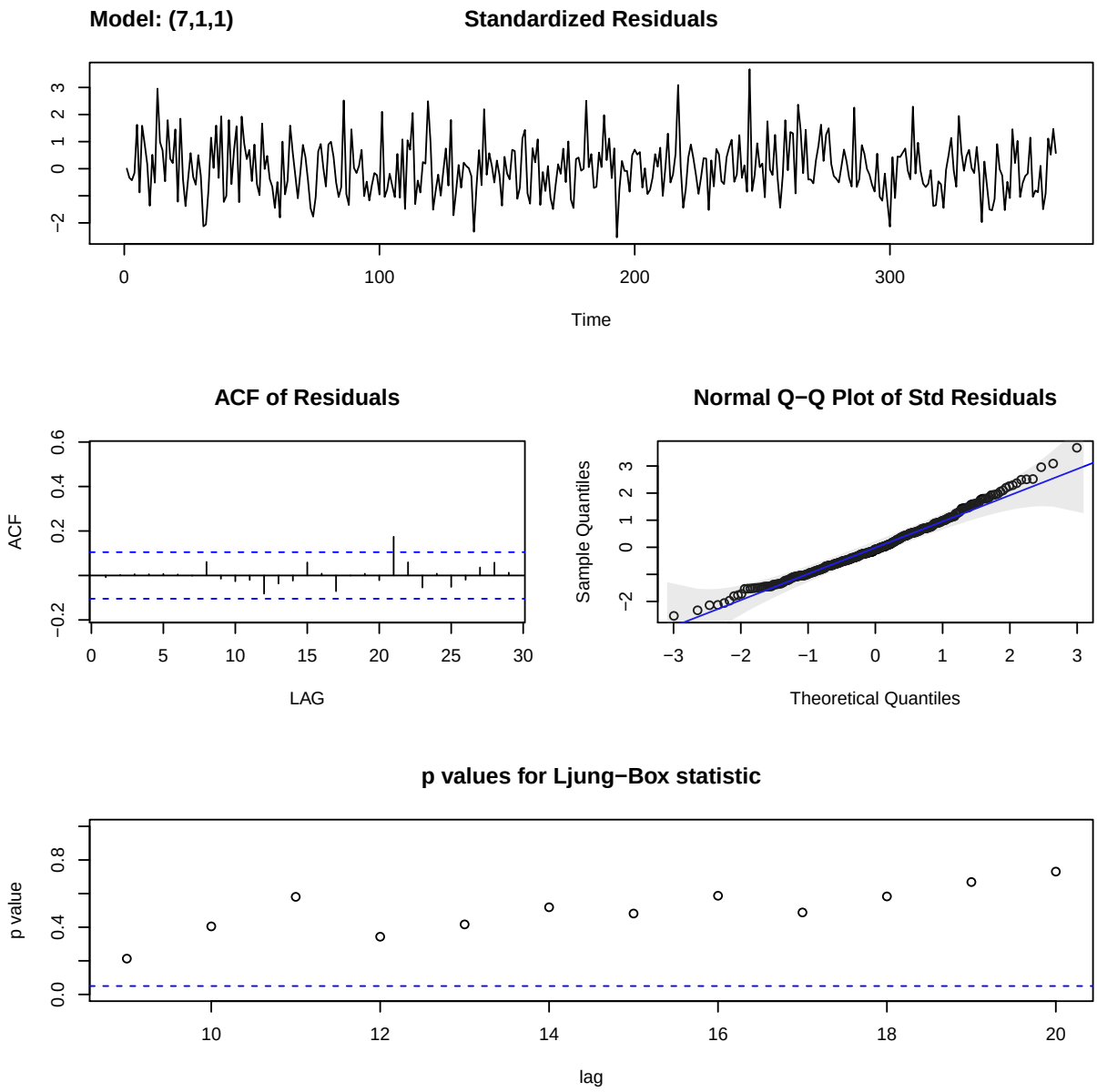


Figure 10: first_model

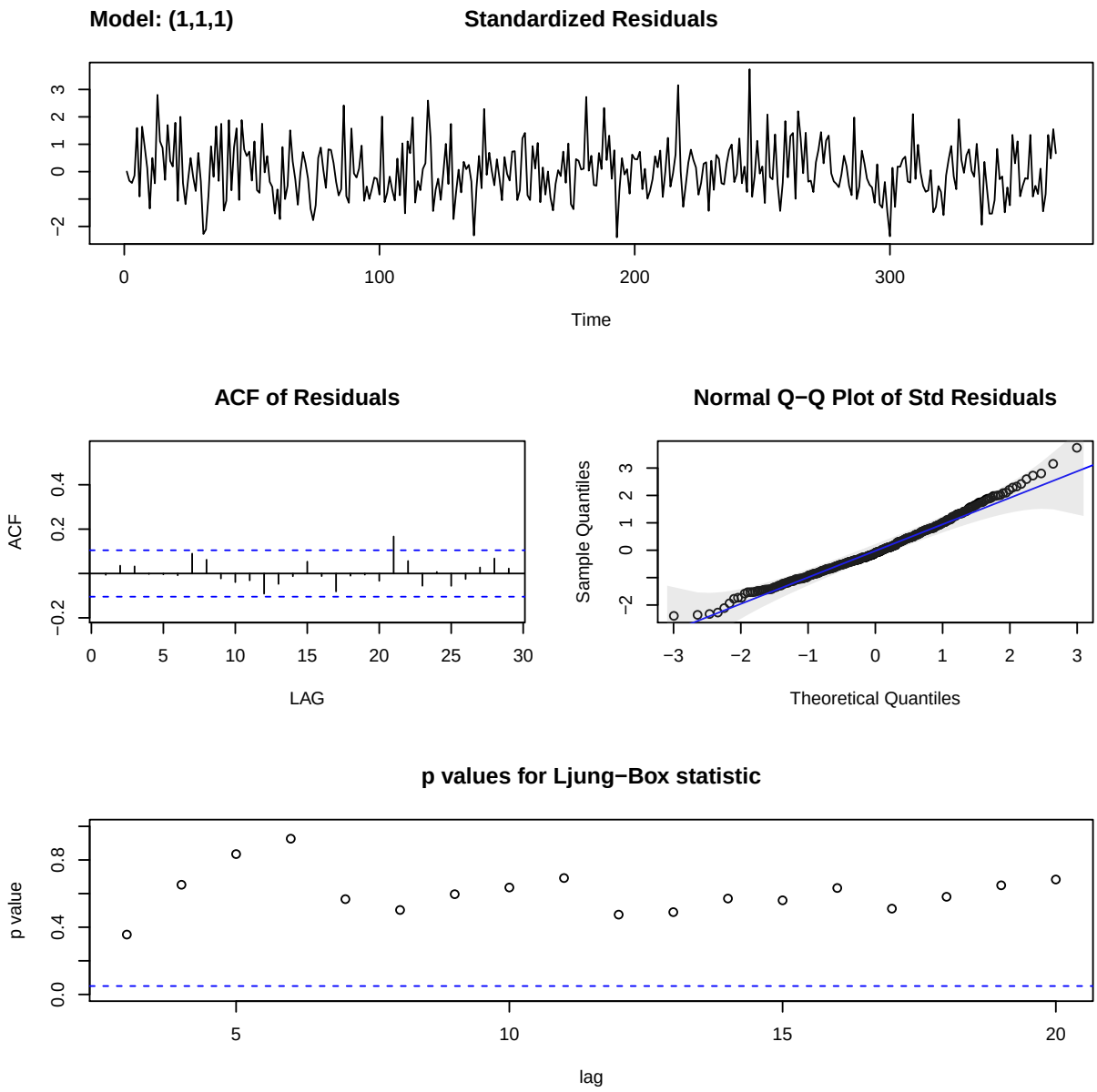


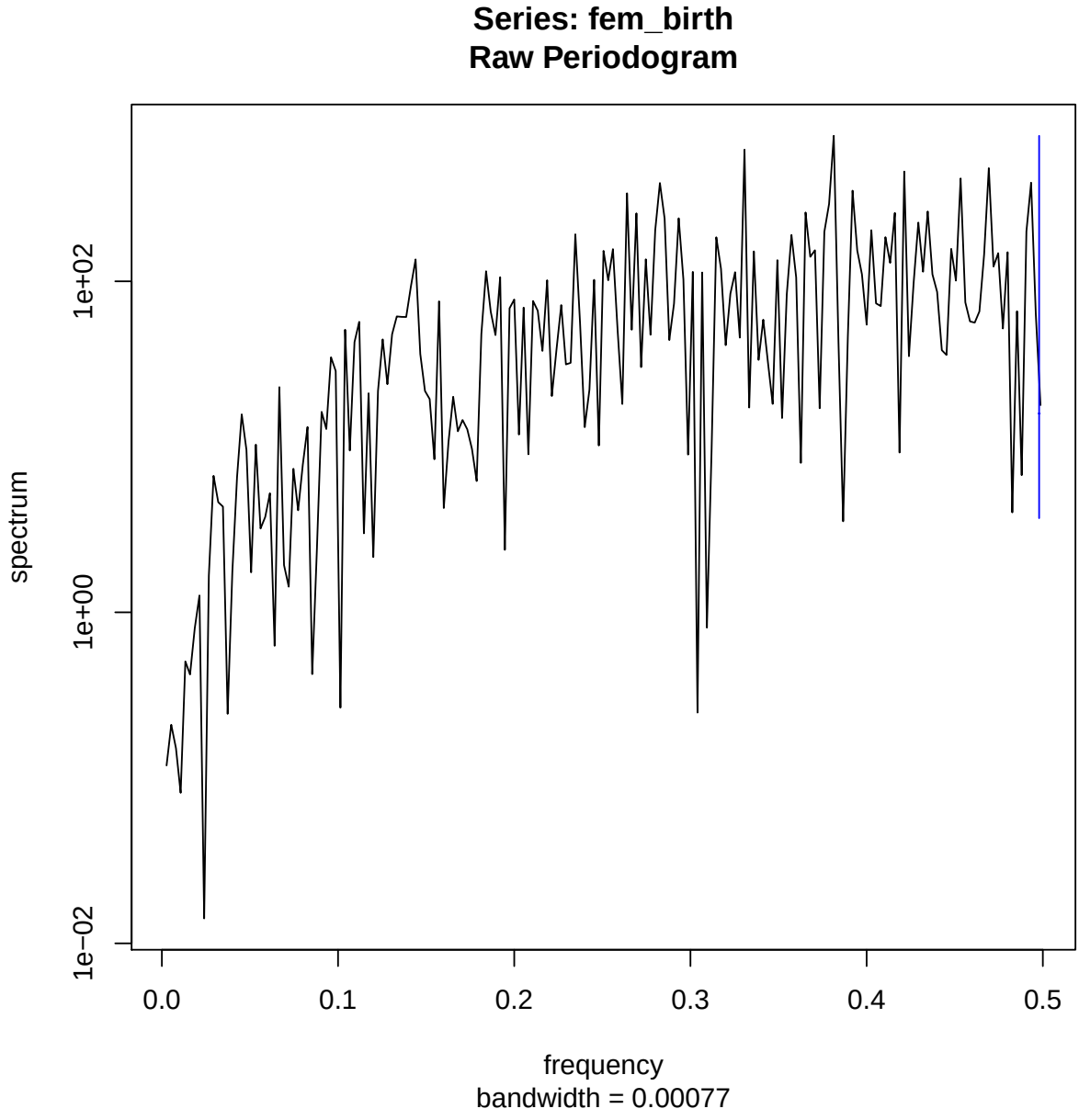
Figure 11: third_model

The test MSE for the ARIMA(20, 1, 21) model was 46.5991992308594. The test MSE for the ARIMA(7, 1, 1) model was 47.546990399347. As you can see the test MSE of the ARIMA(20, 1, 21) model was slightly better. We can now move on to forecasting the best model, the ARIMA(20, 1, 21) model. Below is the 10 step ahead forecast for the ARIMA(20, 1, 21 model):

As you can see, the forecast does not look like it is doing a great job, and the confidence interval is much too large. We have pretty clear evidence that an ARIMA method will not be able to model the female birthrates dataset.

Spectral Analysis

Let's look at the periodogram of our detrended data, detrended here using differencing once.



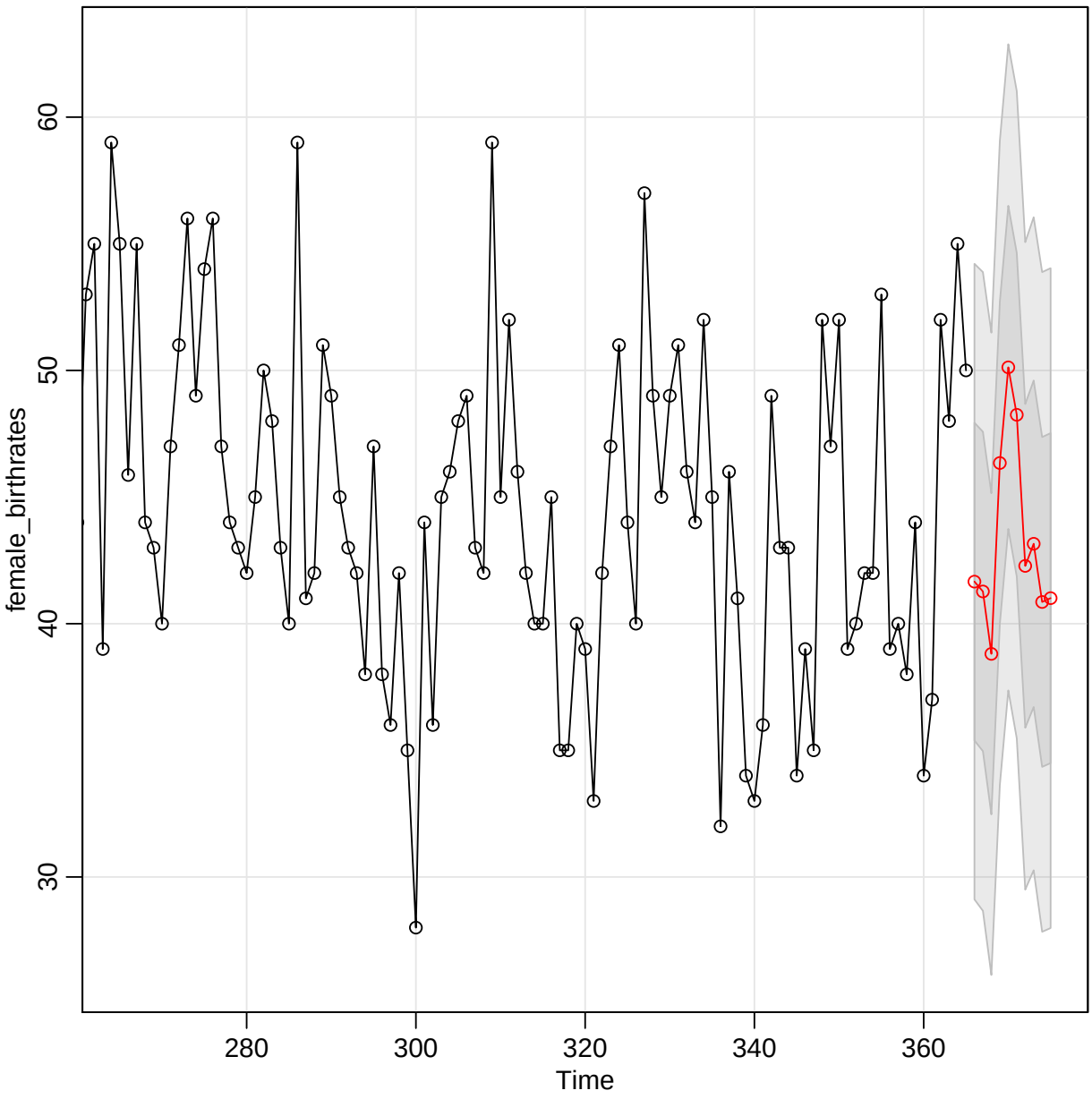
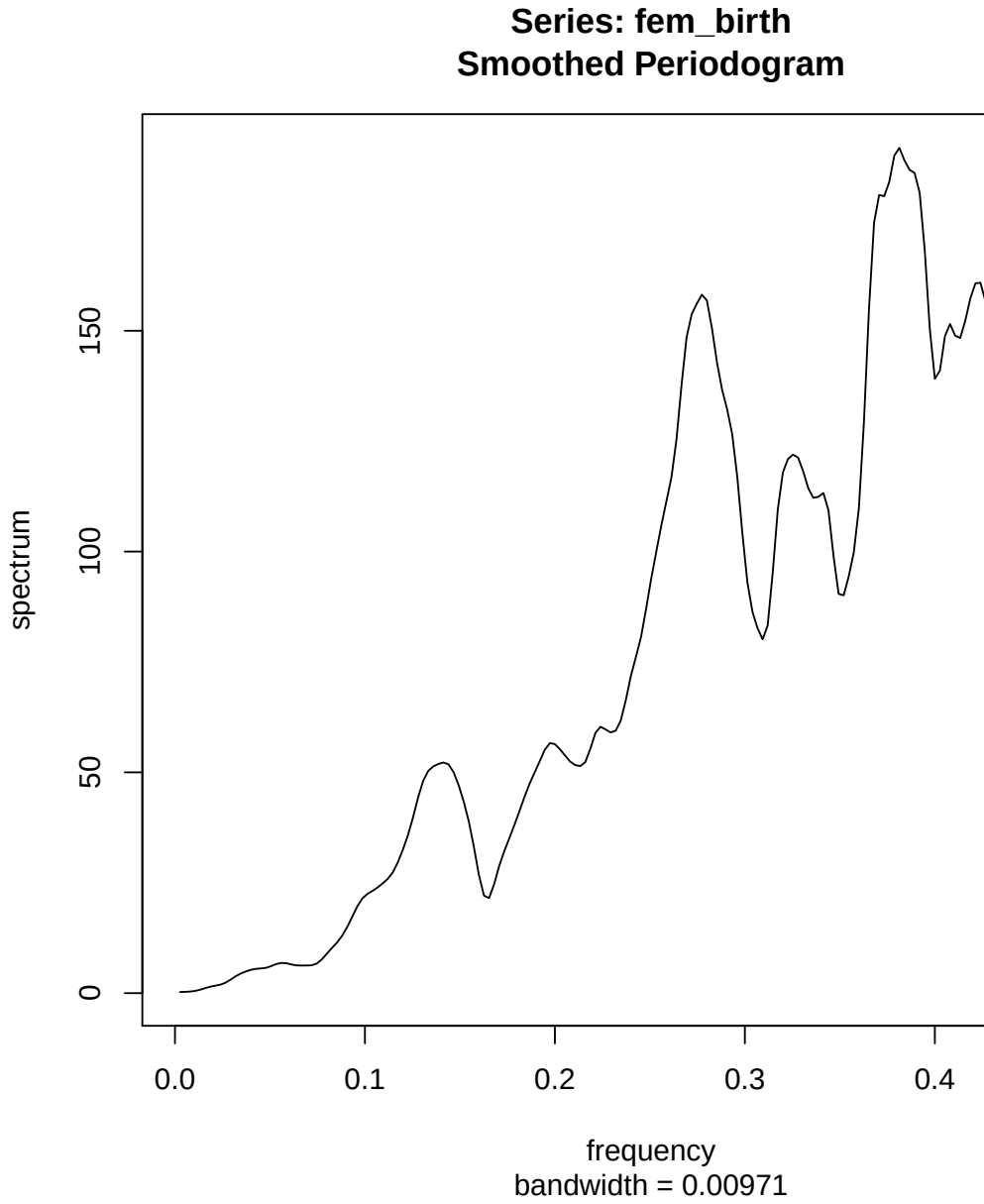


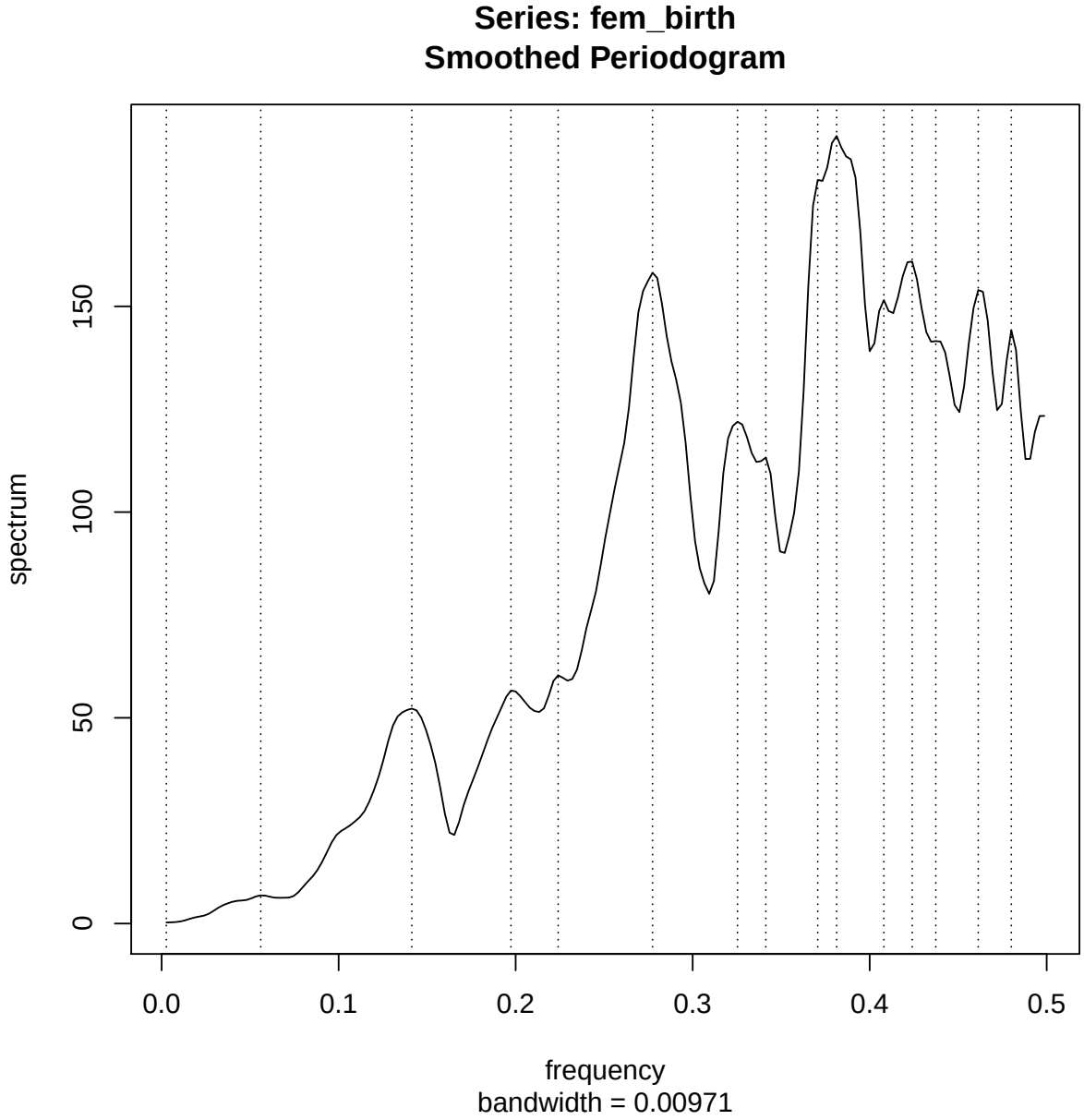
Figure 12: arima_forecast



Then, take the smoothed periodogram.

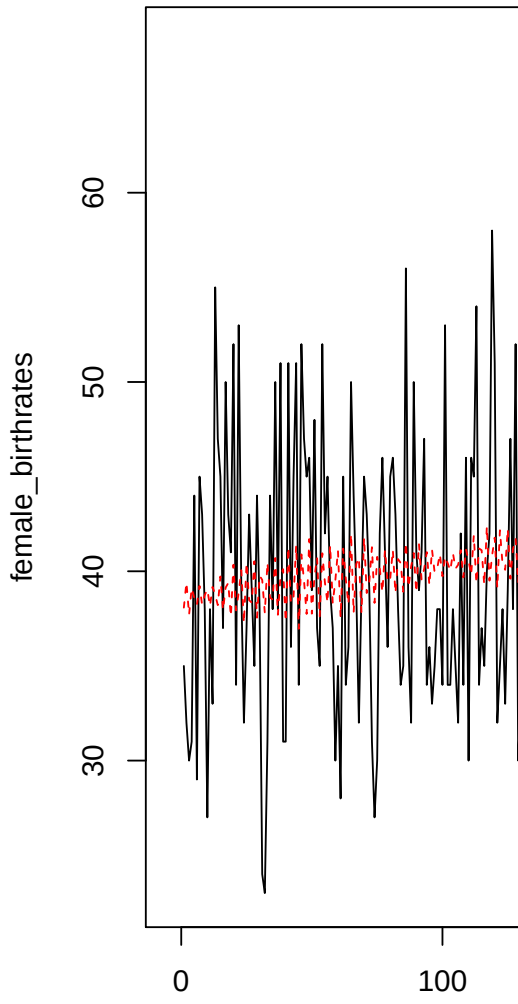
From this smoothed periodogram, we can choose the top three lags and feed the modal.

We append the first fundamental frequency and exhibit all local maxima on the periodogram before selecting the



top three lags.

The top three lags we observe are: 0.3813333 0.3706667 0.4240000

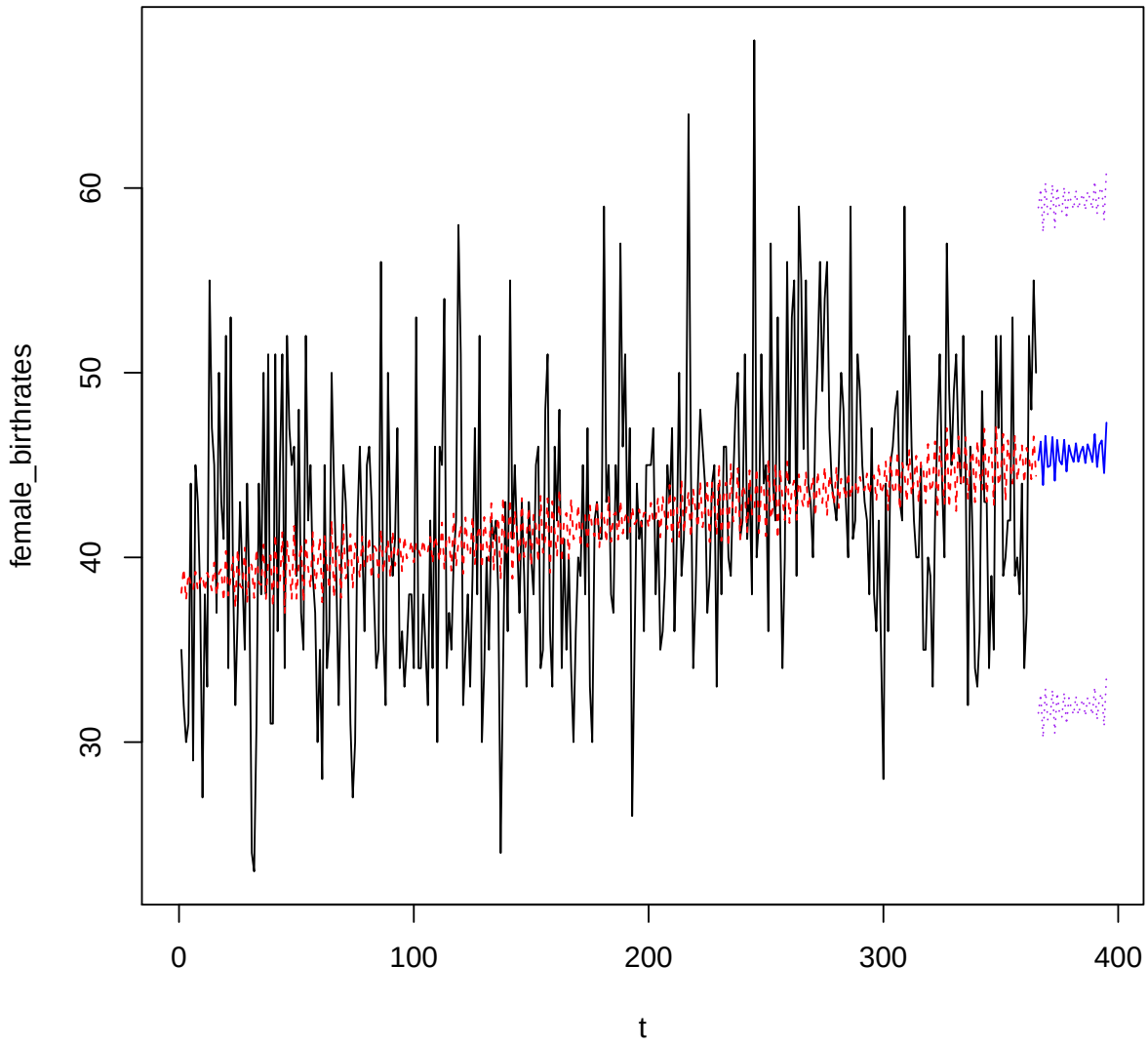


We can then compare with the parametric spectral estimator (the red dotted line):

We can use our three frequencies to generate features (sin terms and cosine terms). We use the following model:\

$$t = \alpha + \beta t + \sum_{j=1}^{3} c_j cos(2\pi\omega * jt) + d_j cos(2\pi\omega * jt) + w_t$$

Then, lastly, we'll forecast the next 30 obervations(1 month) with a 95% confidence level on our plot of the gener-



ated features.

The bounds of our confidence interval encapsulate most of the observed values from the entire dataset at every point, and thus aren't particularly informative.

Conclusion

Given that we only have data for one year, accounting for the spike in September, which could be seasonality if looked at over multiple years, is out of the scope of our analysis. It's impossible to tell from our dataset how much of the upward trend demonstrated by slope of our OLS regression line is due to a long-term increase in birthrates, and how much of it is due to a seasonal increase in birthrates as an individual year progresses from January. It's worthwhile to analyze similar birthdate data over several years. The spike in September suggests that people are concieving around the holidays. It's possible that the outlier at 266 days is caused by conception on or around New Years Eve. This discussion of birthrates on particular dates also suggest that it would be interesting to look at data regarding gestation time from conception to birth.

Appendix

Code Used

```
library(dplyr)
library(astsa)
library(tseries)
library(TSA)
library(tsoutliers)

#read in data
raw_birthrates = read.csv("../data/daily-total-female-births-in-cal.csv")

#remove final row
clean_birthrates <- filter(raw_birthrates, Date!="Daily total female births in California")

#define column names
colnames(clean_birthrates) <-  c('date',"daily_female_births")

#convert dates to Data data type
clean_birthrates$date <- as.Date(clean_birthrates$date)

#make dates row names
birthrates <- clean_birthrates$daily_female_births
female_birthrates <- data.frame(birthrates)
rownames(female_birthrates) <- clean_birthrates$date

#convert to time series
female_birthrates <- ts(female_birthrates)


#create time series plot
pdf('../images/orignal_ts.pdf')
plot(female_birthrates, main = "Original Time Series")
dev.off()

#show outlier
outliers <- tso(female_birthrates, types = c("TC", "AO", "LS", "IO", "SLS"))

pdf('../images/outliers.pdf')
plot(outliers)
dev.off()

#remove outlier
#get index and time of outlier
outlier_indx <- outliers$outliers$ind

#length of ts
n <- length(female_birthrates)

#find outlier effect
ao <- outliers("AO", outlier_indx)
ao_effect <- outliers.effects(ao, n)
```

```r
coefhat <- as.numeric(outliers$outliers['coefhat'])
ao_effect <- coefhat*ao_effect

#generate outlier affect time series
ao_effect_ts <- ts(ao_effect, frequency = frequency(female_birthrates), start= start(female_birthrates))

#generate time series without outlier
birthrates_wo_ao <- female_birthrates - ao_effect_ts

#plot original time series and time series without outlier
pdf('../images/outlier_affect.pdf')
par(mfrow = c(3, 1))
plot(female_birthrates, main = "Original Time Series")
plot(birthrates_wo_ao, main = "Time Series without Outlier")
plot(ao_effect_ts, main = "Outlier Plot")
dev.off()

#generate birthrates time series without outlier
rownames(birthrates_wo_ao) <- clean_birthrates$date

#replace female_birthrates data frame
female_birthrates <- birthrates_wo_ao

#plot time series with no outlier
pdf('../images/ts_no_outlier.pdf')
plot.ts(female_birthrates, main = "Time Series Plot of Female Birthrate Data (Outlier Removed)")
dev.off()

#write in clean data
write.csv(female_birthrates, file = "../data/clean_data.csv", row.names = TRUE)

#plot acf and pacf of ts
pdf('../images/original_ts_acf_pacf.pdf')
par(mfrow = c(2, 1))
acf(female_birthrates, main = "ACF of Birthrates Time Series")
pacf(female_birthrates, main = "PACF of Female Birthrates")
dev.off()

#plot ts with regression line
pdf('../images/trend_in_ts.pdf')
plot.ts(female_birthrates, main = "Time Series with OLS regression line")
abline(reg = lm(female_birthrates ~ time(female_birthrates)), col = 'red')
dev.off()


#get model
reg = lm(female_birthrates ~ time(female_birthrates))

#plot detrended time series vie OLS regression
pdf('../images/detrended_ts.pdf')
plot(resid(reg), type = 'l', main = "Detrended via Simple OLS", ylab = "Birthrates Detrended", xlab = "
abline(lm(resid(reg) ~ time(resid(reg))), col = 'red')
dev.off()
```

```r
pdf('../images/detrended_ts_acf_pacf.pdf')
par(mfrow = c(2, 1))
acf(resid(reg), main = "ACF of Birthrates Detrended via OLS")
pacf(resid(reg), main = "PACF of Birthrates Detrended via OLS")
dev.off()

#take the first difference of the time series
diff_vals <- diff(female_birthrates)

#plot diff of time series with linear region line
pdf('../images/first_diff_ts.pdf')
ts.plot(diff(female_birthrates), main = "First Difference of Birthrates Time Series")
abline(reg = lm(diff(female_birthrates) ~ time(diff(female_birthrates))), col = 'red')
dev.off()

pdf('../images/firstdiff_ts_acf_pacf.pdf')
par(mfrow = c(2, 1))
acf(diff(female_birthrates), main = "ACF of First Difference of Birthrates")
pacf(diff(female_birthrates), main = "PACF of First Difference of Birthrates")
dev.off()

#get second diff
second_diff_vals <- diff(diff(female_birthrates))

#plot second diff of time series with linear region line
pdf('../images/second_diff_ts.pdf')
ts.plot(diff(diff(female_birthrates)), main = "Second Difference of Birthrates Time Series")
abline(reg = lm(diff(diff(female_birthrates)) ~ time(diff(diff(female_birthrates)))), col = 'red')
dev.off()


pdf('../images/seconddiff_ts_acf_pacf.pdf')
par(mfrow = c(2, 1))
acf(diff(diff(female_birthrates)), main = "ACF of Second Difference of Birthrates")
pacf(diff(diff(female_birthrates)), main = "PACF of Second Difference of Birthrates")
dev.off()


#fit model

#first arima
pdf('../images/first_model.pdf')
first_model <- sarima(female_birthrates, 7, 1, 1)
dev.off()

first_aic <- paste('AIC:', first_model$AIC, sep = ' ')
first_aicc <- paste('AICc:', first_model$AICc, sep = ' ')
first_bic <- paste('BIC:', first_model$BIC, sep = ' ')

#model diagnostics
sink('../results/first_model.txt')
first_aic
first_aicc
```

```
first_bic
sink()

#second arima
pdf('../images/second_model.pdf')
second_model <- sarima(female_birthrates, 20, 1, 21)
dev.off()

second_aic <- paste('AIC:', second_model$AIC, sep = ' ')
second_aicc <- paste('AICc:', second_model$AICc, sep = ' ')
second_bic <- paste('BIC:', second_model$BIC, sep = ' ')

#model diagnostics
sink('../results/second_model.txt')
second_aic
second_aicc
second_bic
sink()

#third arima
pdf('../images/third_model.pdf')
third_model <- sarima(female_birthrates, 1, 1, 1)
dev.off()

third_aic <- paste('AIC:', third_model$AIC, sep = ' ')
third_aicc <- paste('AICc:', third_model$AICc, sep = ' ')
third_bic <- paste('BIC:', third_model$BIC, sep = ' ')

#model diagnostics
sink('../results/third_model.txt')
third_aic
third_aicc
third_bic
sink()

#forecast
pdf('../images/arima_forecast.pdf')
first_forcast <- sarima.for(female_birthrates, n.ahead = 10, p= 20, d = 1, q = 21)
dev.off()

pdf('../images/arima_forecast_2.pdf')
second_best_forcast <- sarima.for(female_birthrates, n.ahead = 10, p= 7, d = 1, q = 1)
dev.off()

train <- female_birthrates[1:355]
test <- female_birthrates[356:365]

pdf('../images/train_arima_forecast.pdf')
train_forcast <- sarima.for(train, n.ahead = 10, p= 20, d = 1, q = 21)
dev.off()

arima_preds <- train_forcast$pred
```

```r
arima_test_mse1 <- sum((1/10)*(test - arima_preds)^2)

pdf('../images/train_second_arima_forecast.pdf')
second_train_forecast <- sarima.for(train, n.ahead = 10, p= 7, d = 1, q = 1)
dev.off()

second_arima_preds <- second_train_forecast$pred

arima_test_mse2 <- sum((1/10)*(test - second_arima_preds)^2)

ar_test_1 <- paste("Best ARIMA Model Test MSE:", arima_test_mse1, sep = ' ')
ar_test_2 <- paste("Second Best ARIMA Model Test MSE:", arima_test_mse2, sep = ' ')

sink('../results/arima_test_results.txt')
ar_test_1
ar_test_2
sink()

##Spectral Analysis Appendix
#load data and detrend using differencing
source('./CleanData.R')
fem_birth <- diff(female_birthrates)

#make periodogram of the data
pdf('../images/periodogram.pdf')
spec.pgram(fem_birth)
dev.off()

#For smoothed periodogram, try different values for the kernel & taper
pdf('../images/smoothed-pgram-1.pdf')
spec.pgram(fem_birth,kernel('daniell',3), taper = 0.1)
dev.off()
pdf('../images/smoothed-pgram-2.pdf')
spec.pgram(fem_birth,kernel('modified.daniell',3), taper = 0.2)
dev.off()
pdf('../images/smoothed-pgram-3.pdf')
spec.pgram(fem_birth,kernel('modified.daniell',5), taper = 0)
dev.off()
pdf('../images/smoothed-periodogram.pdf')
spec.pgram(fem_birth,kernel('modified.daniell',c(1,6,1)), taper = 0.1,log="no")
dev.off()

#exhibit all local maxima
pdf('../images/pgram-local-maxima.pdf')
pgram <- spec.pgram(fem_birth,kernel('modified.daniell',c(1,6,1)), taper = 0.1,log="no")
key_freq_ind <- c(1, which(diff(sign(diff(pgram$spec)))==-2) + 1)
key_freq <- pgram$freq[key_freq_ind]
abline(v=key_freq, lty=3)
dev.off()

#choose the top three lags
top_freq <- key_freq[order(pgram$spec[key_freq_ind], decreasing = T)][1:3]
```

```r
#compare to the parametric spectral estimator
pdf('../images/parametric-spectral-estimator.pdf')
pgram <- spec.pgram(fem_birth,kernel('modified.daniell',c(1,6,1)), taper = 0.1,log="no")
pgram_ar <- spec.ar(fem_birth, plot=F)
lines(pgram_ar$freq, pgram_ar$spec, lty=2, col="red")
dev.off()

#generate features
t <- 1:length(female_birthrates)
periodic_terms <- do.call(cbind, lapply(top_freq, function(freq) {
  cbind(cos(2 * pi * freq * t), sin(2 * pi * freq * t))
}))
df <- data.frame(female_birthrates, t, periodic_terms)
fit_final <- lm(female_birthrates ~ ., df)
#plot
pdf('../images/parametric-spectral-estimator.pdf')
plot(t, female_birthrates, type="l")
lines(t, fit_final$fitted.values, lty=2, col="red")
dev.off()

# forecast the next 30 obervations with a 95% confidence level.
t_new <- (tail(t, 1) + 1):(tail(t, 1) + 30)
periodic_terms_new <- do.call(cbind, lapply(top_freq, function(freq) {
  cbind(cos(2 * pi * freq * t_new), sin(2 * pi * freq * t_new))
}))
df_new <- data.frame(t_new, periodic_terms_new)
colnames(df_new) <- colnames(df)[-1]
predictions <- predict.lm(fit_final, newdata=df_new,interval="prediction", level=.95)
#plot predictions and interval
pdf('../images/spectral-predictions.pdf')
plot(t, female_birthrates, type="l", xlim=c(0, tail(t_new, 1)))
lines(t, fit_final$fitted.values, lty=2, col="red")
lines(t_new, predictions[, "fit"], col="blue")
matlines(t_new, predictions[, 2:3], col = "purple", lty=3)
dev.off()
```

Second differenced model

Model Statistics for the 3 ARIMA models

Here is the AIC, AICc, and BIC of the ARIMA(20, 1, 21) model:

AIC: 4.89305737044942 AICc: 4.93083315673372 BIC: 4.3418126823685

Here is the AIC, AICc, and BIC of the ARIMA(7, 1, 1) model:

AIC: 4.87895885933218 AICc: 4.88614096598027 BIC: 3.97512071188627

Here is the AIC, AICc, and BIC of the ARIMA(1, 1, 1) model:

AIC: 4.86904410372145 AICc: 4.87482796977929 BIC: 3.90109805457282

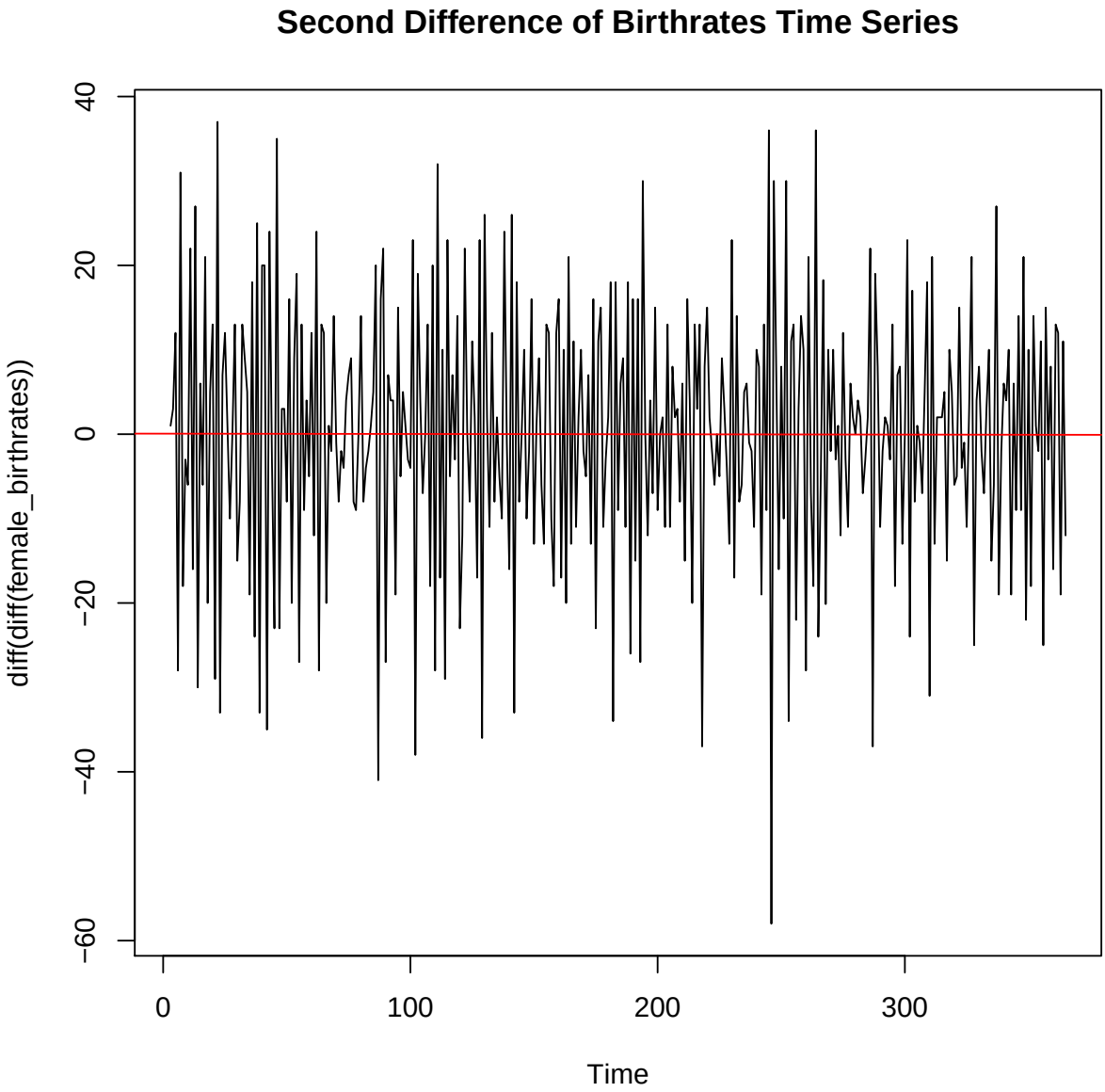
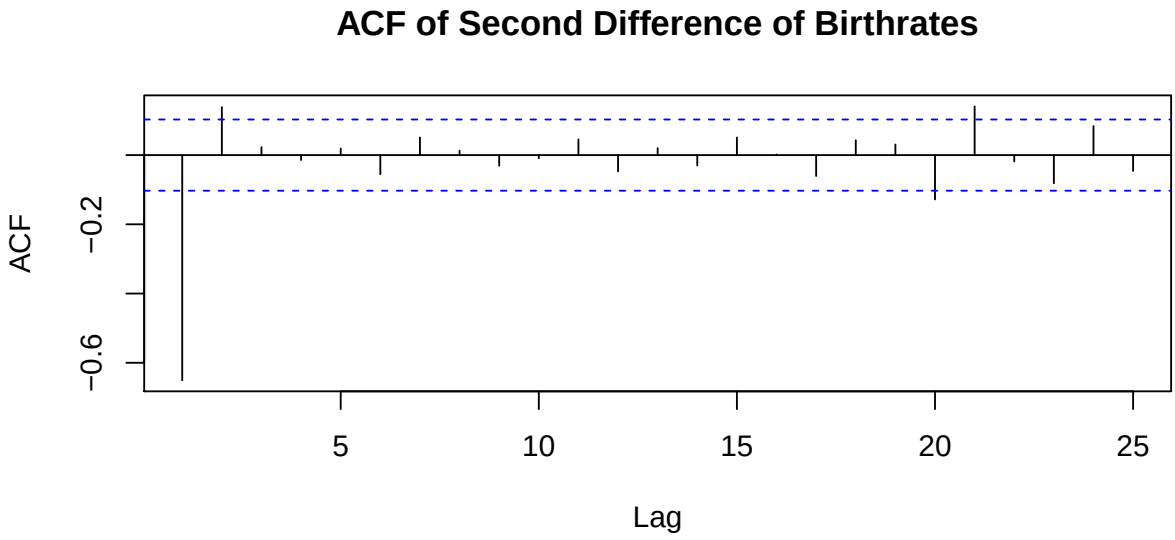


Figure 13: second_diff_ts



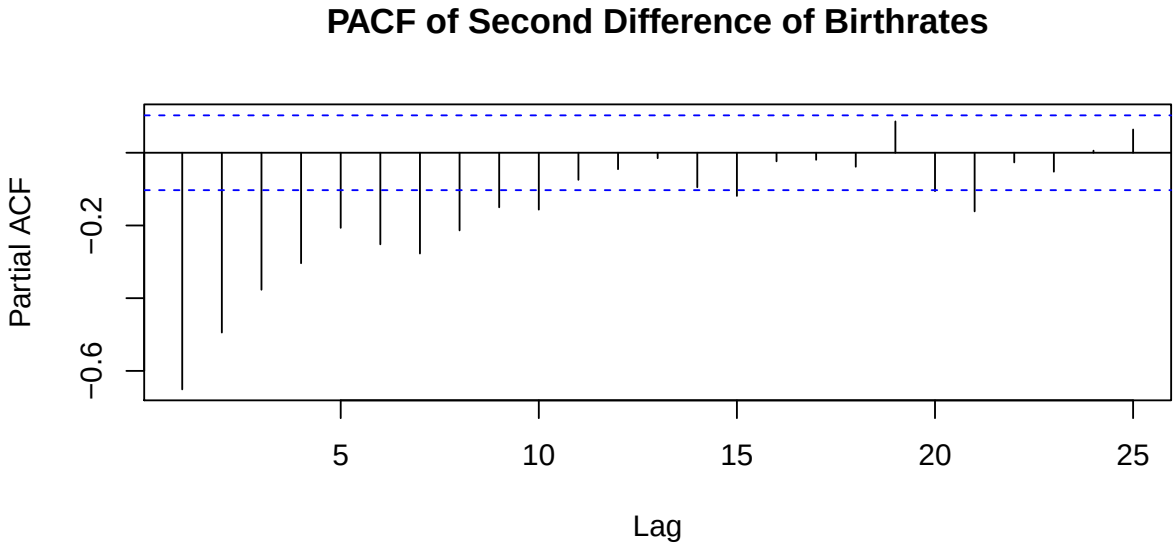


Figure 14: seconddiff_ts_acf_pacf

Forecast plots using the training data of the 2 ARIMA models

Below you can see the plot when I forecasted using the training set on the ARIMA(20, 1, 21) model:

And now below you can see the plot when I forecasted using the training set on the ARIMA(7, 1, 1) model:

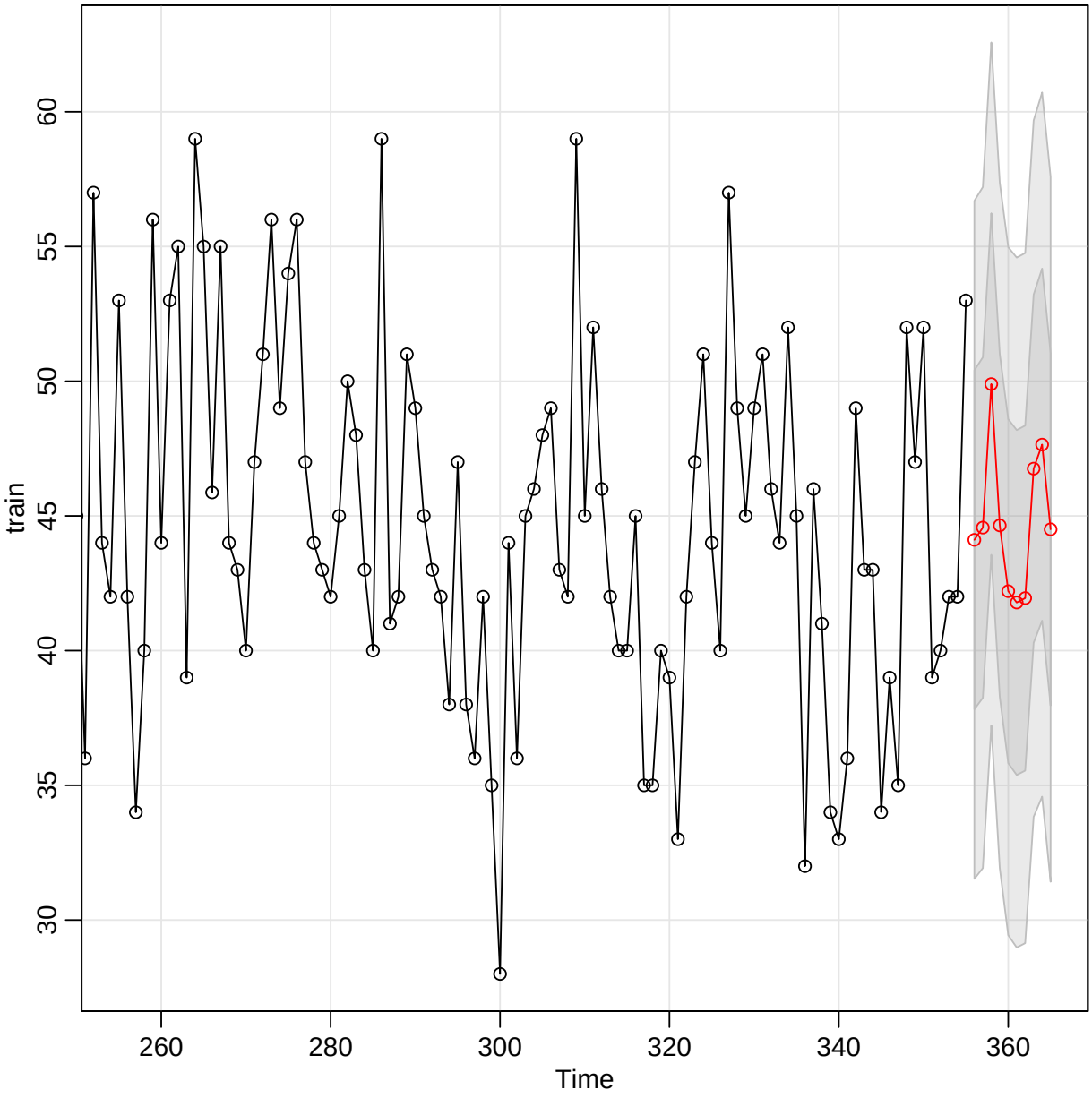


Figure 15: train_arima_forecast

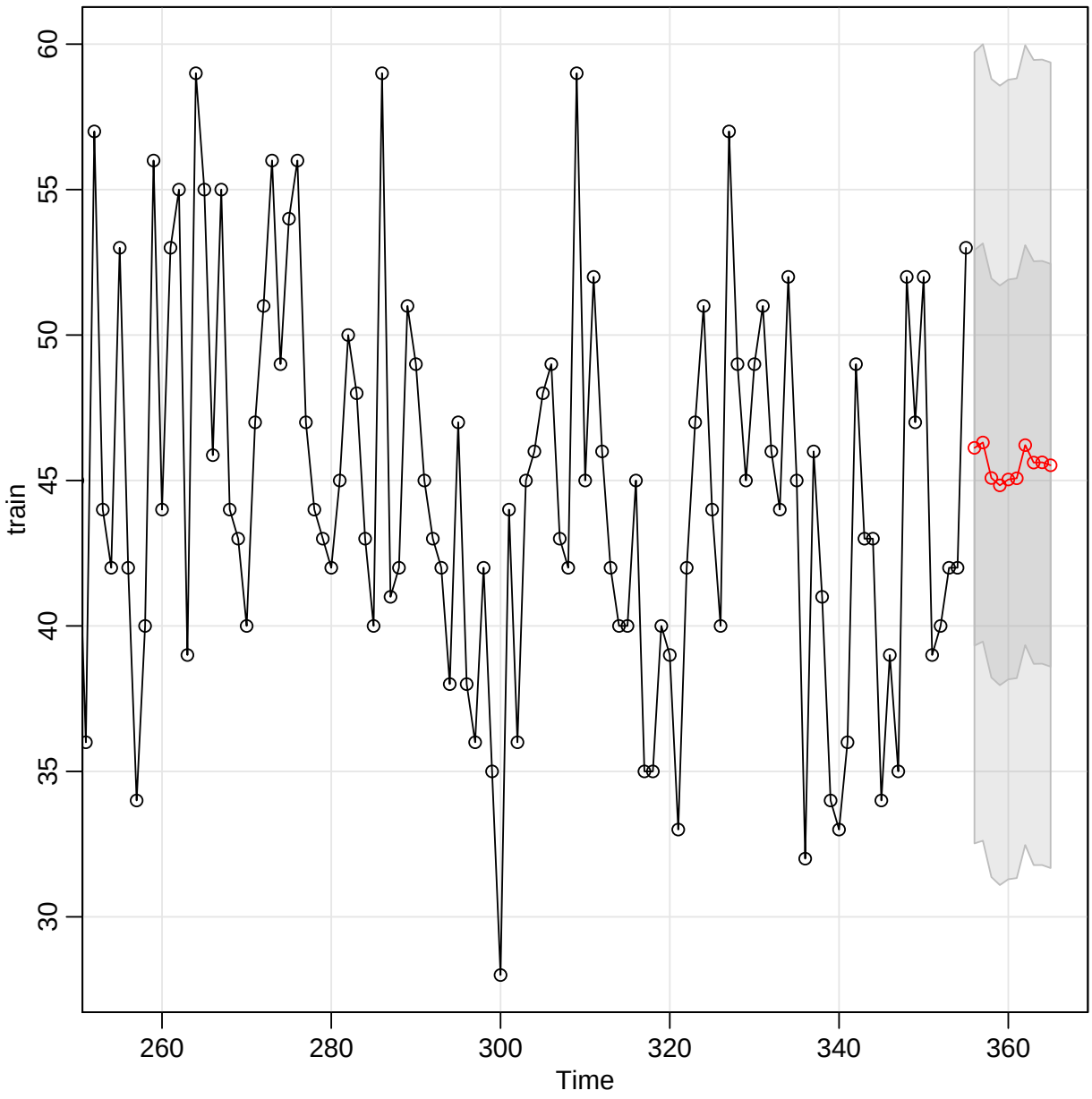


Figure 16: train_second_arima_forecast