Modeling Monthly CPI Percent Change in Norway using Time Series Models with Covariates

Couchot, S. Lagresle, V. Mahmood, Y.

November 17th, 2023

Abstract

In our analysis, we explored three covariates and applied two time series models to accurately forecast the Consumer Price Index (CPI). Our aim was to identify the optimal model and influential covariates for precise forecasting. Through a comprehensive evaluation considering visual analyses, covariances via heatmaps, ACF- and PACF-plots, parameter significance, and AIC, we selected the GARCH(1,1) model with the covariates "Index of household consumption of goods" and "Producer Price Index." These were chosen for their financial significance to CPI fluctuations. Our key finding is in the obtained confidence intervals, providing a likely range for anticipated CPI values. Overall, our study contributes to a clearer understanding of modeling considerations for the complex system of CPI forecasting.

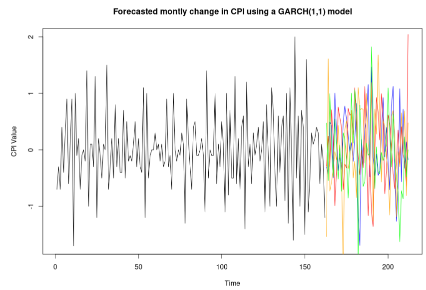
Introduction

In times of economic uncertainty, understanding and predicting inflation trends are crucial for making informed decisions. This project offers a straightforward approach to model and forecast the monthly percent change in the Norwegian Consumer Price Index (CPI) using time series models, specifically ARIMA or GARCH, with up to three covariates, contingent on their significance.

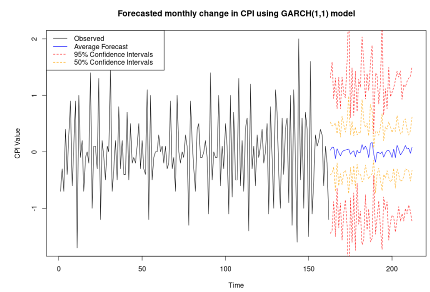
Results

The model used for forecasting is a GARCH(1,1) model given by:

$$\begin{aligned}\sigma_t^2 =& 0.52675728 + 0.45108092X_{t-1}^2 - 0.57416461\sigma_{t-1}^2 \\ &- 0.06459942X_{2,t-1}X_{t-1} - 2.81344979X_{3,t-1}X_{t-1}\end{aligned} \tag{1}$$



(a) Four realizations of the forcast



(b) Confidence ranges and mean forecast

Figure 1: Forecasting using GARCH(1,1) model

Only using the covariates "Index of household consumption of goods" and "Producer Price Index".

Discussion

Data Exploration and Visualization

Refer to the appendix for details on data sourcing.

Visualizing the data, we observe that the CPI data (2) exhibits no noticeable trend, resembling white noise, as supported by the histogram (10). Covariates show weak trends; unemployment (6) has a weak seasonal trend resembling a sine function, and Commodity Consumption (4) exhibits an increasing linear trend amid volatility. The Producer Price Index data has a weak seasonal trend and an unexplained spike at the end.

We explore covariate interplay, emphasizing their influence on each other and the CPI, through a covariance heatmap (16). Notable covariance is observed between Commodity Consumption and the Producer Price Index, and the other covariates does not seem to have any covariance, it is zero.

A near zero covariance suggests the absence of a consistent linear association among the variables. It's crucial to note that the absence of covariance does not necessarily mean the absence of all relationships. Although linear relationships might not be evident among the data, non-linear or more intricate dependencies may still be present.

Data Transformation

We aim to apply transformations to our for several reasons. The main reasoning behind this is to ensure that our data is stationary (mean and variance are constant over time). This assumption is especially important for the ARIMA model, as it assumes a stationary mean. The GARCH model on the other hand, does not require the underlying time series to be stationary.

Another reason to transform the data is to remove trends, seasonality, and most importantly, normalizing the data. Normalizing can prove helpful, as many statistical methods assume normality.

Determining the appropriate data transformation in time series analysis is not always obvious. Choosing the correct transformation often requires careful examination of the data. As we have done our data analysis and seem weak trends of linearity and seasonality, we will log-difference our data. It involves taking the natural logarithm of the data and then differencing the logarithms. This transformation is used to stabilize the variance and remove trends and seasonality. We want to stabilize the variance in case we are going to use the GARCH model, and we want to remove trends and seasonality to ensure stationary data.

Post-transformation, data plots (3, 5, 7, 9) and histograms (11, 14, 13, 15) indicate near-normal distribution, confirming the success of our transformation.

The heatmap (16) reveals clearer positive covariance between covariates and CPI, implying a linear relationship. This strengthens assumptions for further analysis, allowing us to consider the CPI as linearly dependent on our covariates, allowing us to model the CPI linearly in terms of our covariates.

Parameter Estimation

Before we can choose which model to use, ARIMA or GARCH, we need to determine their parameters. A detailed description of the models and their parameters is given in the appendix, under the section "Model Description".

We initiate the estimation process by determining p, d, and q. Determining d is relatively straightforward, set at 1 as it corresponds to the degree of differencing applied to our data.

The estimation of p and q is more intricate. There are many ways to estimate these, but one common way of doing it is by studying the ACF (AutoCorrelation Function) and PACF (Partial AutoCorrelation Function) plots of the data. The patterns revealed in these plots offer insights into suitable values for p and q.

The ACF helps identify the moving average (MA) component (q) and the PACF helps identify the autoregressive (AR) component (p).

The ACF functions for the CPI data 18, and transformed CPI data 19 suggest that the MA component should be set to 1. This is because the ACF has a sharp decline at lag 1, and remains closely to, or within the confidence level. This is true for both the transformed and not transformed data. Therefore set q = 1

The PACF functions for the CPI data 26 and transformed CPI data 27 seems to be within the confidence level from the start. This is true for the not transformed data. Looking at the transformed data, we see that after 1-2 lags, the value of the PACF is within and close to the significant level. Therefore set p = 1.

For the other parameters to be estimated mentioned in the appendix, their estimated coefficients, standard errors and their significance is given in 3 for the GARCH and 2 for the ARIMA. These are estimated by optimizing (minimizing) a likelihood function. The inbuilt optim() function for this was used in R. In this function there is an alternative which returns a hessian matrix. Inverting this hessian matrix, and taking the root of the diagonal of this matrix, gives the standard errors.

Diagnostics and Model Selection

To choose between GARCH(1,1) and ARIMA(1,1,1) models, we use AIC, a metric balancing model complexity and fit quality. We select the model with the smallest AIC value. AIC values in 1 guide the choice, favoring the GARCH(1,1) model. This conclusion make sense regarding the data exploration, as the CPI did not show any trend, but showed a lot of volatility, and then the forecasting will be more about describing the volatility of the data in a meaningful way.

Which covariates we should look at in our analysis is given by a hypothesis test. 3 gives the p-values for the different parameters in the GARCH model. With the significance level set at $\alpha = 0.5$, the p-test is taken to a 95% confidence level.

From this we see that only the parameters β_2 and β_3, corresponding to the data Commodity Consumption and Producer Price Index should be taken into account. According to our analysis, unemployment does not have a significance effect on the CPI.

The final model is hence given by 1

where X is the differenced CPI data, X_2 is the Commodity Consumption data, and X_3 is the Producer Price Index data.

Inflation Rate Forecasting

Using the method of innovations, we forecast 162 time steps in the future for the CPI data, and run 100 different realizations. We plot 4 of these runs in 1a, where the first 50 forecasted values are plotted.

The condition variance is calculated, which is then used to calculate the innovations. The innovations are then simulated from a normal distribution (as our data is approximately normal), with mean 0 and standard deviation $\sqrt{\sigma_t^2}$. These are then our forecasted values. The variability in our estimations is a result of how the innovations are estimated.

From the 100 runs, we can calculate the confidence intervals, and the average estimation. In **??** we have plotted two different confidence intervals, and the average forecast. The 95% confidence interval is of must interest for our forecast.

From a visual perspective, the forecasted values appear valid as they align with the patterns of the preceding data and do not disrupt the overall "trend". The primary significance lies not in the predictions alone, as the data is so unpredictable and volatile, but rather in the confidence intervals. These intervals provide a probable range within which the CPI values are expected to fall.

Conclusion

The analysis involved the examination of three distinct covariates and two different time series models. The objective was to identify the optimal model and the most influential covariates for accurate CPI forecasting. The chosen model, determined through visual analysis of both raw and transformed data, covariances through heatmaps, assessment of ACF- and PACF-plots, consideration of parameter significance, and evaluation of AIC, was the GARCH(1,1) model. This model incorporated the covariates "Index of household consumption of goods" and "Producer Price Index," which, from a financial perspective, are deemed relevant to CPI fluctuations. Furthermore, the selection of the GARCH model is justified by the high volatility in the CPI data, making it more meaningful to model volatility rather than the mean, which remains relatively constant. Consequently, the focal point of interest in the forecasting results lies not in the individual CPI values for each realization but rather in the obtained confidence intervals. These intervals offer a likely range within which the CPI values are anticipated to fall.

Use of ChatGPT

Our text was a bit long, so ChatGPT was used to shorten the text from around 7 pages to 5. Further, it was used to comment whole the code, and in debugging processes regarding the code.

Data Sourcing

The data is sourced from "Statistisk Sentralbyrå" (SSB), which is the primary authority responsible for compiling and disseminating official statistics in Norway. Hence, it is reasonable to infer that the data used maintains a high level of quality. The following outlines the procedure and links on which data is used, and how the data was retrieved.

- **Consumer Price Index:** Monthly change (per cent), From MO1 2010 to MO8 2023[1].
- **Labour force survey:** Unemployment (LFS) (1000 persons), From MO1 2010 to M08 2023, Not seasonally adjusted, both sexes[2].
- **Index of household consumption of goods:** Consumer Price Index (2015=100), From MO1 2010 to MO8 2023[3].
- **Producer Price Index:** Producer Price Index (2015=100), From MO1 2010 to MO8 2023[4].

Appendix

Calculation descriptions

ACF

The ACF is calculated by the formula below:

$$\rho_k = \frac{\sum_{t=k+1}^{T}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sqrt{\sum_{t=k+1}^{T}(X_t - \bar{X})^2 \cdot \sum_{t=k+1}^{T}(X_{t-k} - \bar{X})^2}} \tag{2}$$

[1]Link: https://www.ssb.no/en/statbank/table/03013/
[2]Link: https://www.ssb.no/en/statbank/table/13760/
[3]Link: https://www.ssb.no/statbank/table/05333/
[4]Link: https://www.ssb.no/en/statbank/table/12462/

PACF

The PACF is calculated by using the Durbin-Levinson algorithm:

$$\phi_{kk} = \frac{r_{k+1} - \sum_{j=1}^{k-1} \phi_{k-1,j} r_{k-j+1}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} r_{j+1}} \tag{3}$$

Model descriptions

ARIMA

An ARIMA(p,d,q) is generally written as:

$$\begin{aligned} Y_t =& \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} \\ &+ W_t - \theta_1 W_{t-1} - \theta_2 W_{t-2} - ... - \theta_q W_{t-q} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} \end{aligned} \tag{4}$$

Here, p, d, and q represent the orders of the autoregressive, differencing, and moving average components, respectively. These are to be determined.

θ and ϕ are the moving average and the autoregressive coefficients, respectively. These are to be estimated.

Y_t represents our time series data, and X_t represents our covariate data. It is evident that β corresponds to the coefficients of the covariates. We can assume linearity in this case, as observed in the heatmap shown in 17.

We write the ARIMA model in state space form, with state equation:

$$\mu_{t+1} = T \cdot \hat{\mu}_t$$

and the observation equation:

$$\epsilon_t = D - T \cdot \hat{\mu}_t - \beta_1 D_1 - \beta_2 D_2 - \beta_3 D_3$$

where μ is the mean, $\hat{\mu}$ is the predicted mean, T is the transition matrix, β are the coefficients of the covariates, D is the CPI data, and D_1, D_2 and D_3 is the covariate data.

It is reasonable to use mean as our model equation, as the ARMAX model is all about working with the mean. Secondly, the mean is updated using the prediction error, which reflects the model's belief about the evolution of the underlying process, and hence, the prediction error is reasonable to take as our observation equation.

We use the Kalman filter to update our estimates and their uncertainty. The Kalman Filter is to provide an optimal estimate of the true state of a system, given uncertain and noisy observations. This goes well with the state space representation of our model, as the state equation is an unobserved true state of the system, and the observation equation is related to the true state of the model.

The Kalman gain is an important step in the filter, that determines how much the new observation should be incorporated into the updated estimate. If the prediction is highly uncertain, the Kalman gain will be small, and the new observation will have less influence on the updated state estimate, and conversely.

GARCH(p,q)

An GARCH(p,q) is generally written as:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + \sum_{k=1}^{3} \gamma_k X_{k,t-1} \epsilon_{t-1} \tag{5}$$

ω is the beseline term, or constant level of volatility in the time series. This is to be estimated.

α_i are the ARCH coefficients, that represent the impact of the past squared residuals (ϵ_{t-i}^2). on the current conditional variance. This is to be estimated.

β_j are the GARCH coefficients, that represent the impact of the past conditional variances (σ_{t-j}^2) on the current conditional variance. This is to be estimated.

It is evident that γ corresponds to the coefficients of the covariates. We can assume linearity in this case, as observed in the heatmap shown in 17.

The GARCH model is all about modelling the variance of the data, and is good for modelling volatile data. The model is hence very simple, and all we need to to is to calculate the conditional variance for each subsequent time point using the GARCH model equation. A general representation of this model equation is given above in σ_t^2

Model Comparison

As a generalization of those two models, ARIMA models focus on modeling the mean of the time series, while GARCH models focus on modeling the volatility. They can be complementary and are often used together in time series analysis. ARIMA is used for capturing the underlying trend and seasonality, while GARCH is used for modeling the volatility around that trend.

In our analysis, we are using GARCH. This conclusion make sens with the comments on the CPI data made on the data exploration section. Indeed, there were not really a main trend to catch with this dataset and the prediction work was more about predicting volatility than trend. It is therefore logical that the AIC score of the GARCH model is more than twice lower than the ARIMA AIC score.

Moreover, ARIMA asumes a linear relationship in the data, which may not be appropriate for all time series, especially CPI. Consequently, it may not perform well on data with complex nonlinear patterns or changing volatility.

Finally, let's remind that GARCH models have been specifically designed to model and forecast the volatility of financial time series data. Inflation rate seems to be close to financial data, so again it makes sens that GARCH model is the best.

Figures

Model	**AIC**
ARIMA(1,1,1)	337.6094
GARCH(1,1)	23.37445

Table 1: AIC Values for ARIMA(1,1,1) and GARCH(1,1)

Parameter	**Coefficient**	**T-Statistic**	**P-Value**	**Significant**
phi	2.907574e-01	0.16536139	8.688736e-01	FALSE
theta	4.566159e-05	-0.70499405	4.818652e-01	FALSE
sigma2	4.001808e-01	5.74062446	4.781237e-08	TRUE
beta1	3.191171e-01	0.49164410	6.236620e-01	FALSE
beta2	-1.932574e-01	-5.17587122	6.893005e-07	TRUE
beta3	2.626465e-02	-0.05350794	9.573957e-01	FALSE

Table 2: ARIMA(1,1) Model Parameters with uncertainty

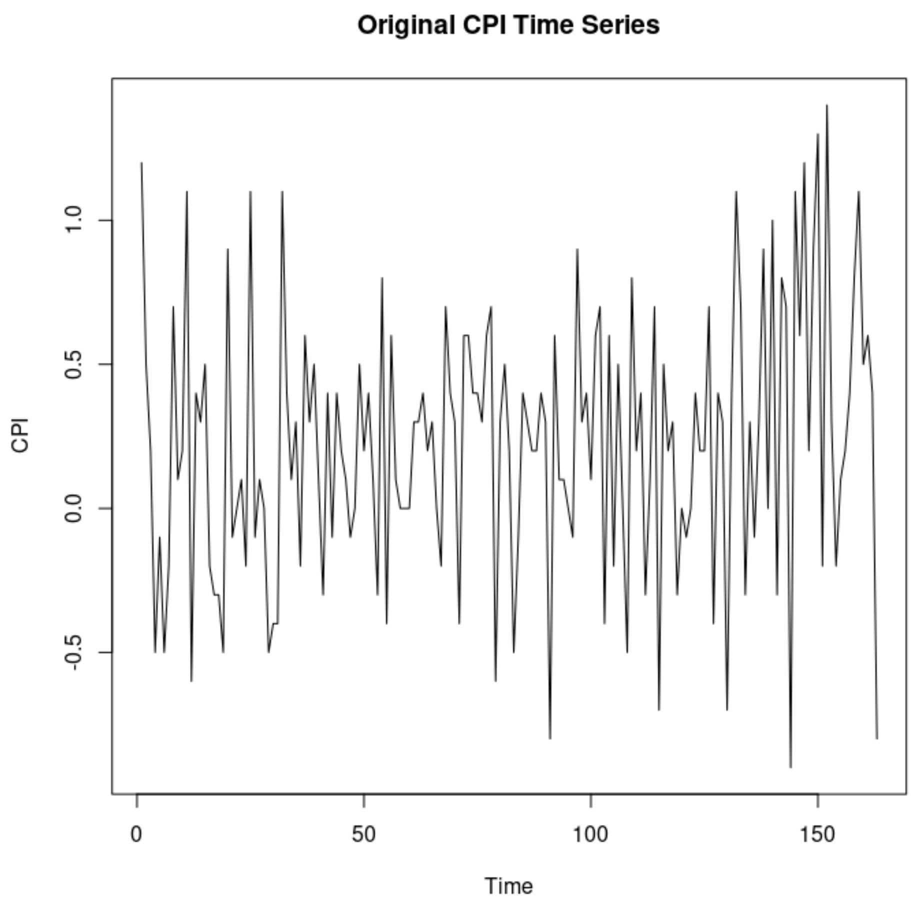


Figure 2: Plot of CPI data

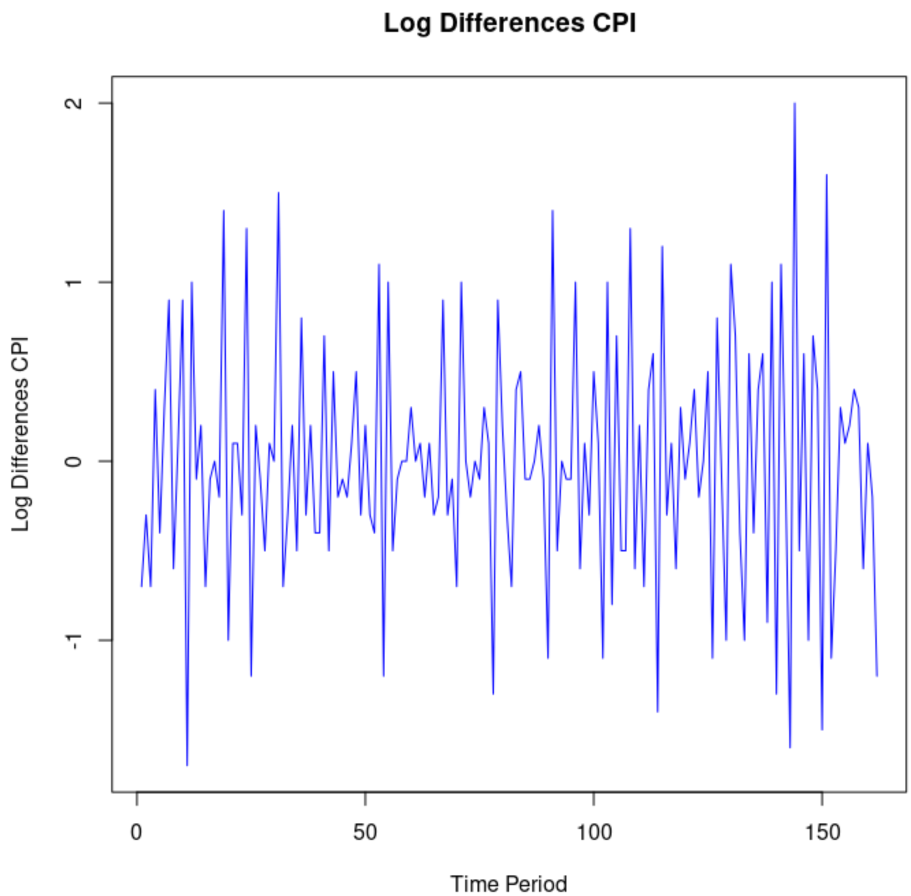


Figure 3: Plot of Transformed CPI data

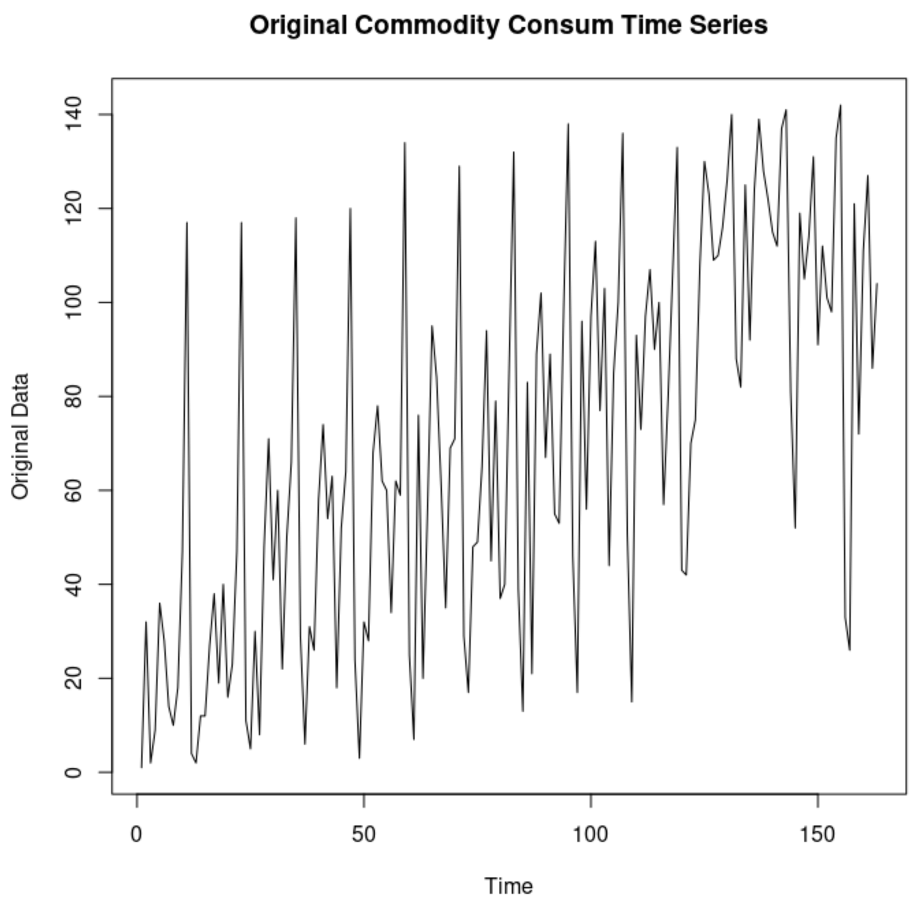


Figure 4: Plot of Commodity Consumption data

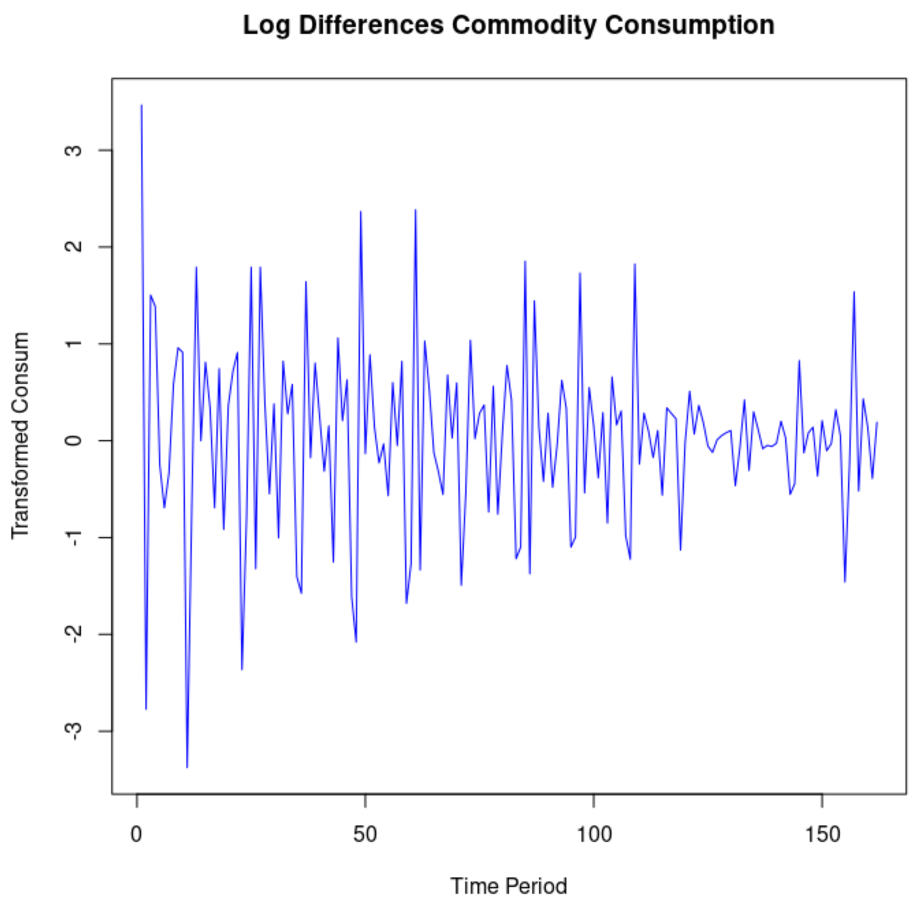


Figure 5: Plot of Transformed Commodity Consumption data

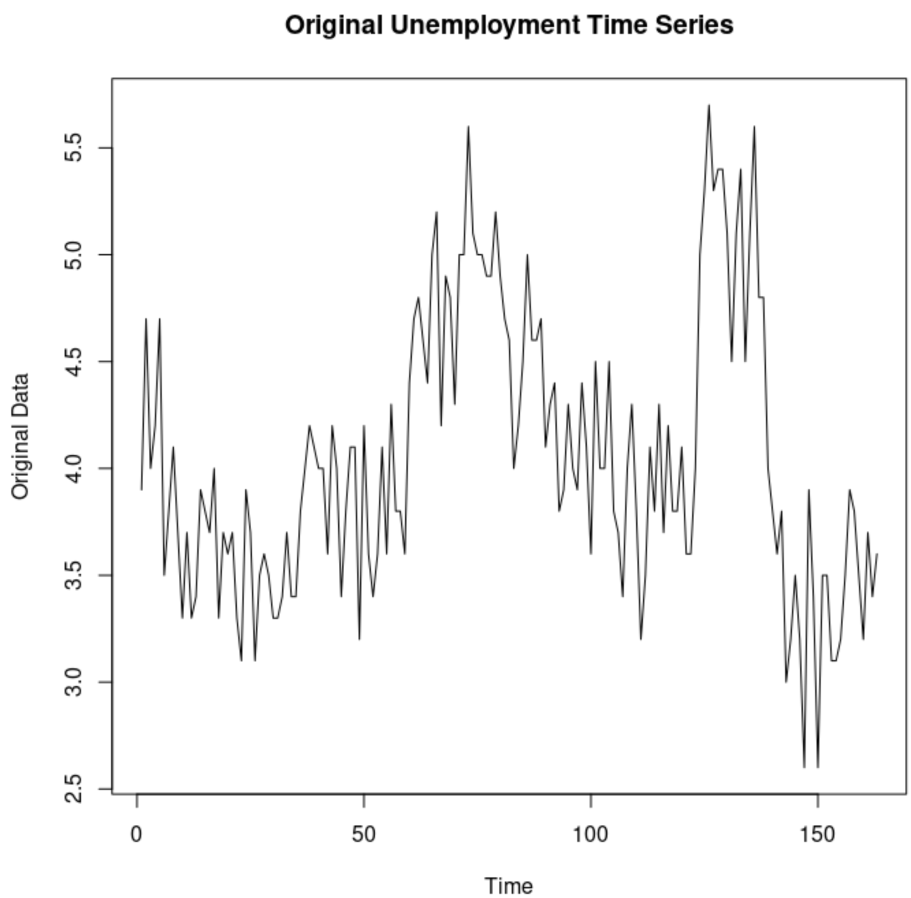


Figure 6: Plot of Unemployment data

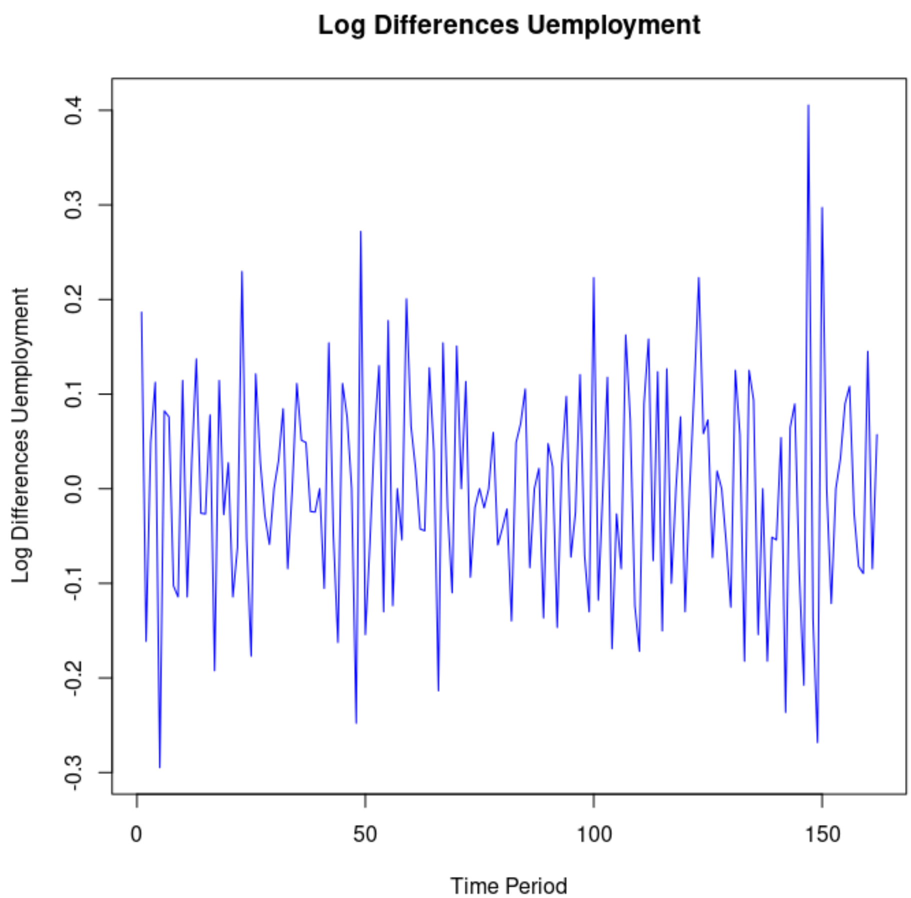


Figure 7: Plot of Transformed Unemployment data

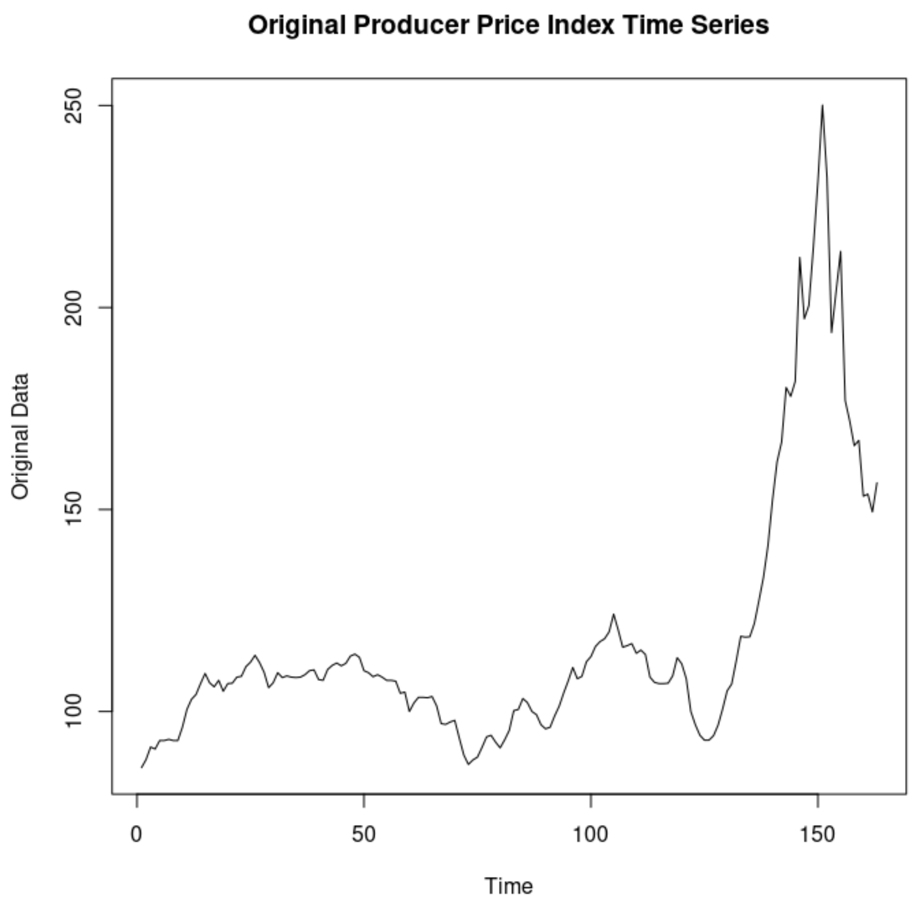


Figure 8: Plot of Producer Price index data

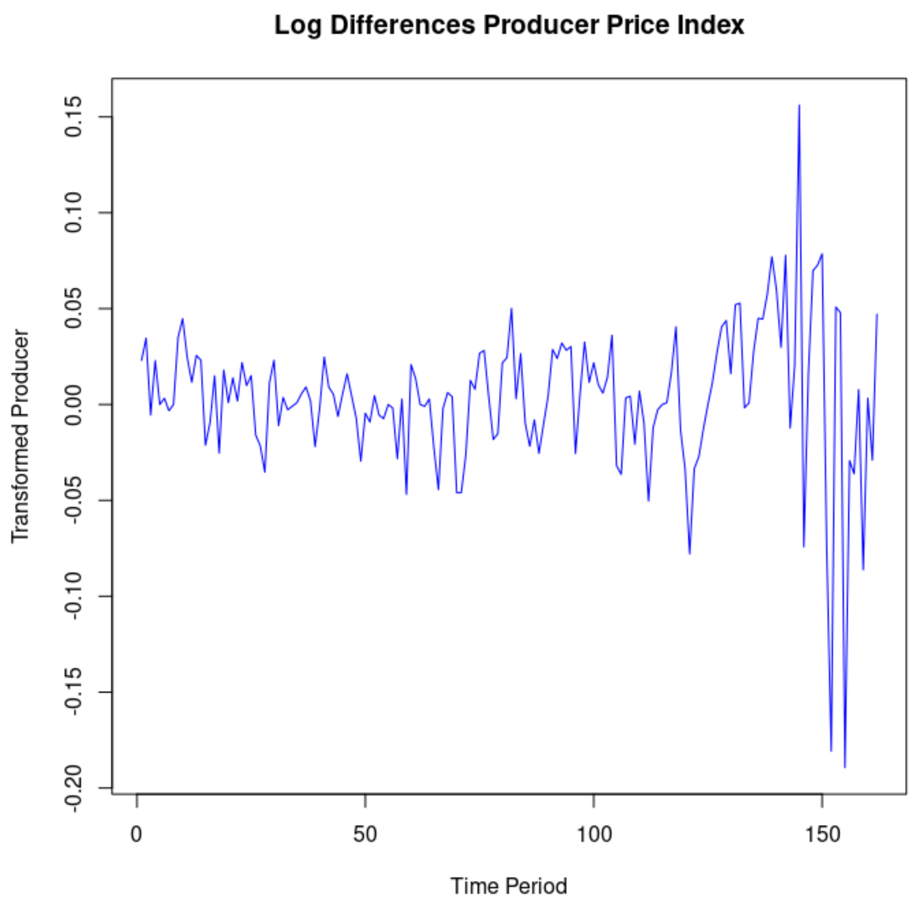


Figure 9: Plot of Transformed Producer Price index data

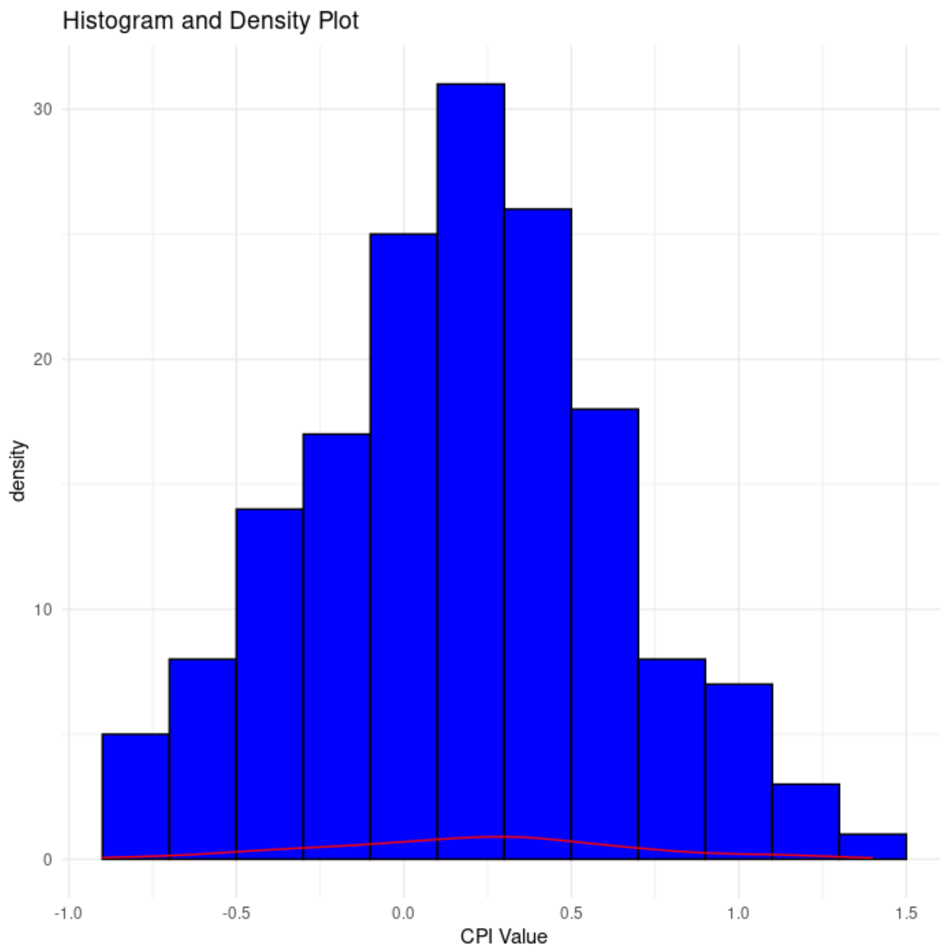


Figure 10: Distribution for CPI data

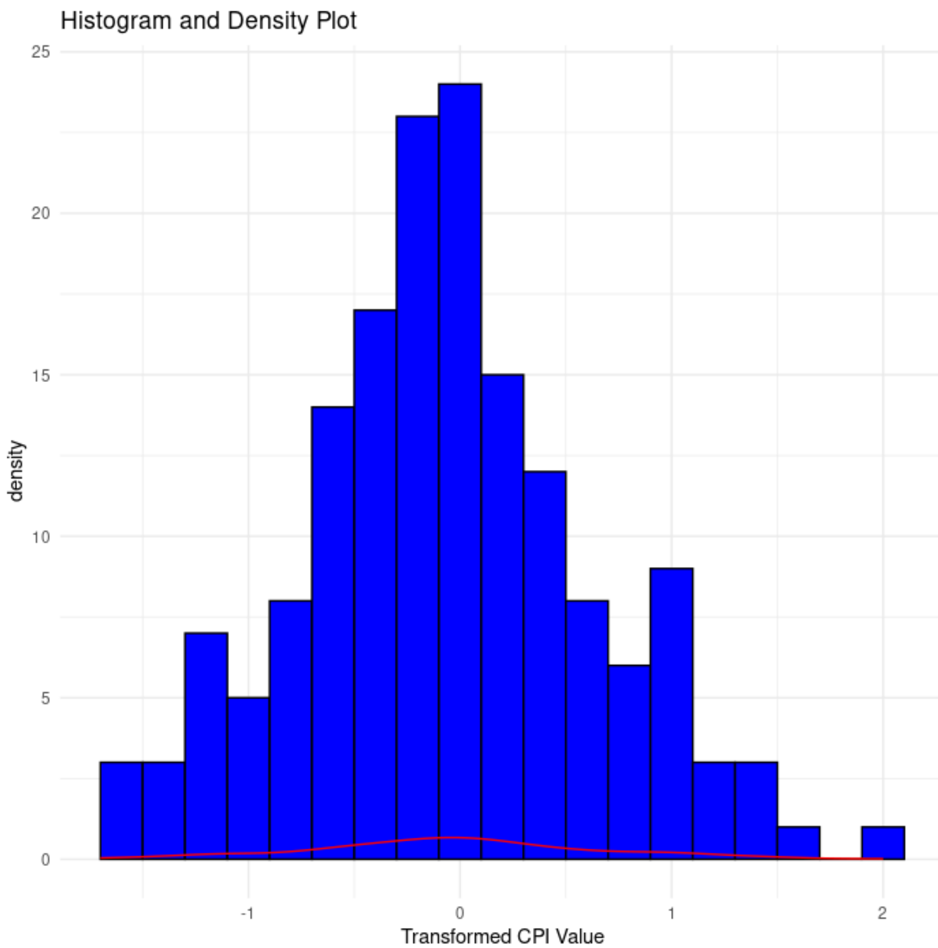


Figure 11: Distribution for Transformed CPI data

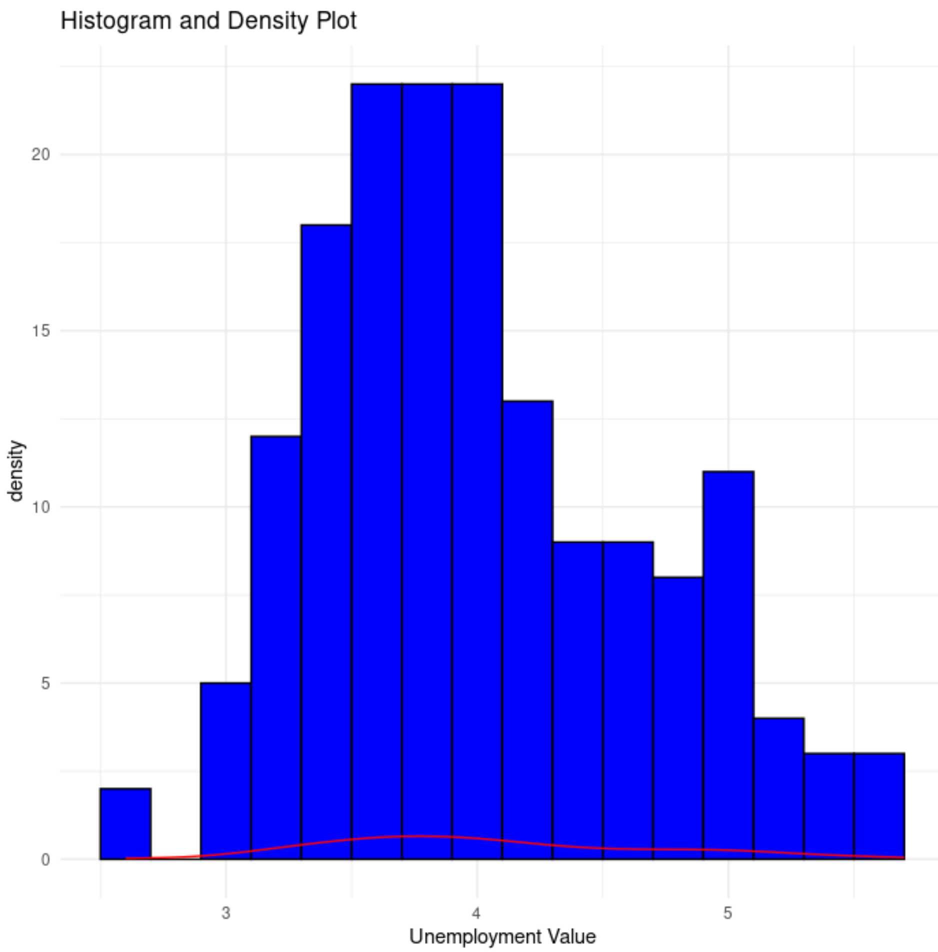


Figure 12: Distribution for Unemployment data

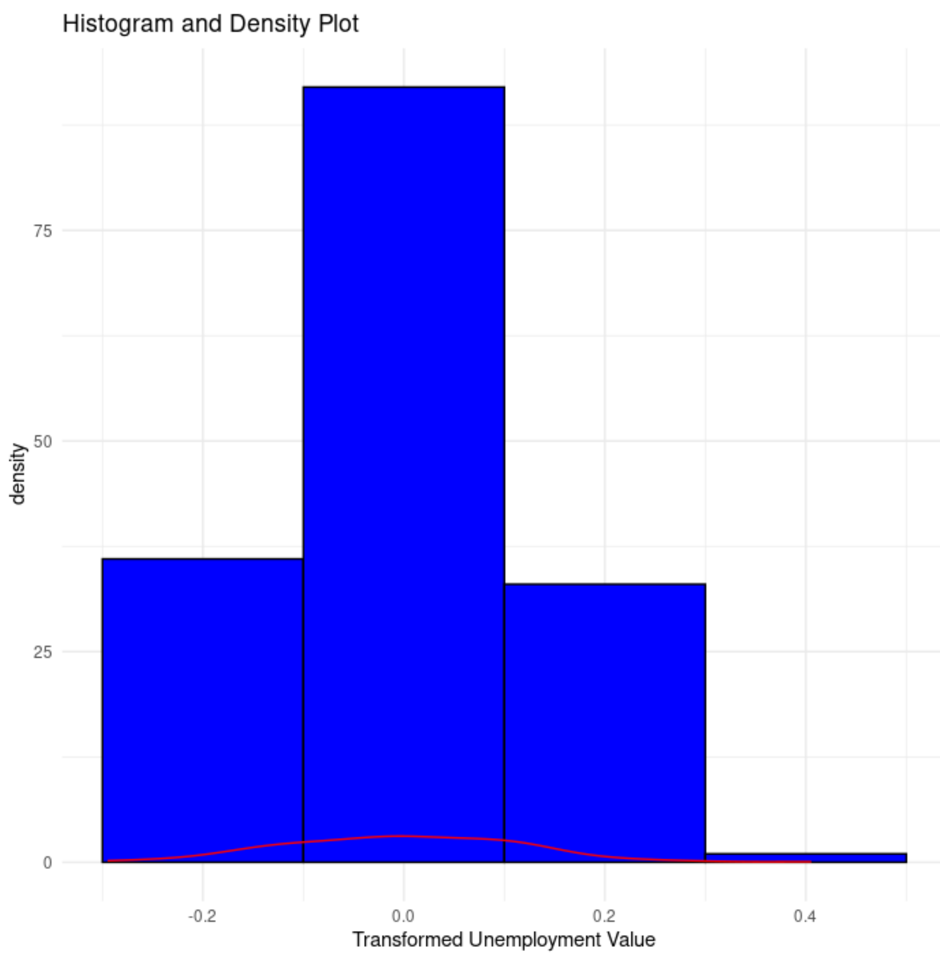


Figure 13: Distribution for Transformed Unemployment data

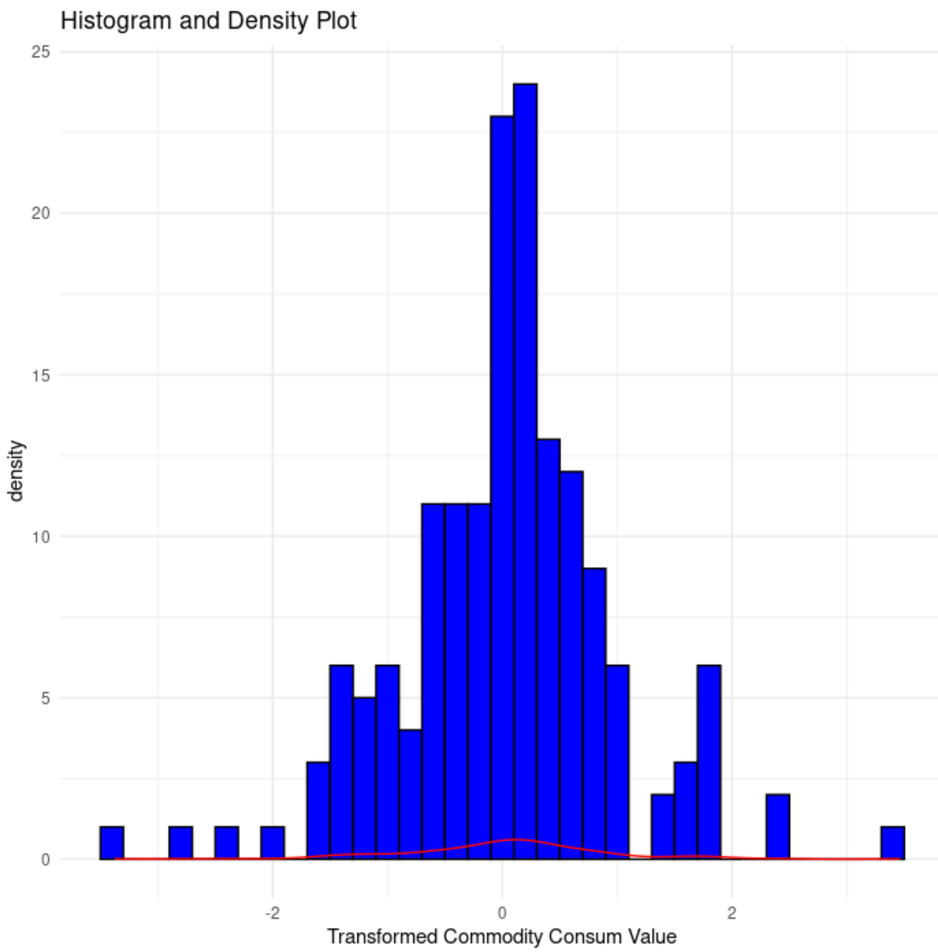


Figure 14: Distribution for Transformed Product Consumption data

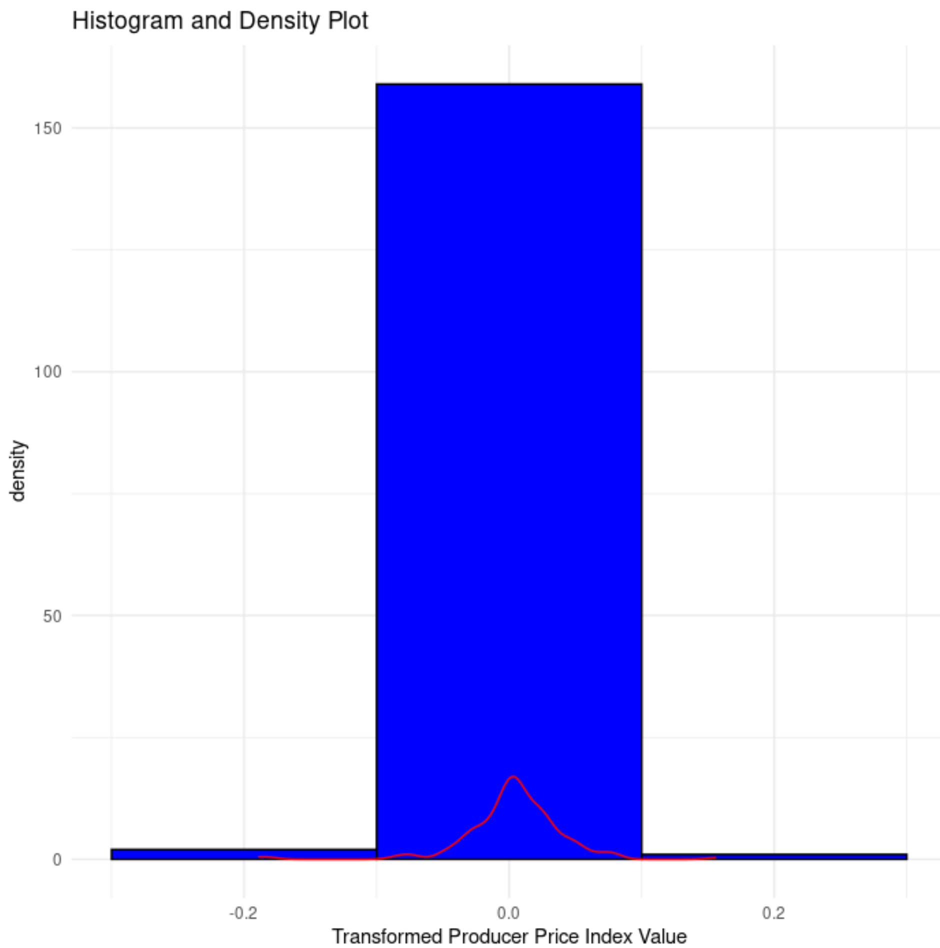


Figure 15: Distribution for Transformed Producer Price Index data

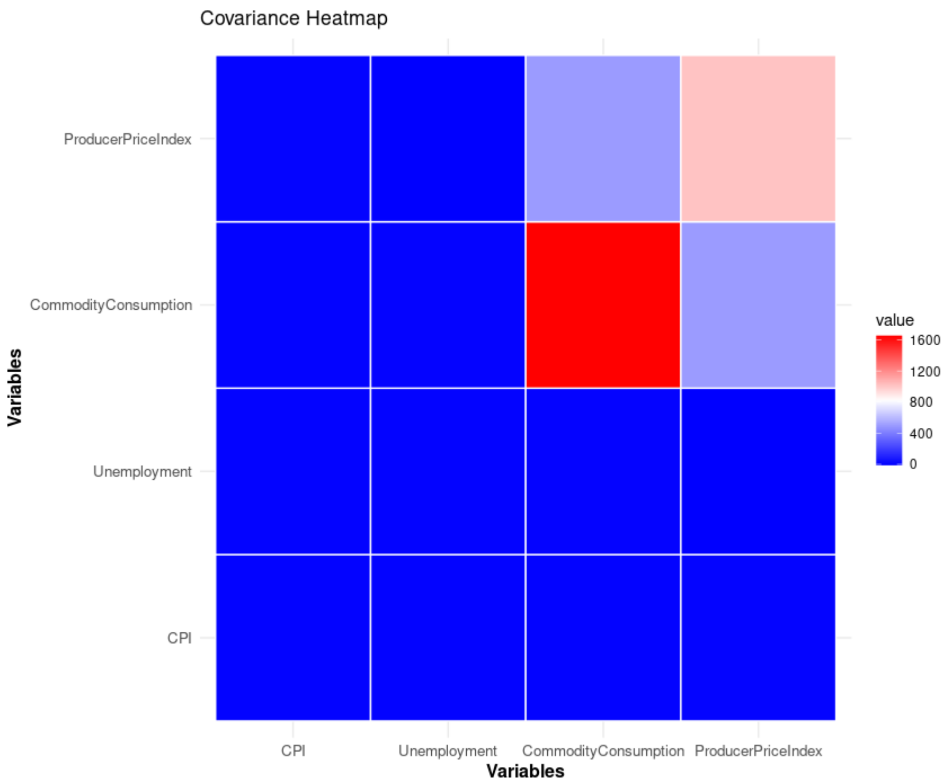


Figure 16: Covariance heatmap for data

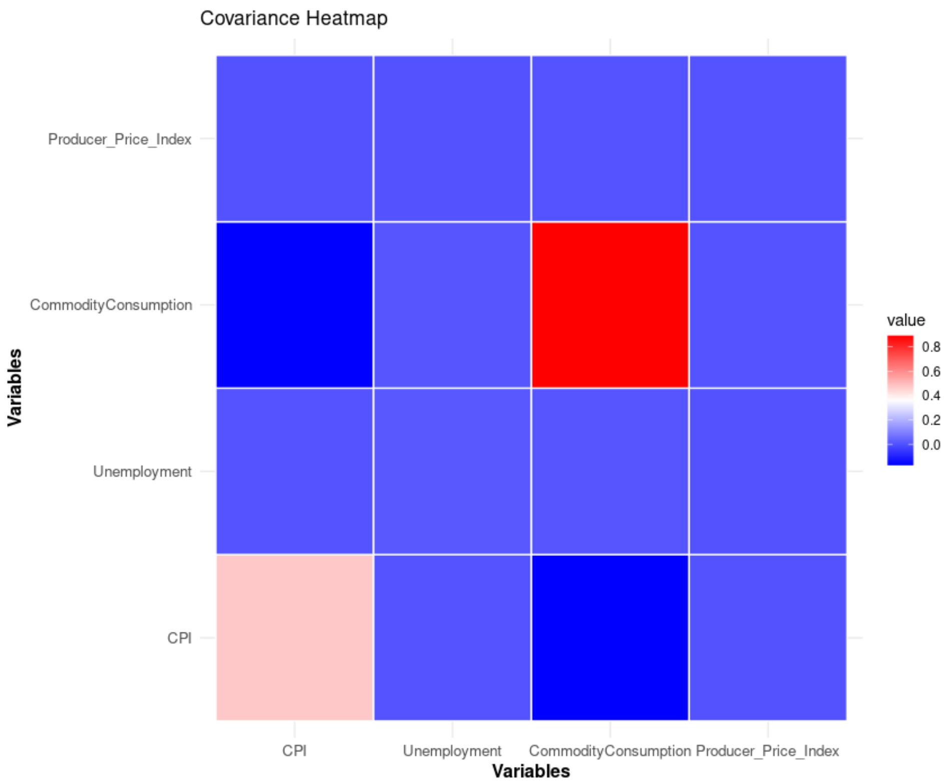


Figure 17: Covariance heatmap for transformed data

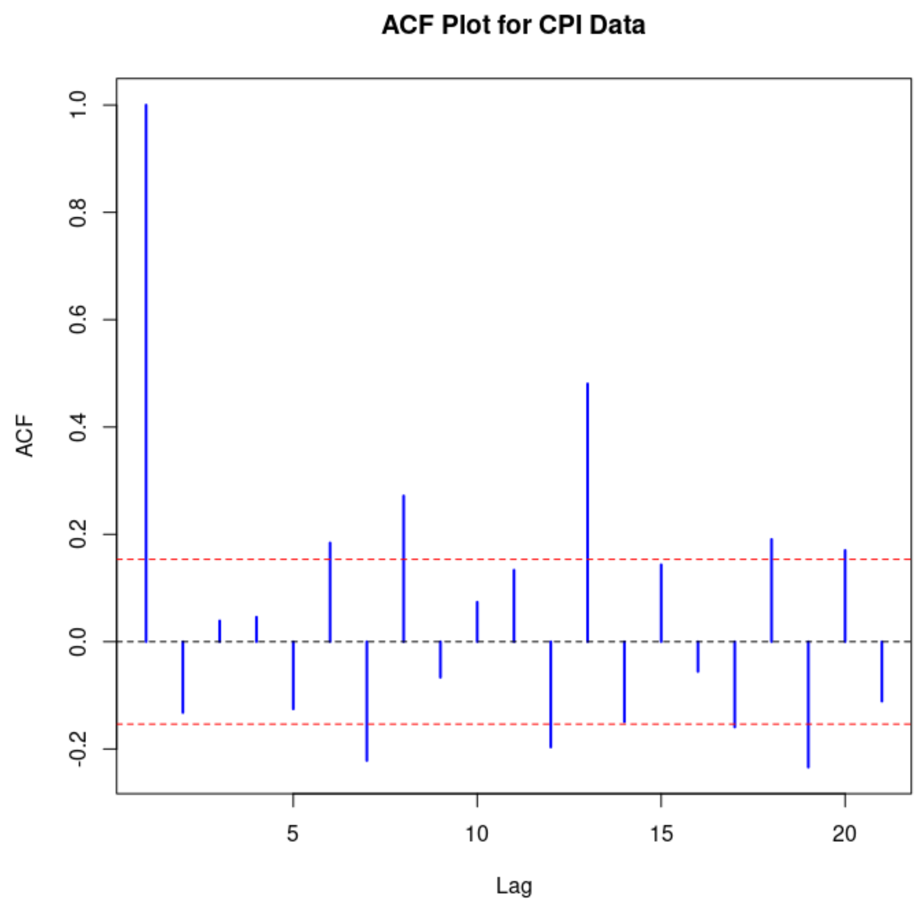


Figure 18: ACF for CPI data

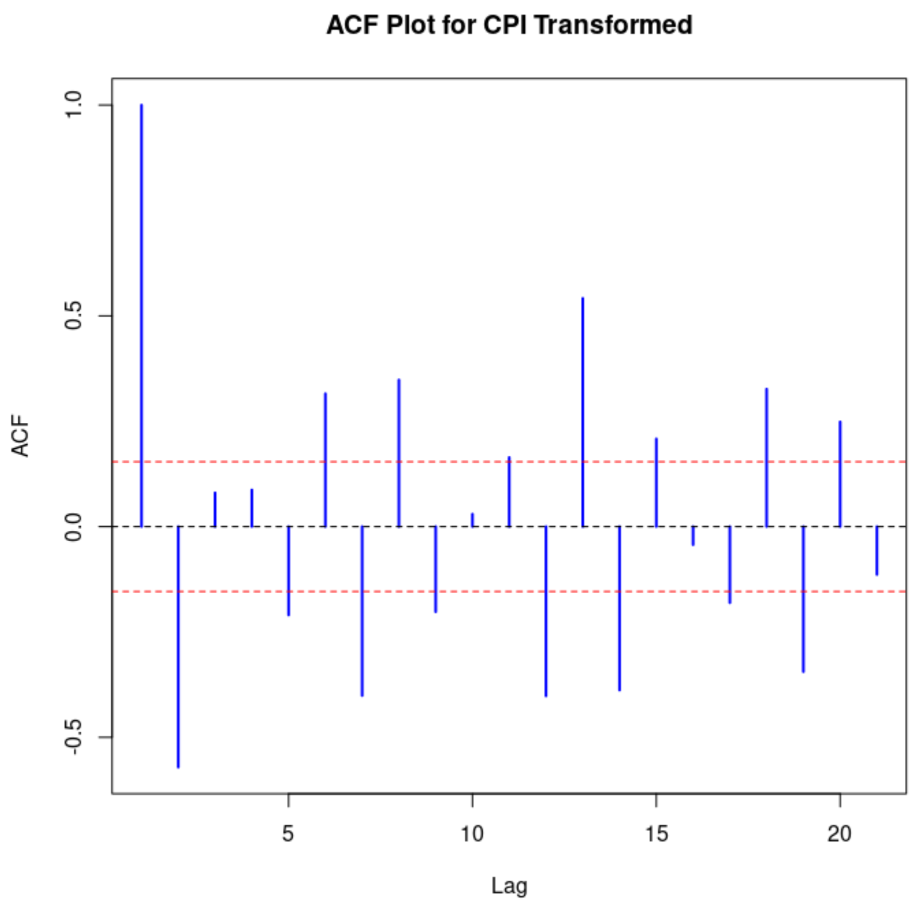


Figure 19: ACF for Transformed CPI data

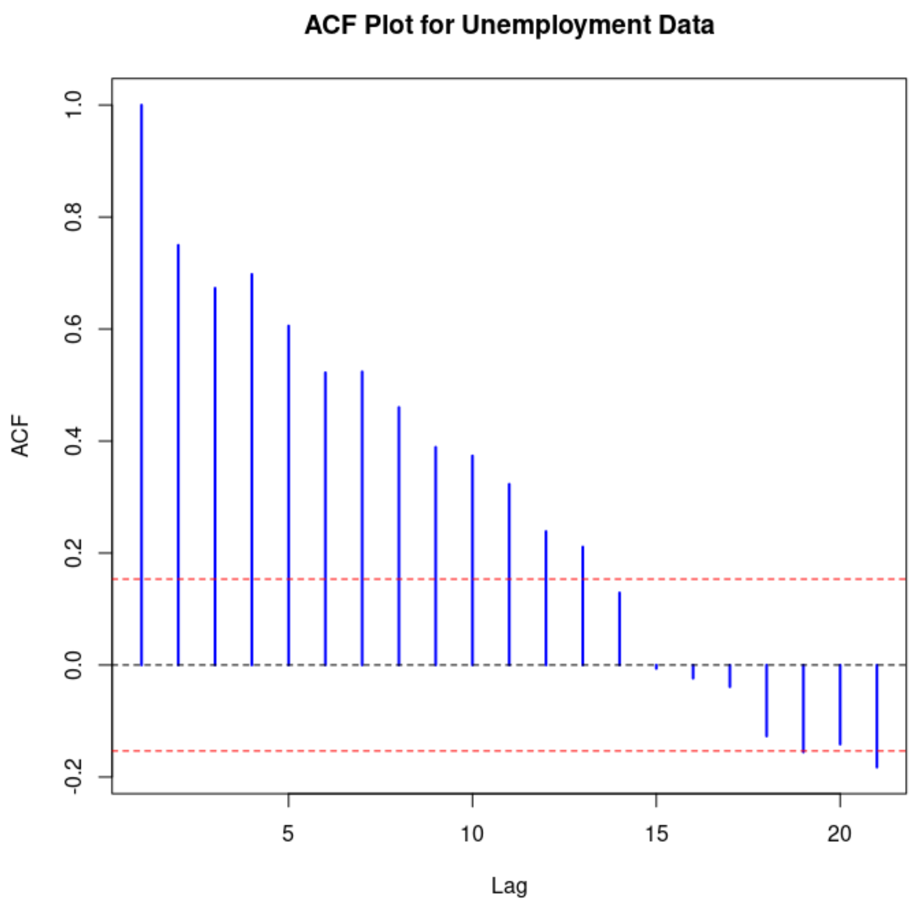


Figure 20: ACF for Unemployment data

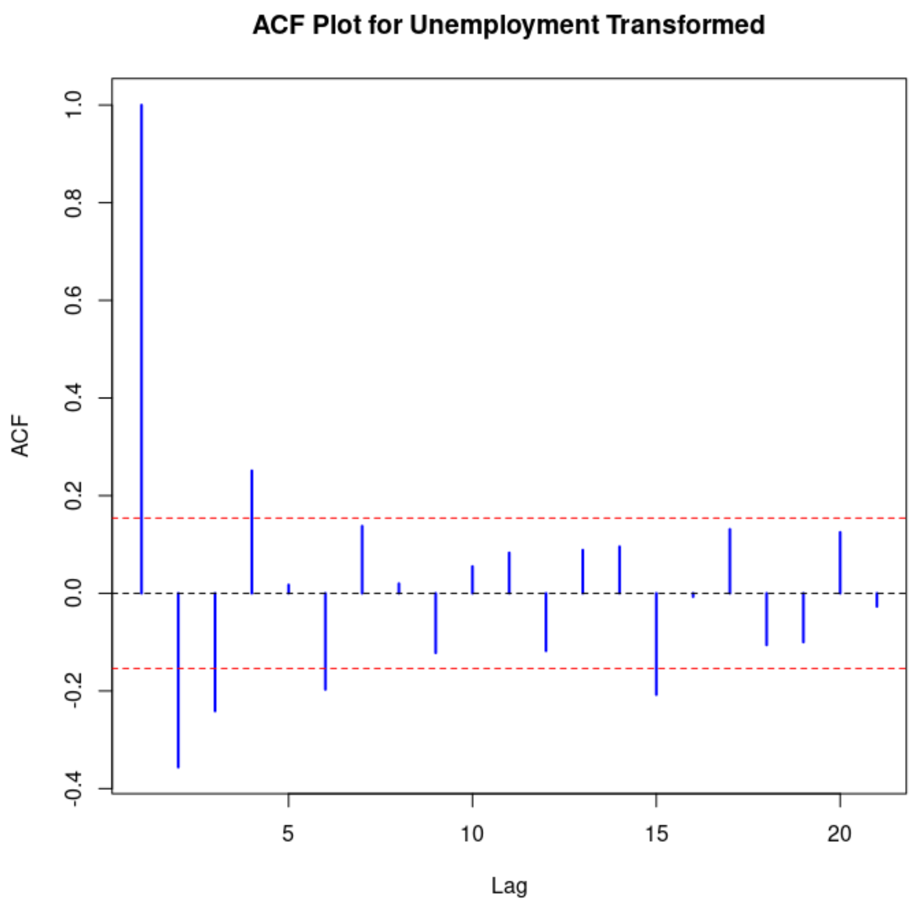


Figure 21: ACF for Transformed Unemployment data

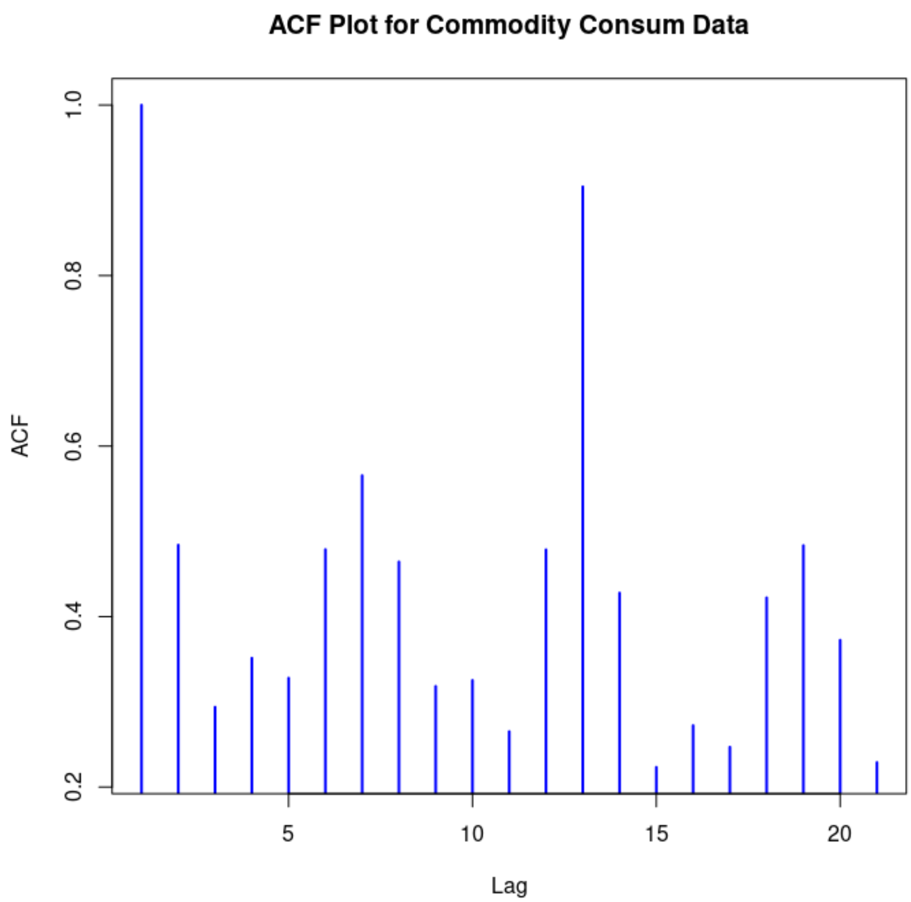


Figure 22: ACF for Product Consumption data

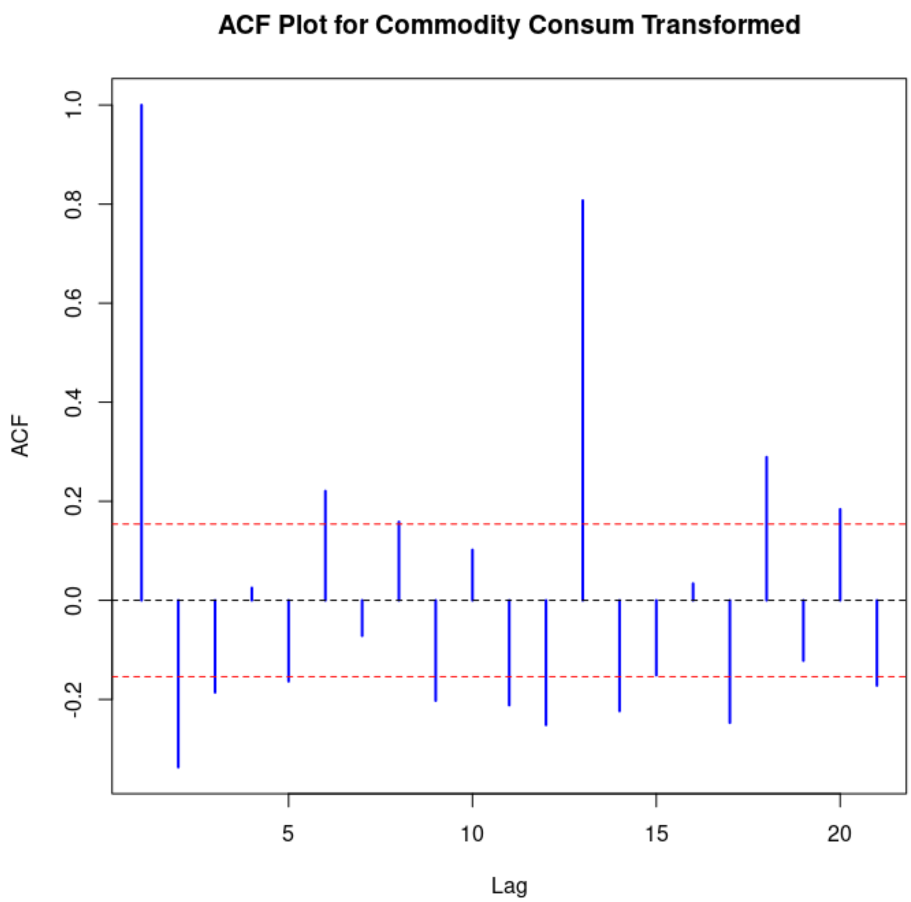


Figure 23: ACF for Transformed Product Consumption data

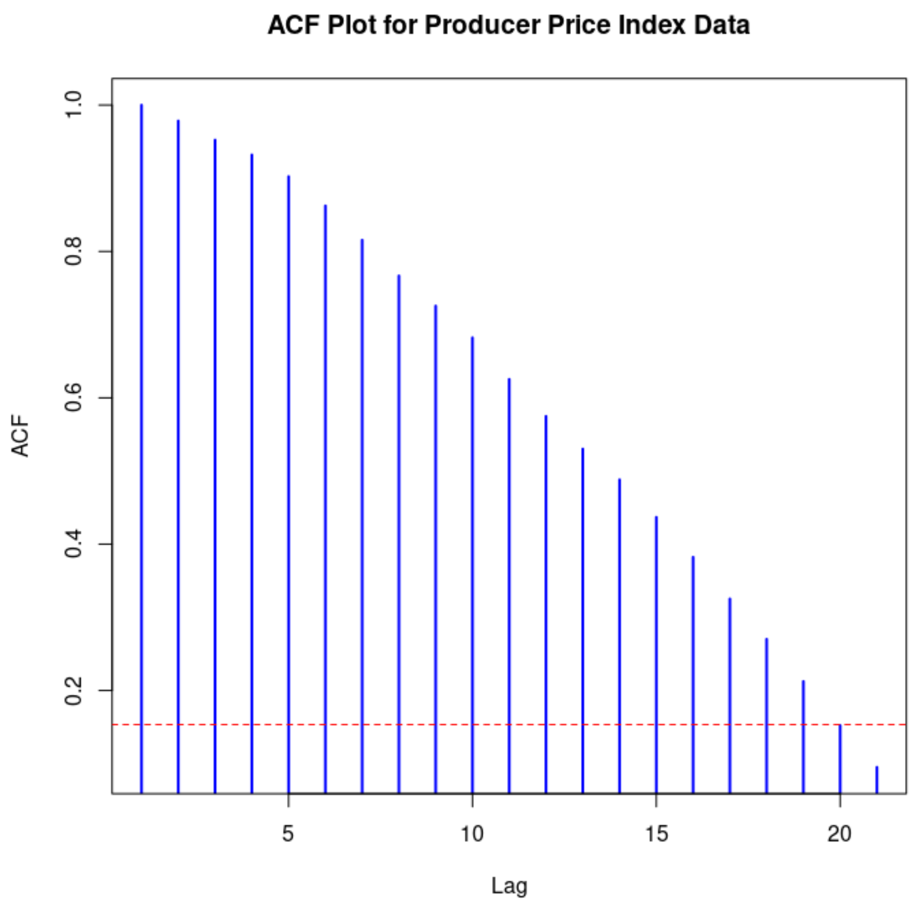


Figure 24: ACF for Producer Price Index data

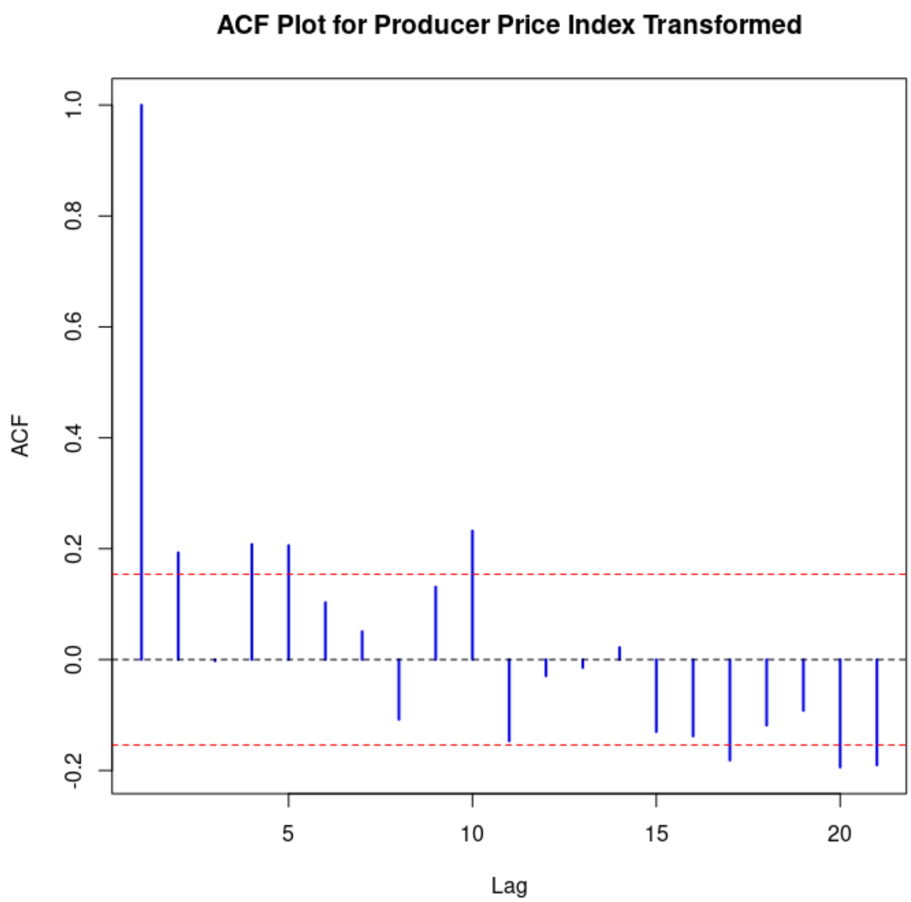


Figure 25: ACF for Transformed Producer Price Index data

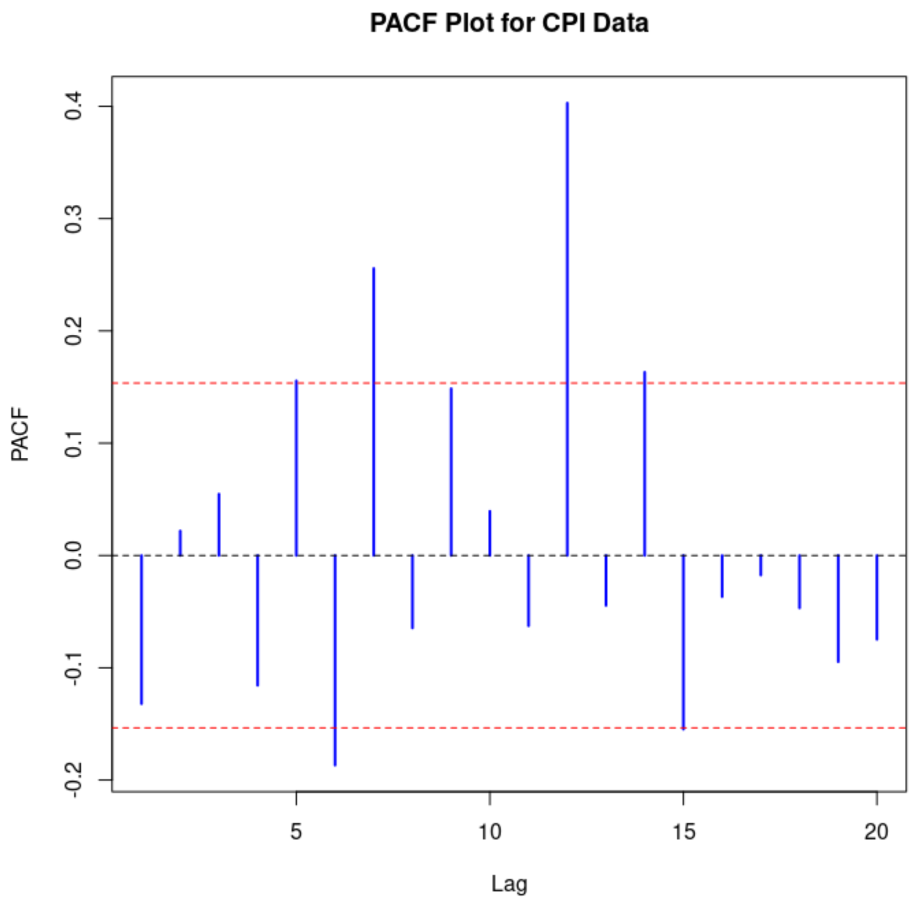


Figure 26: PACF for CPI data

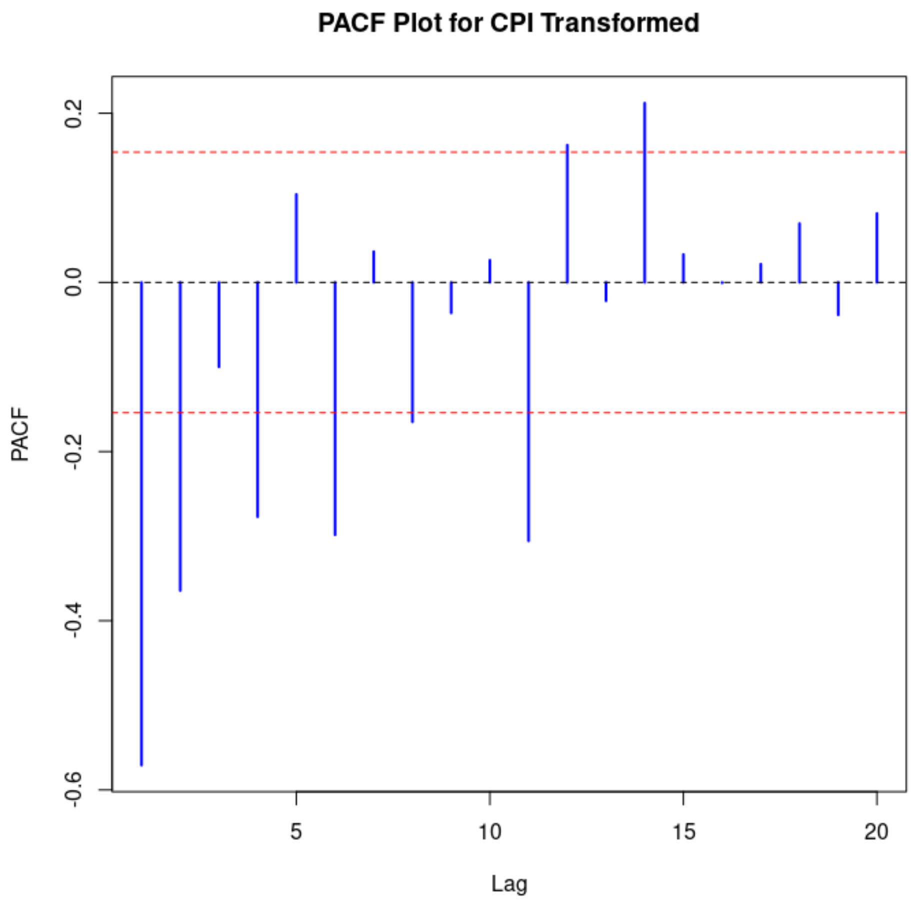


Figure 27: PACF for Transformed CPI data

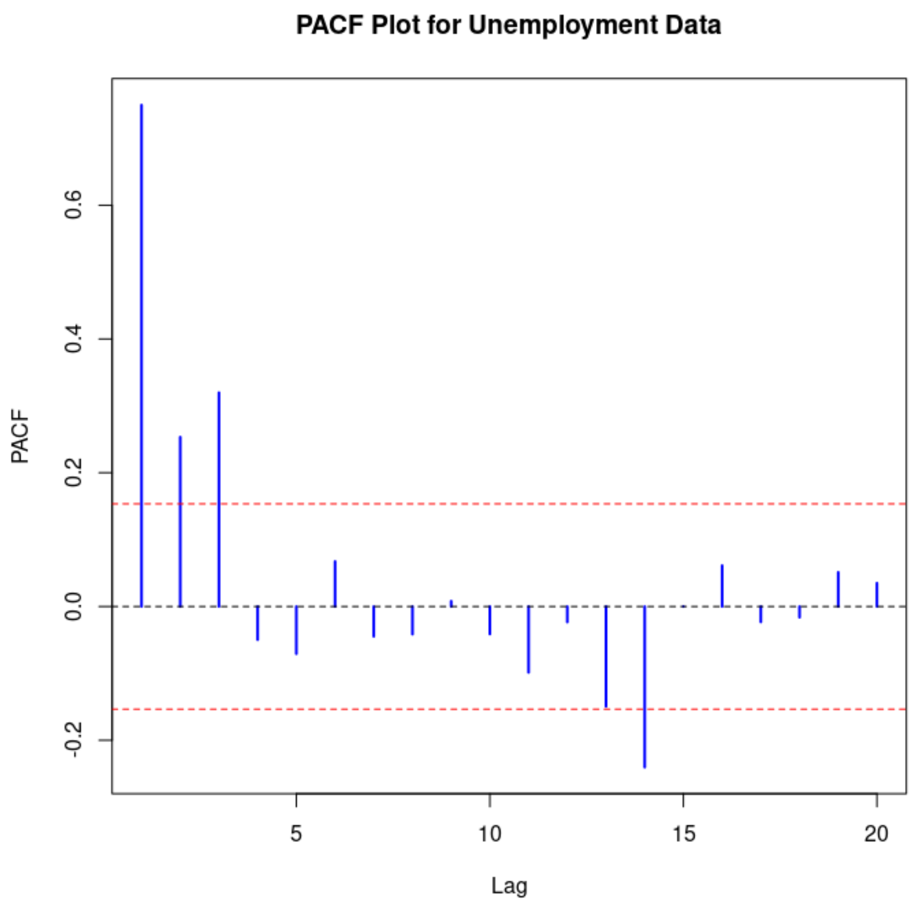


Figure 28: PACF for Unemployment data

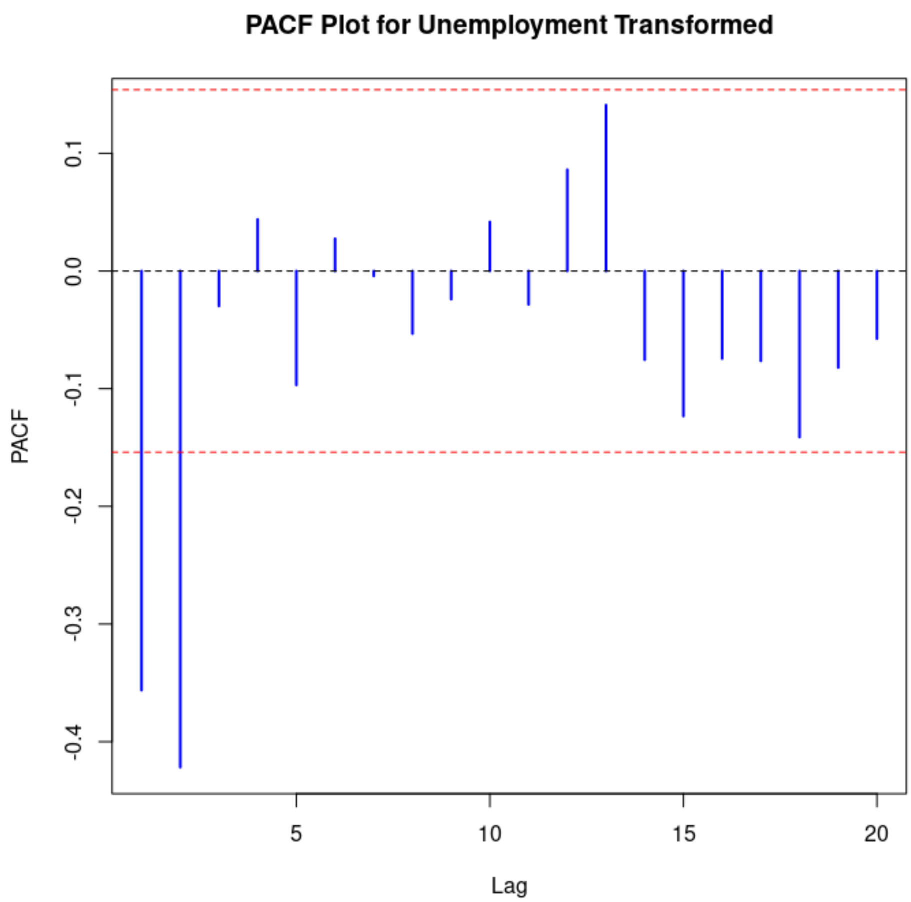


Figure 29: PACF for Transformed Unemployment data

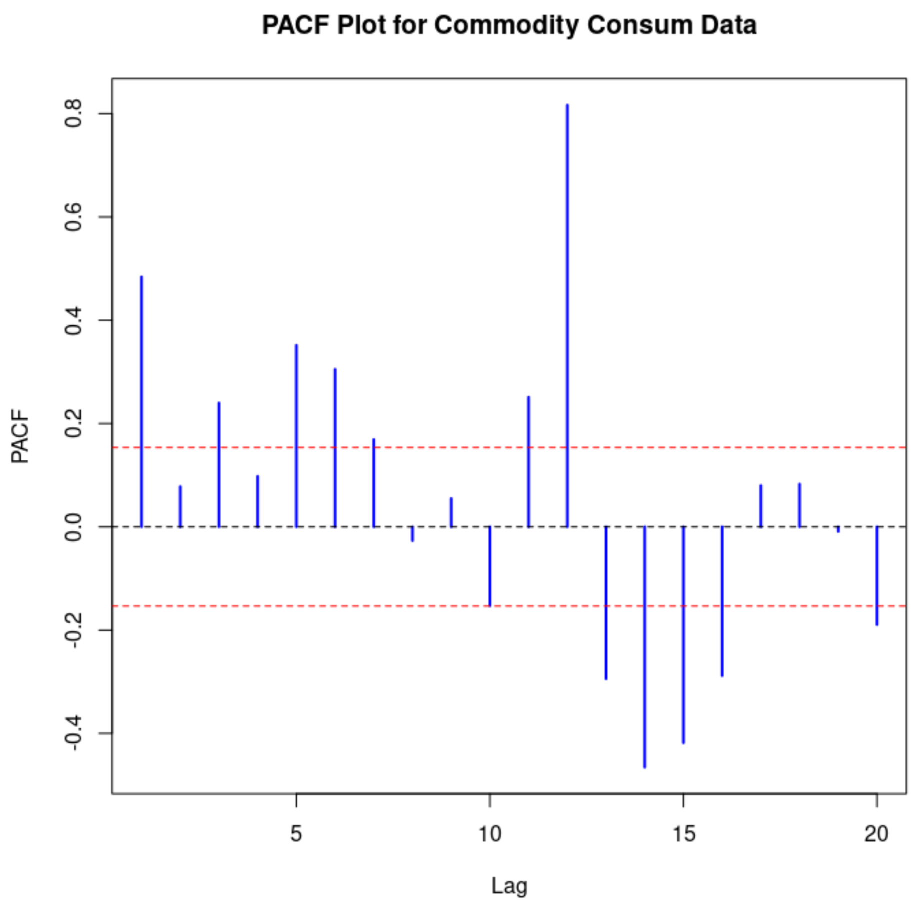


Figure 30: PACF for Product Consumption data

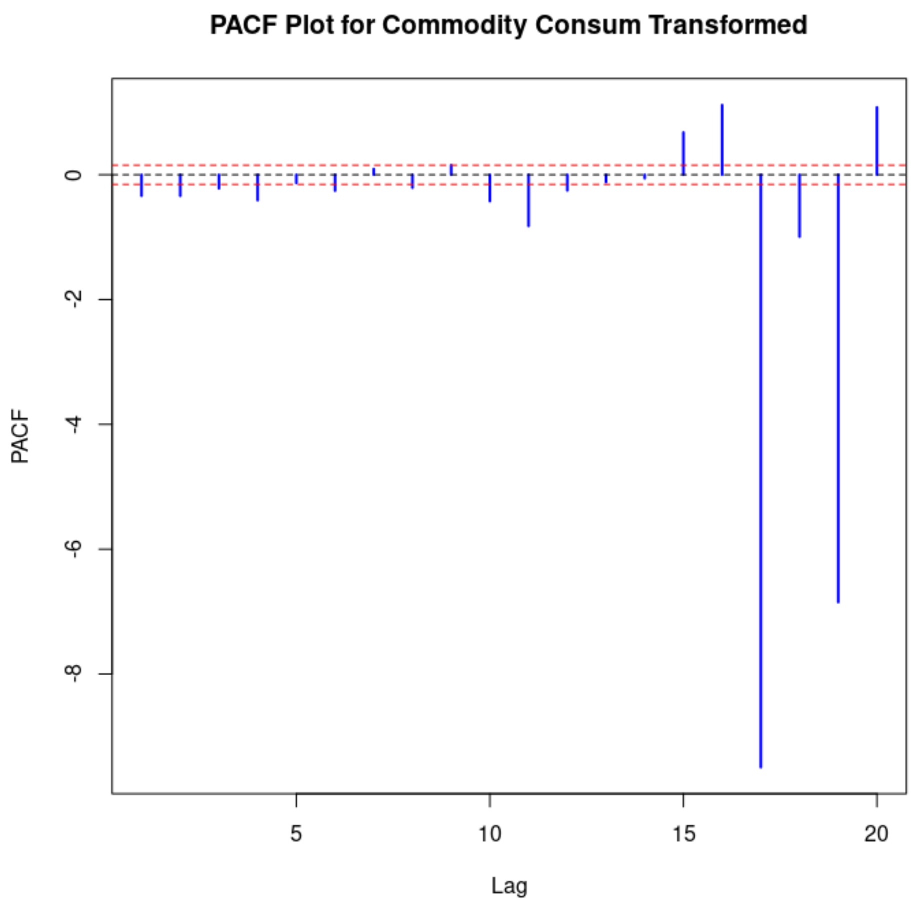


Figure 31: PACF for Transformed Product Consumption data

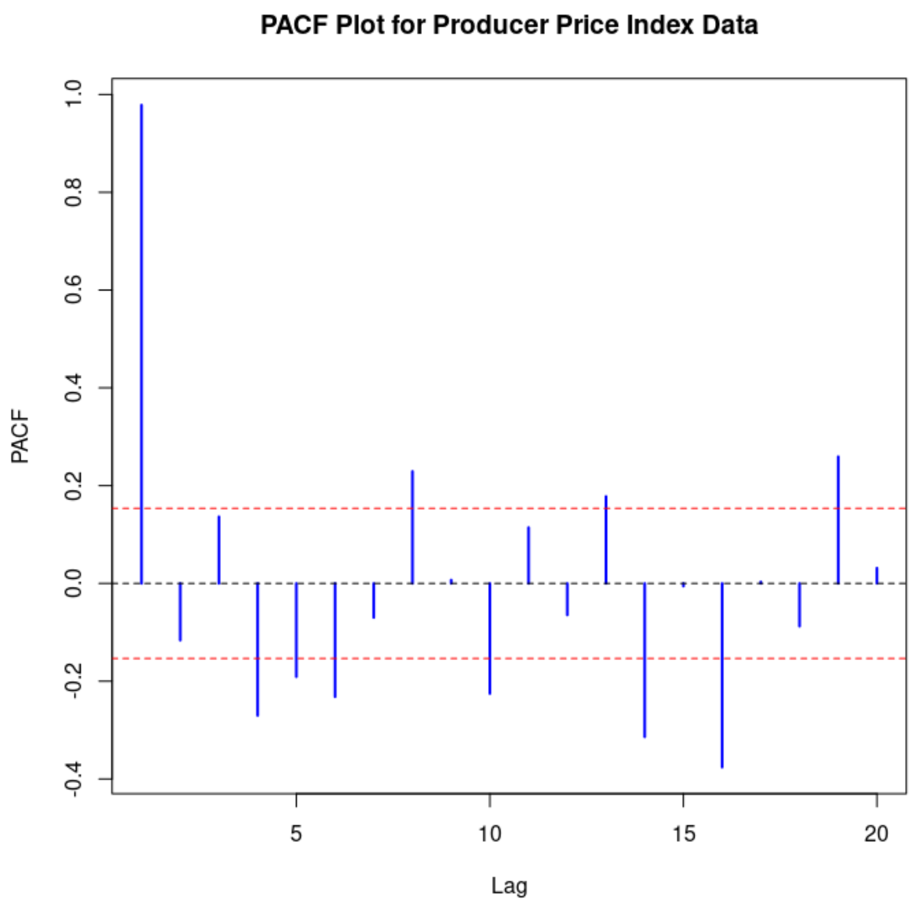


Figure 32: PACF for Producer Price Index data

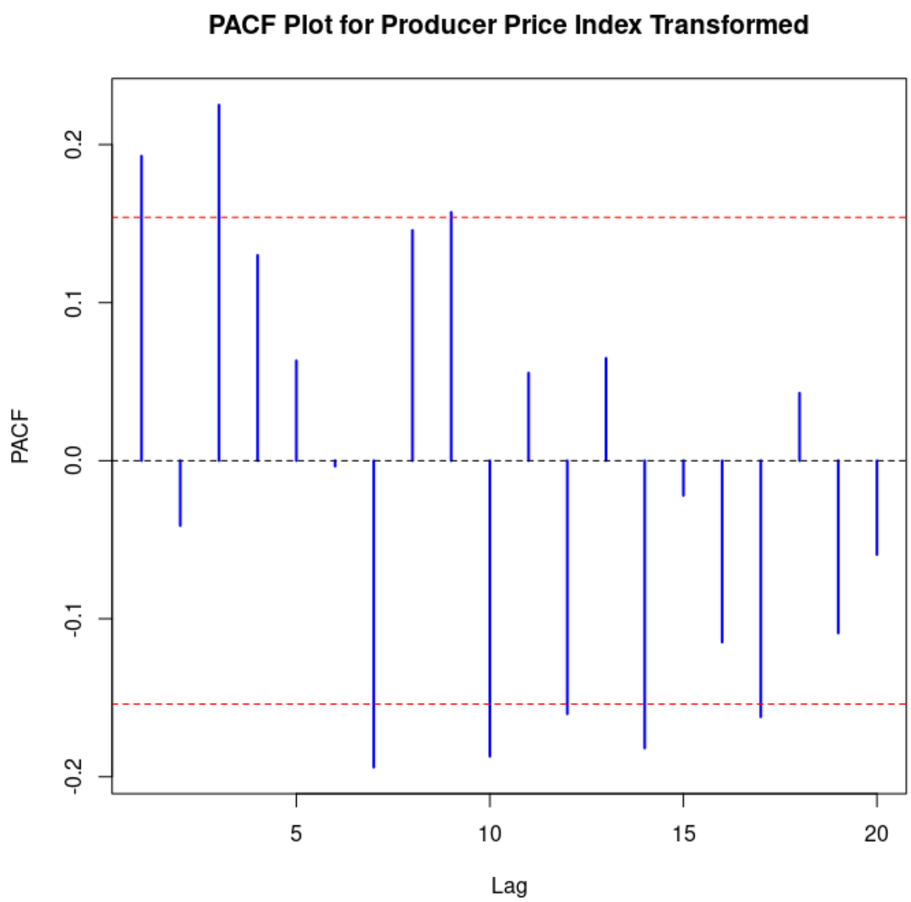


Figure 33: PACF for Transformed Producer Price Index data

Parameter	**Coefficient**	**T-Statistic**	**P-Value**	**Significant**
omega	0.52675728	4.8059271	3.592250e-06	TRUE
alpha	0.45108092	-0.4971228	6.198023e-01	FALSE
beta	-0.57416461	-5.9279190	1.902203e-08	TRUE
beta1	-0.16064596	-1.3963991	1.645781e-01	FALSE
beta2	-0.06459942	-4.2870057	3.160258e-05	TRUE
beta3	-2.81344979	-2.5596788	1.142726e-02	TRUE

Table 3: GARCH(1,1) Model Parameters with uncertainty

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
`Cpi values`	-0.9000	-0.1000	0.2000	0.2215	0.5000	1.4000
`Unemployment`	2.600	3.600	4.000	4.053	4.500	5.700
`Consumption`	1.00	33.50	67.00	69.13	103.50	142.00
`Price Index`	86.1	100.3	108.4	117.4	114.8	250.1

Table 4: Summary statistics Data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
`Cpi values`	-1.70000	-0.40000	0.00000	-0.01235	0.40000	2.00000
`Unemployment`	-0.2947995	-0.0845574	0.0000000	-0.0004941	0.0811690	0.4054651
`Consumption`	-3.37588	-0.45903	0.08010	0.02867	0.49134	3.46574
`Price Index`	-0.189359	-0.010231	0.004135	0.003693	0.023047	0.156114

Table 5: Summary statistics Transformed Data