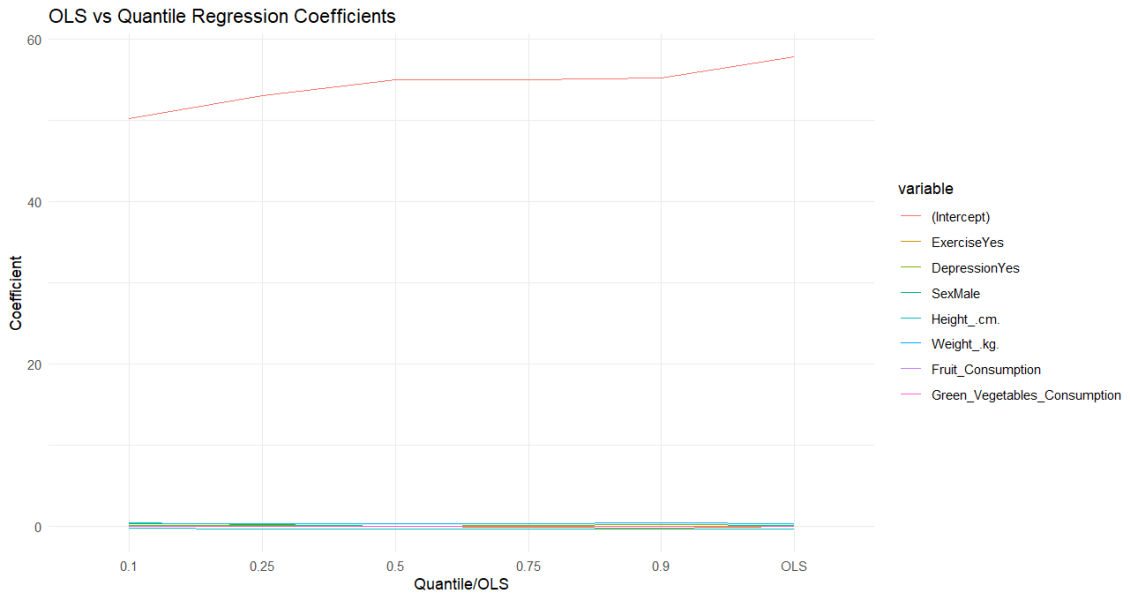
Quantile Regression

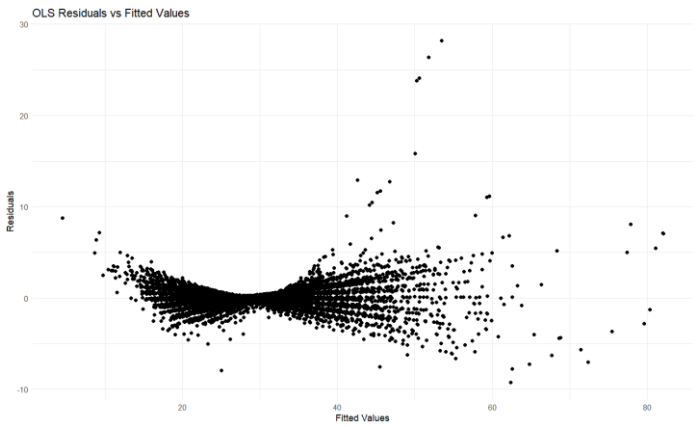
We chose to deal with Cardiovascular Diseases Risk Prediction Dataset (which we also used in advanced statistics course). This data contains information on various health and demographic factors such as age, gender, cholesterol levels, blood pressure, smoking habits, and physical activity, which are used to predict the risk of developing cardiovascular diseases. We reduced the number of items in the dataset to 30,000 to get results in a reasonable run-time.

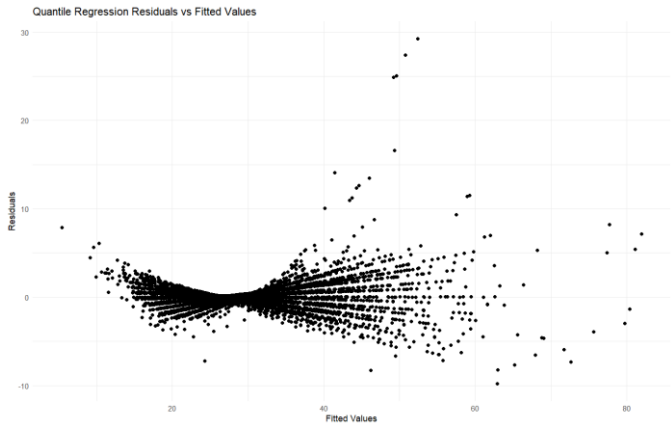
We first applied a linear regression model to identify which features are statistically significant in relation to BMI and created a model using only these features. For example, we removed ArthritisYes, which indicates the presence of arthritis, because it had a P-value of 0.91, suggesting it was not statistically significant.
Afterwards, we applied quantile regression on the selected set of features to examine the relationships between BMI and these variables across different quantiles of the BMI distribution. We saw various interesting finding:

- **Height and Weight are the strongest predictors of BMI:**
  Across all models (OLS and quantile regressions), Height_.cm. and Weight_.kg. have highly significant p-values close to zero. This is expected as BMI is directly calculated from height and weight - $BMI = \frac{\text{weight [kg]}}{\text{height}[m]^2}$.
- **Exercise is associated with lower BMI, especially in higher BMI quantiles:**
  In the OLS model, ExerciseYes has a significant negative association with BMI (p-value = 5.80e-06), indicating that those who exercise tend to have lower BMI on average. However, in the quantile regression, the negative effect of exercise on BMI becomes more pronounced at higher quantiles - in the 0.9 quantile, the coefficient is -0.194 with a very significant p-value, suggesting exercise is particularly important for those with higher BMIs.
- **Depression shows a positive association with BMI:**
  DepressionYes consistently shows a positive relationship with BMI across all models, with significant p-values. In quantile regression, this effect is stronger in the upper quantiles, indicating that depression might be more strongly linked to BMI for individuals with higher BMIs.
- **Sex has varying significance across BMI levels:**
  In the OLS model, SexMale is highly significant with BMI (p-value = 2.52e-08). However, in the quantile regression, the effect of sex diminishes at higher BMI quantiles. In the 0.9 quantile, the p-value is 0.52, indicating that sex may not be as influential in higher BMI quantiles. We didn't see it in OLS model!
- **Fruit and green vegetable consumption impact BMI differently across quantiles:**
  In the OLS model and lower quantiles (0.1, 0.25), both Fruit_Consumption and Green_Vegetables_Consumption shows negligible effects on BMI (both has 1 p-value in 0.1 quantile). However, at the 0.9 quantile, Green_Vegetables_Consumption has a significant p-value of 0.01056, indicating that green vegetable consumption does have a significant impact on BMI among individuals at the higher end of the BMI distribution.

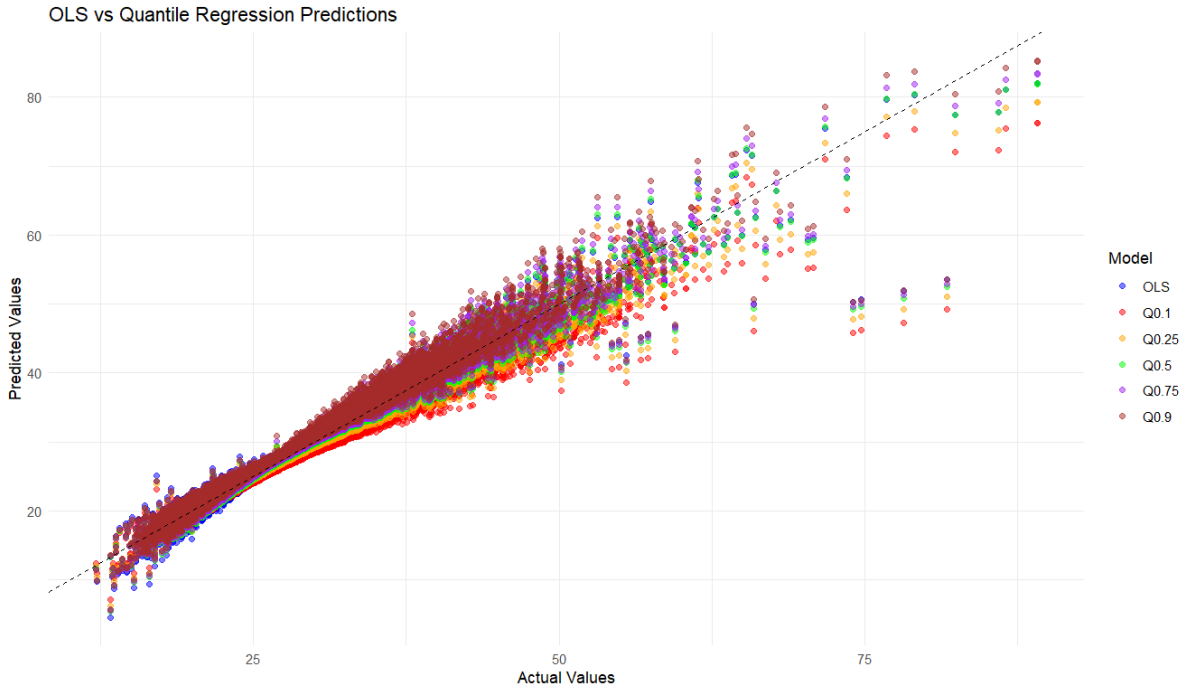


We can see that the intercept line shows a notable rise to the 0.5 quantile, indicating that a significant portion of the variation in BMI is captured by the intercept up to the median. After the 0.5 quantile, the intercept stabilizes, suggesting minimal additional variation explained by it in the higher quantiles. For the other variables, the coefficients remain relatively stable across all quantiles. This implies that their impact on BMI is consistent across different points in the BMI distribution and moreover, the OLS model seems to capture their overall effects correctly.

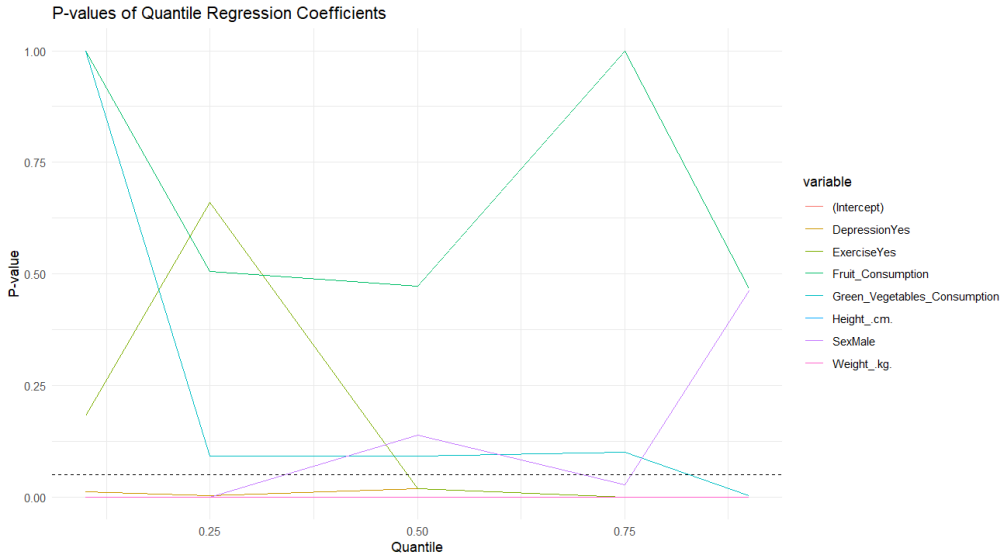




In both the OLS and Quantile Regression residual plots, there seems to be a funnel-shaped pattern with increasing fitted values, indicating heteroscedasticity - non-constant variance in the residuals. This pattern suggests that for larger fitted values, the residuals tend to be more spread out. This behavior indicates that the model might be less accurate for higher fitted values. In addition, both models show outliers, especially for higher fitted values (around 60 and above).



From this plot we can see the strengths of quantile regression in handling extreme values by adjusting predictions according to the distribution of BMI values. OLS appears to be slightly better at middle ranges but doesn't adapt well to the tails, particularly in overestimating low BMI values and underestimating high BMI values. Quantile regression provides a more nuanced understanding of the data at different levels, allowing for better handling of extremes and variability in BMI across different segments of the dataset. That's not a surprise - from the residuals plot comparison between OLS and quantile regression, we noticed a clear widening in the residuals at higher fitted values, which associated with higher BMI values. Thus, both models struggle to accurately predict high BMI values, leading to larger errors.



The p-values from Wald test indicate that: DepressionYes, ExerciseYes, Weight_kg, and SexMale are consistently significant predictors across quantiles, therefore important for predicting BMI. Height_cm shows varying significance, being more impactful in higher quantiles. Green_Vegetables_Consumption and Fruit_Consumption generally do not show significant effects, suggesting they may not be strong predictors in this model.

Panel Data

We analyzed the Cigarette Consumption dataset from the Ecdat package, comprising a panel of 46 observations spanning the years 1963 to 1992. The dataset contains 1380 observations and pertains to the United States.

The dataset includes the following variables (at least 5 independent variables):
state: State abbreviation
year: Year of observation
price: Price per pack of cigarettes
pop: Total population
pop16: Population above the age of 16
cpi: Consumer Price Index(100=1983)
ndi: Per capita disposable income
sales: Cigarette sales in packs per capita
pimin: Minimum price in adjoining states per pack of cigarettes

Models

Our fixed effects model is:

$$log(sales) = \beta_0 + \beta_1 log(price) + \beta_2 \log(\text{pimin}) + \beta_3 \log(\text{ndi}) + \beta_4 \log(\text{pop}) + \sum_{i=1}^{n} \alpha_i year_i + \epsilon$$

We converted the year variable into a factor to create time dummies, which helps control for time-specific effects that could influence the dependent variable (log(sales)). The formula defined here specifies the relationship to be modeled: it includes the logarithm of cigar sales per capita (log(sales)) as the dependent variable and several independent variables such as the logarithms of price per pack (log(price)), minimum price per pack (log(pimin)), disposable income per capita (log(ndi)), and population (log(pop)). ϵ is the error term which accounts for the deviations of the observed values from the predicted values.

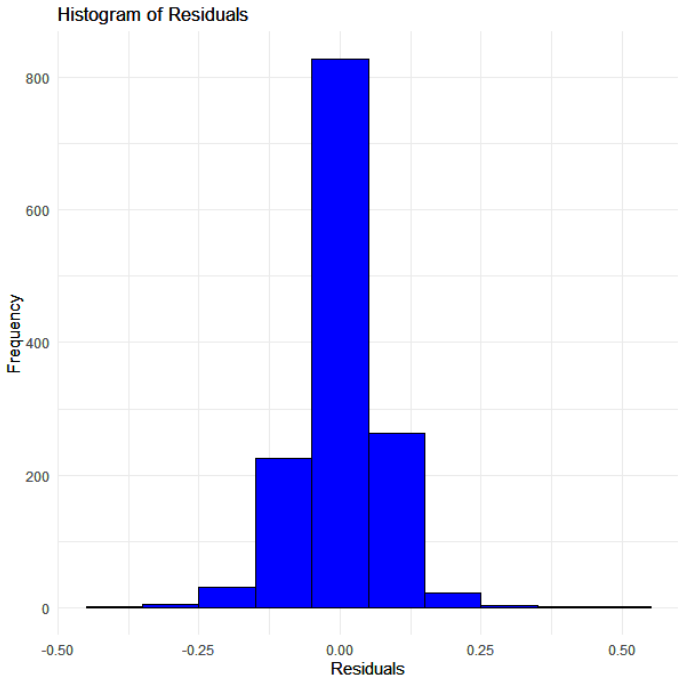
The random effects model includes the same independent variables as the fixed effects model, but it also incorporates a random effect (u_i) for each state, capturing individual-specific variations.

Hausman Test

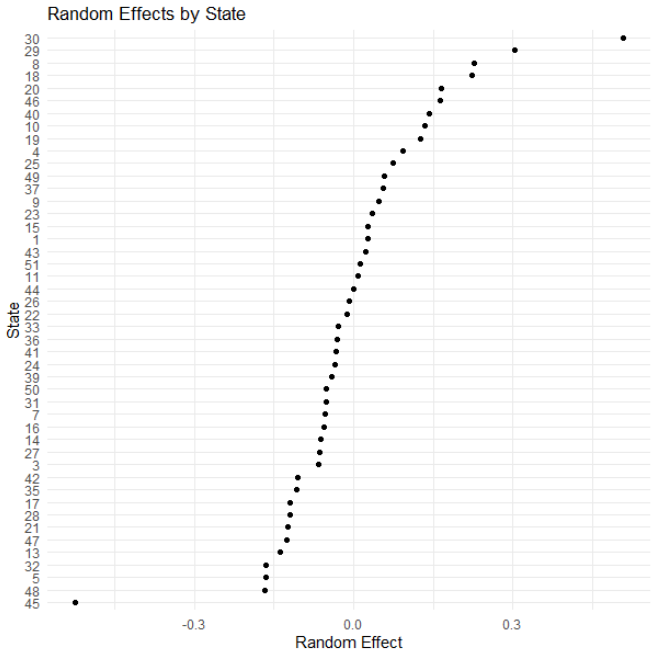
The Hausman test yielded a p-value of 1, suggesting that there is no significant difference between the random effects and fixed effects models in explaining the variation in cigarette sales per capita. This indicates that both modeling approaches provide comparable estimates for the relationship between price, minimum price, disposable income, population, and cigarette sales. Thus, the random effects model potentially offers greater efficiency and broader generalizability due to its assumption of common unobserved effects across states.

Plots:

- We were asked only to add intuitions for our work during the semester for this part.



The histogram of residuals shows that the distribution is primarily centered around zero, indicating that most residuals are small, with fewer large errors. This suggests that the model fits the data well. The residuals appear to follow a normal distribution with constant variance. As illustrated by the histogram, the distribution closely resembles a normal distribution, with most of the data points clustered around zero. This characteristic is essential for validating the residuals, as it aligns with the assumptions of normality and minimal bias, supporting the model's accuracy and reliability.



The scatter plot displays the random effects by state derived from a statistical model, with each point representing the magnitude of the random effect for a particular state. The plot illustrates the

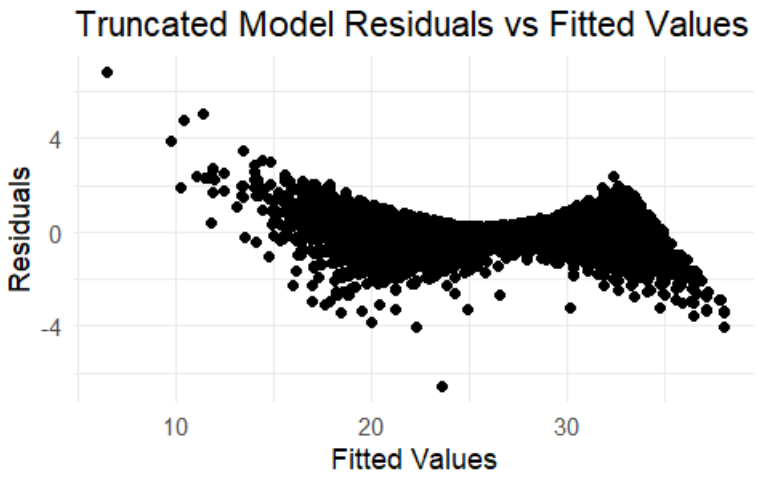
variation in random effects across different states. Similar to the histogram of residuals, most states cluster around a random effect value of zero. Notably, states 45 and 30 fall outside the range of [-0.4, 0.4], which is consistent with the overall distribution resembling a normal curve centered around zero. This pattern further supports the assumption of normality in the random effects distribution, affirming the robustness of the model.

Truncated Data

We chose to truncate the CVD dataset which we also used in the quantile regression section. We performed the truncating from above, by removing all BMI values that are considered as obese – class 2 or above. I.E we removed all values with BMI ≥ 35. We chose to do so because in OLS and quantile regression models there were a lot of errors especially in large values of BMI. Moreover, we firstly performed censored regression, but this model failed to give good results.

The OLS, quantile regression and the truncated regression are applied to the same formula. The explained variable is BMI, and the explaining variables are Exercise, Depression, Sex, Height_.cm., Weight_.kg., Fruit_Consumption and Green_Vegetables_Consumption.

Almost all variables have low p-values, but some key differences emerge in the truncated models. The first observation is that Green_Vegetables_Consumption becomes significant in the truncated model ($p = 0.00115$), even though it was borderline significant in the OLS model. This suggests that vegetable consumption may be more relevant in the context of lower BMIs. However, from the quantile regression, we know that Green_Vegetables_Consumption is significant only in high quantiles, e.g., the top decile. In fact, green vegetable consumption has a significant impact on BMI among individuals at the higher BMI values. Therefore, we decided to make another truncation of the CVD dataset, but this time we removed all values with BMI considered as overweight (i.e., BMI ≥ 25). The results were striking. For Green_Vegetables_Consumption, we obtained a p-value of 0.8295, which indicates that **green vegetable consumption has little to no significance in individuals with a BMI below 25.** In that truncation (only BMI < 25), we also saw that **exercise became less significant, while the effect of sex on BMI remained strongly significant, implying sex may be a more robust predictor of BMI in the lower range** – just as we saw in quantile regression**.**



Also in this case, there's evidence of heteroscedasticity. Where the spread of residuals changes across different ranges of fitted values. The residuals exhibit greater spread for lower fitted values (around 10–17) compared to the higher fitted values.

Compared to the OLS model, the truncated model shows a lower residual variation ($\sigma = 0.532$), as it is restricted to a smaller range of BMI values.

Time Series Models

We chose to utilize time series data from the Yahoo Finance library, specifically focusing on two major indices: the S&P 500 (represented by 'SPY') and the NASDAQ-100. Initially, we planned to apply an Autoregressive Distributed Lag (ADL) model to analyze the relationship between these indices. To determine the suitability of the ADL model, we conducted the Augmented Dickey-Fuller (ADF) test to check for stationarity in the data. If the data did not exhibit stationarity, indicating the ADL model may not be appropriate, we intended to explore alternative models, such as the ARIMA model, for better fitting and forecasting.

Explaining the data: financial dataset related to stock market indices or securities, with each row representing a daily of the stock market. an explanation of each column:

- ticker: This column contains the ticker symbol, which is a unique identifier for a financial security or index. Example value: ^FTSE refers to the FTSE 100 Index
- ref_date: reference date for the data, indicating the specific day when the data was recorded
- price_open: the opening price of the index on the specified date.
- price_high: the highest price the index reached during the trading session on that day.
- price_low: the lowest price the index reached during the trading session on that day.
- price_close: The closing price is the last price at which the index traded before the market.
- volume: representing the number of shares or contracts traded during the trading session for the index/security.
- price_adjusted: adjusted closing price accounts for any actions such as dividends, stock splits, or other corporate actions that might affect the stock's value.

We obtained stock market data spanning the last five years, specifically from September 19, 2019, to September 19, 2024. In preparation for our analysis, we performed data cleaning to address any missing values (NaN). Next, we applied a logarithmic transformation to the **price_adjusted** column to stabilize variance and interpret the data in terms of growth rates. Additionally, we created a differenced price column based on the log-transformed data to capture short-term changes and ensure stationarity. We also introduced a time trend variable (**t_trend**) as a sequential count from 1 to the total number of rows in the dataset. This time trend represents the progression of time and can be incorporated into further analysis as needed.

We conducted stationarity testing using the Augmented Dickey-Fuller (ADF) test on both indices, specifically examining whether the log-transformed time series (log_SPY and log_NASDAQ) were stationary or non-stationary. The ADF test evaluates the presence of a unit root in the series, which indicates non-stationarity. The null hypothesis of the ADF test posits that the series has a unit root (i.e., it is non-stationary). If the p-value of the test is below a chosen significance level (e.g., 0.05), we reject the null hypothesis, implying that the series is stationary.
Upon running the ADF test on both SPY and NASDAQ, we obtained p-values of 5.89e-09 for SPY and 3.635e-08 for NASDAQ. These values are significantly smaller than the 0.05 significance threshold, meaning that the ADF test fails to reject the null hypothesis for both indices. As a result, both series exhibit non-stationarity at the tested significance levels.

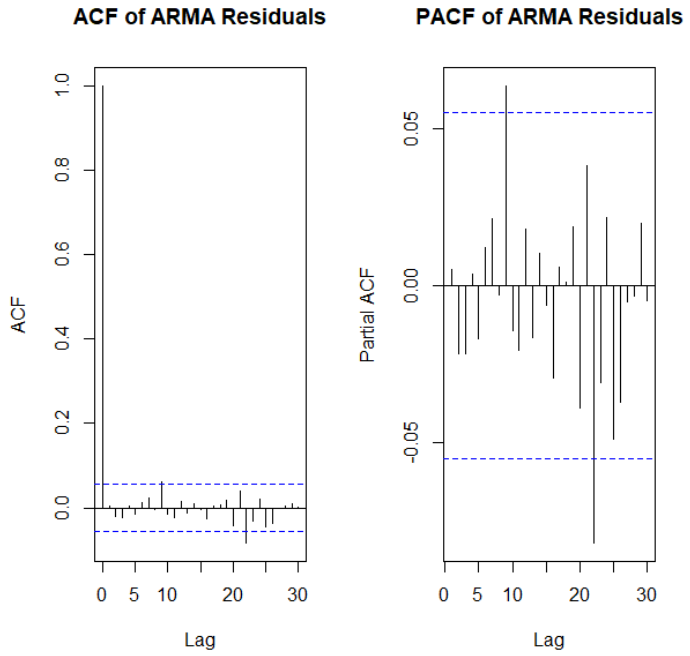
Due to this outcome, we cannot proceed with the ADL (Autoregressive Distributed Lag) model. The ADL model requires the variables to be stationary, or at least cointegrated, to capture the relationship between them effectively. Non-stationary data can lead to spurious regression results, given that both indices were found to be non-stationary, we opted to use an ARIMA model to address the issue. To determine the best-fitting ARIMA model, we implemented a function that

optimizes the parameters (p, d, q) based on the minimum Akaike Information Criterion (AIC). Additionally, the function ensures that the residuals from the selected model exhibit white noise behavior by performing a Ljung-Box test, which tests for autocorrelation in the residuals.

The optimization process identified the best ARIMA model with the following parameters: AR = 5 (autoregressive terms), I = 0 (no differencing), and MA = 4 (moving average terms). This results in an ARMA(5, 4) model, as no differencing was necessary. The model's AIC was exceptionally low at -6858.5914, indicating a strong fit to the data. Furthermore, despite some convergence warnings during the fitting process, the model showed a high log-likelihood, suggesting a good overall fit. Residual analysis confirmed the appropriateness of the model, with minimal autocorrelation in the residuals (ACF1 = 0.000348). This indicates that the model successfully captured most of the autocorrelation in the data, ensuring that the residuals approximate white noise, which is a crucial criterion for a well-fitted time series model. Overall, the ARMA(5, 4) model provided an effective solution for the non-stationary indices.

ACF and PACF of the ARMA residuals:

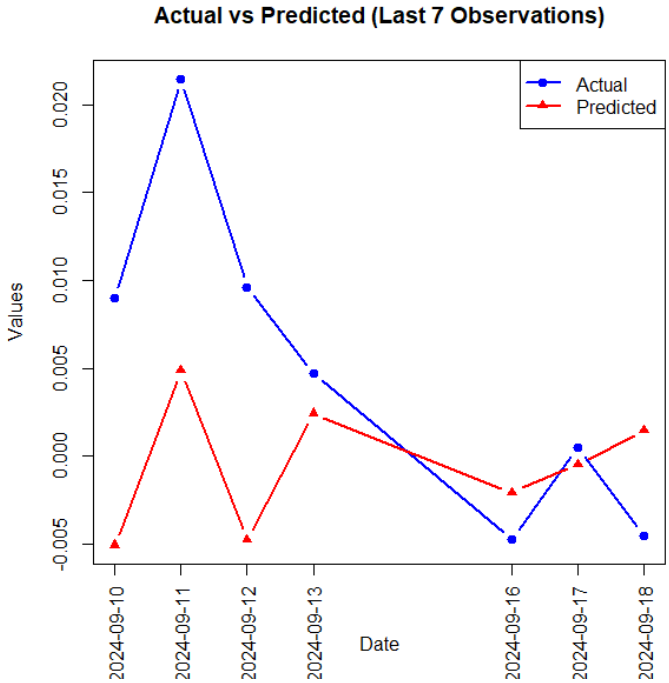
These plots are utilized to analyze the residuals of the ARMA(5, 4) model fitted to the differenced log NASDAQ data. The primary objective of these plots is to assess whether the residuals exhibit characteristics of white noise, which is a fundamental assumption for a well-specified ARMA model. Ensuring that the residuals resemble white noise confirms that the model has effectively captured the underlying autocorrelation structure of the time series, leaving no significant patterns unaccounted for.



The ACF and PACF plots reveal that the residuals of the ARMA(5, 4) model are largely uncorrelated and exhibit behavior consistent with white noise, with only minor residual autocorrelation present at a few early lags. This suggests that the model effectively captures the underlying structure of the time series data, although a small degree of unexplained autocorrelation persists. The near-white noise pattern of the residuals confirms that the ARMA(5, 4) model is well-suited for this dataset, though there remains potential for further refinement to eliminate the remaining autocorrelation. In conclusion, the residual analysis demonstrates that the ARMA(5, 4) model provides a satisfactory

fit to the data. However, slight improvements in model performance may still be possible. The results from the ACF and PACF plots also indicate that the time series is not purely autoregressive (AR) or purely moving average (MA) in nature but rather a combination of both, as appropriately captured by the ARMA model.

Forecasting ARMA:



The above plot compares the actual and predicted values for the last seven observations of the differenced log NASDAQ series, as forecasted by the ARMA(5, 4) model. While the ARMA model successfully captures the general downward trend in the data, discrepancies between the actual and predicted values are evident. These deviations are particularly pronounced around sharp fluctuations in the actual data, such as on 11-09-2024 and 13-09-2024, where the model underestimates or fails to capture the magnitude of the peaks and dips. This discrepancy is more noticeable during periods of volatility, highlighting the challenge of forecasting short-term market movements with precision. The model's smoother predictions, especially compared to the actual values, indicate that while it performs adequately in identifying general trends, it struggles with accurately predicting the magnitude of rapid changes. This is a common issue in time series forecasting, particularly for financial data, where short-term volatility is often difficult to model and predict due to the inherent noise and uncertainty in market behavior.

In conclusion, although the ARMA(5, 4) model provides reasonable predictions of the general trend, the mismatch between the actual and predicted values suggests limitations in its ability to fully capture the nuances of short-term market fluctuations. This could indicate that further model refinement or the inclusion of additional explanatory variables may be necessary to improve predictive accuracy in such volatile periods.