**Avery Girsky, James Peavy & Moe Kyaw Thu (Group 7)  
December 8th 2023  
ISyE 6402: Time Series Analysis  
Project Report[[1]](#footnote-0)**

**A Comprehensive Analysis of the Dynamics: Student Loans, College Costs, and Economic Factors in the US (1960s-2010s)**

***I. Introduction***

In the landscape of higher education in the United States, existing dynamics between student loans, starting salaries, college costs, and other various factors have undergone significant changes between the 1960s and the 2010s. Against this backdrop, the purpose of this study is to unravel the trends and correlations, answering critical questions about the evolution of student loan debt, changes in college costs, shifts in the student debt-to-income ratio, and the overarching influence of inflation. To undertake such tasks, this study asks the following questions:

1. How has student loan debt changed from the period between the 1960s and the 2010s?
2. In what ways college costs have evolved over the decades from the 1960s to the 2010s?
3. In what ways have college degree starting salaries evolved over the decades from the 1960s to the 2010s?
4. How do the outcomes identified in questions II and III relate to the student loan debt’s overall changes?
5. How does the collective information from the above questions connect with the impact of changes in inflation, GDP, and population?
6. Have there been any significant changes in economic/social policies that could affect both factors?

Before diving into our data, we expect for there to only be basic trends before revealing stationarity. Additionally, we expect a VARX model with Inflation, GDP, and Population data as exogenous variables to perform the best of any VAR modeling. We would also expect there to be statistically significant relationships between at least some lags of all of our data for each time series, with an expectation that loan data and tuition data is heavily related.

***II. Dataset***

1. **Student Loan Dataset**

The dataset for student loans in the US, sourced from Kaggle - a data science competition platform and online community - provides a comprehensive repository of structured data, capturing the intricacies of US student loans across the period from 1958 to 2014. This dataset encompasses loan amounts in billions, borrower demographics, and period, providing a holistic view of the student loan landscape.[[2]](#footnote-1) To complement this, a second dataset acquired from the Federal Reserve Bank of New York extends the temporal scope, offering information specifically on student debt from 2004 to 2022.[[3]](#footnote-2) These datasets collectively serve as the bedrock for a nuanced analysis of the evolving dynamics of student loans.

1. **College Costs**

Moving onto the data on college costs, the National Association of Colleges and Employers (NACE) dataset presents information on the average tuition, fees, room, and board rates assessed for full-time students in degree-granting postsecondary institutions. The data is organized by the level of education (e.g., undergraduate) and the control of the institution (e.g., public or private) with the period, stretching from 1963 to 2022. Overall, it provides a comprehensive understanding of the financial aspects associated with pursuing higher education.[[4]](#footnote-3)

1. **Average Starting Salaries**

Similarly, this study obtains the average starting salaries from NACE’s repository. Overall, it provides a historical dataset encompassing majors such as business, engineering, humanities, and mathematics. Covering data since 1960, this dataset serves as a valuable tool for tracking salary trends over time, shedding light on the economic outcomes of college education.[[5]](#footnote-4)

1. **Inflation Data**

The inflation dataset, provided by the World Bank, acts as a crucial contextual backdrop. Providing an indicator for the inflation rate in the US since 1960, it becomes instrumental in understanding the broader economic conditions that may influence both student loan dynamics and college costs.[[6]](#footnote-5)

1. **GDP Data**

As the study also examines the economy of the US and its relations with other factors, the dataset on GDP provided by the World Bank, is also crucial in understanding such dynamics.[[7]](#footnote-6)

1. **Population Data**

As the study also examines the relationship between factors that are derived from the general population of the US, it is also important to include this as potential exogenous factors in our analysis.[[8]](#footnote-7)

***III. Methodology***

1. **Data Cleaning**

**i. Student Loan Dataset**

For the student loan dataset, this study combines the data provided by Kaggle and the Federal Reserve Bank of New York by using Excel. As stated, while the former covers the period from 1958 to 2014, the latter covers up to 2020. Additionally, the combined dataset is further rounded up to 2 decimal places in order to account for precision and facilitate a more streamlined analysis. This process of merging these datasets involves aligning the corresponding time periods and reconciling any discrepancies in data formats.

**ii. College Costs**

The college tuition data was organized into 120 different time series combinations. These combinations included every combination of the following four categories:

1. Institution type: **all**, public, private, private/non-profit, private/for-profit
2. Inflation adjustment: constant (adjusted to 2021-2022 prices), **current** (at-time cost)
3. Type of cost: **all**, tuition and fees, room, board
4. Program length: **all**, four year degree, two year degree.

To proceed with analysis of tuition costs, we decided to hone our focus onto the total (including tuition and fees, room, and board) cost unadjusted for inflation of all institutions of all lengths. This decision was made to understand the picture for an average college student throughout the last 50 years regardless of any specific categories. Further, the inclusion of inflation adjusted data into multivariate models influenced the decision to avoid inflation adjusted data in this individual time series. Lastly, in order to facilitate more readable code and visualizations, data from the school year y1-y2 would be truncated to read only y1 (i.e. 2021-22 reads as 2021).

**iii. Average Starting Salaries**

The data on average starting salaries was organized into reported averages (not accounting for inflation), percent change (reported average), inflation-adjusted averages, and percent change (inflation-adjusted averages). We took just the data on the reported average and formatted it for easy manipulation. We were unable to find any solid numbers beyond the year 2015 that linked with the data we were using, so any analysis using the starting salaries beyond the year 2015 is not included.

**iv. Inflation Data**

The data on Inflation was given initially as a decimal representation of the percentage (1.01 equates to a 1% increase). We were able to find good data from the 1960s onward so all analysis using inflation was unbounded by our access to inflation information.

**v. GDP Growth Data**

The data on Inflation was given initially as a percentage increase/decrease from the previous year. We were able to find good data from 1961 onward so all analysis using GDP Growth.

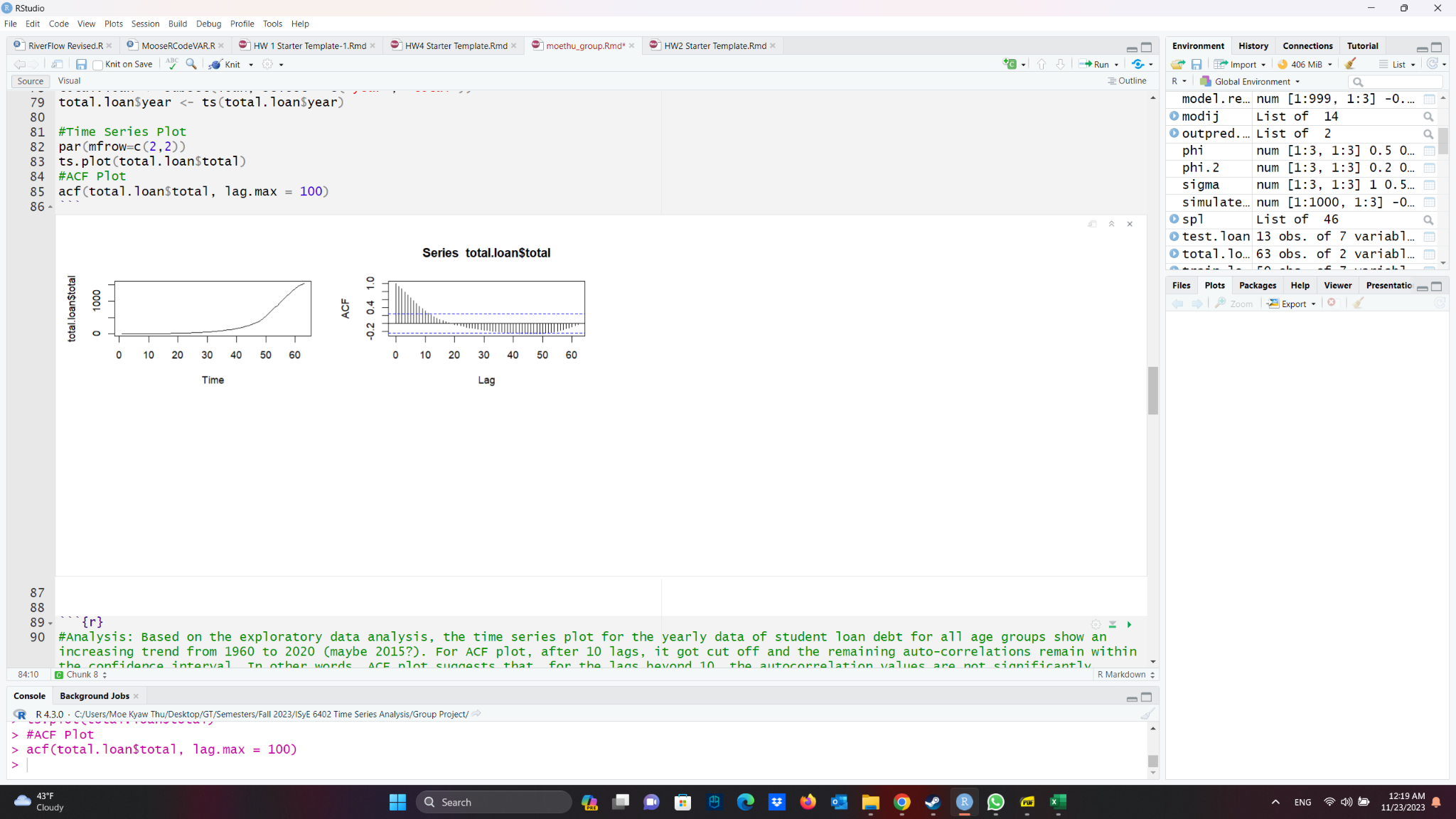
**vi. Population Data**

The data on Population was given as total numbers in 1000s of individuals. For our purposes, we corrected this to be raw numbers by multiplying but 1000. Data was broken up into present and past borders, we used present borders, and data was given continually from 1800 onwards.

***IV. Results***

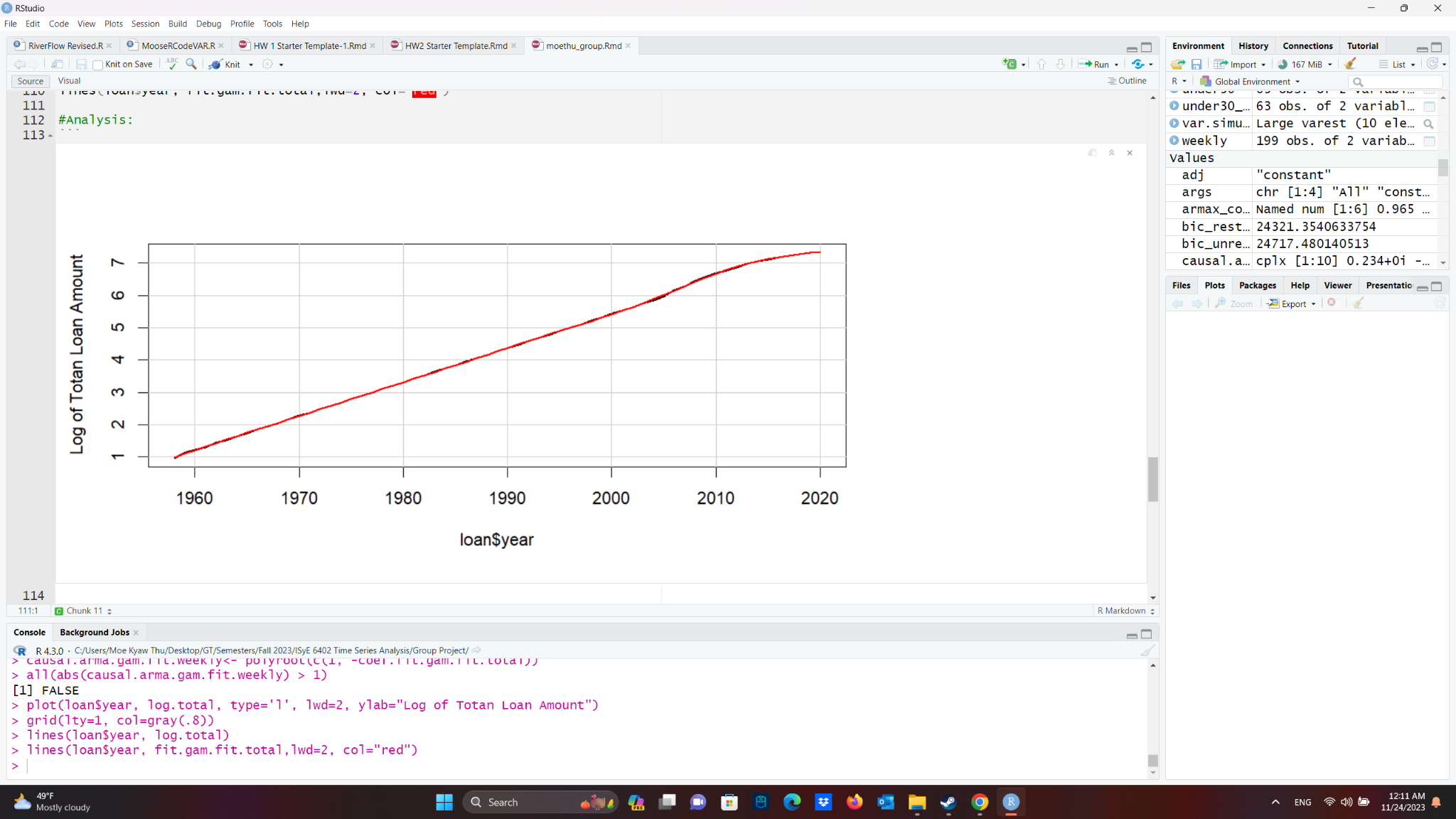
1. **Tracing the Student Loans from the 1960s to the 2020s**

Overall, the exploratory data analysis reveals a discernible upward trend in the time series plot depicting yearly data on student loans across all age groups from 1958 to 2020. Turning attention to the autocorrelation function (ACF) plot, a notable observation; beyond 20 lags, the ACF plot exhibits a truncation, with subsequent autocorrelation values consistently falling within the confidence interval. This demonstrates that the autocorrelation values lack statistical significance and are essentially interpreted as random fluctuations, In essence, the exploratory data analysis reveals a clear upward trend, and beyond 20 lags, the autocorrelations do not significantly deviate from zero, reinforcing the notion that these correlations being statistically indistinguishable from random noise.



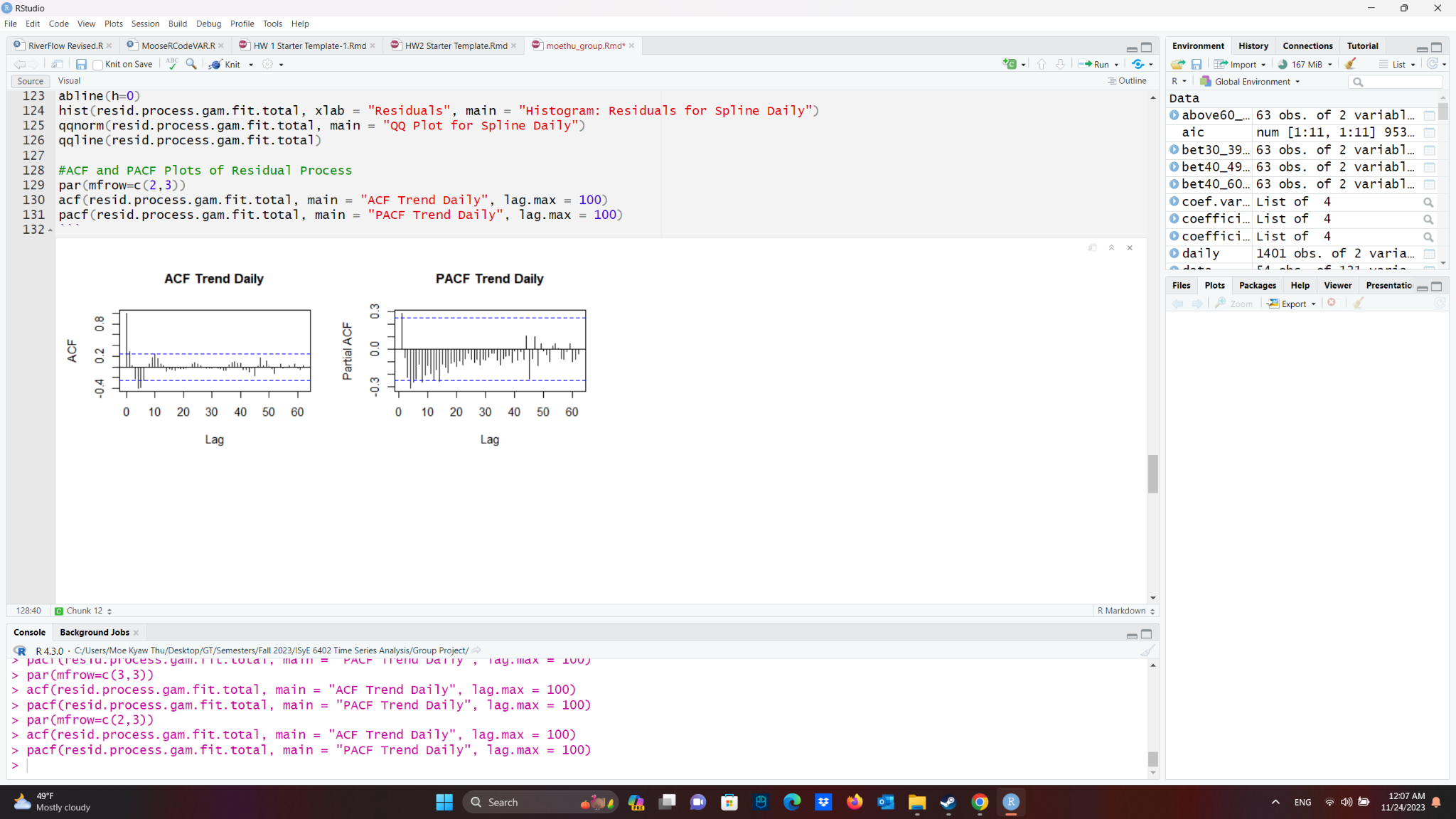
**Fig.1.** Trend and Seasonality Analysis of Total Student Loans

In order to further deepen the analysis of the student loan trend data, this study also applies spline regression analysis. The spline regression model is paired with a Gaussian family and an identity link function that effectively captures the relationship between the logarithm of the total student loan and the smooth term, the index of the total student loan amount. Additionally, both the intercept and the smooth term are highly significant, and the model fits the data extremely well, explaining all the variability in the response variable.[[9]](#footnote-8) The output is further demonstrated by fitting in the spline regression output with the original time series and the fit shows an almost perfect fit (see Figure 2).



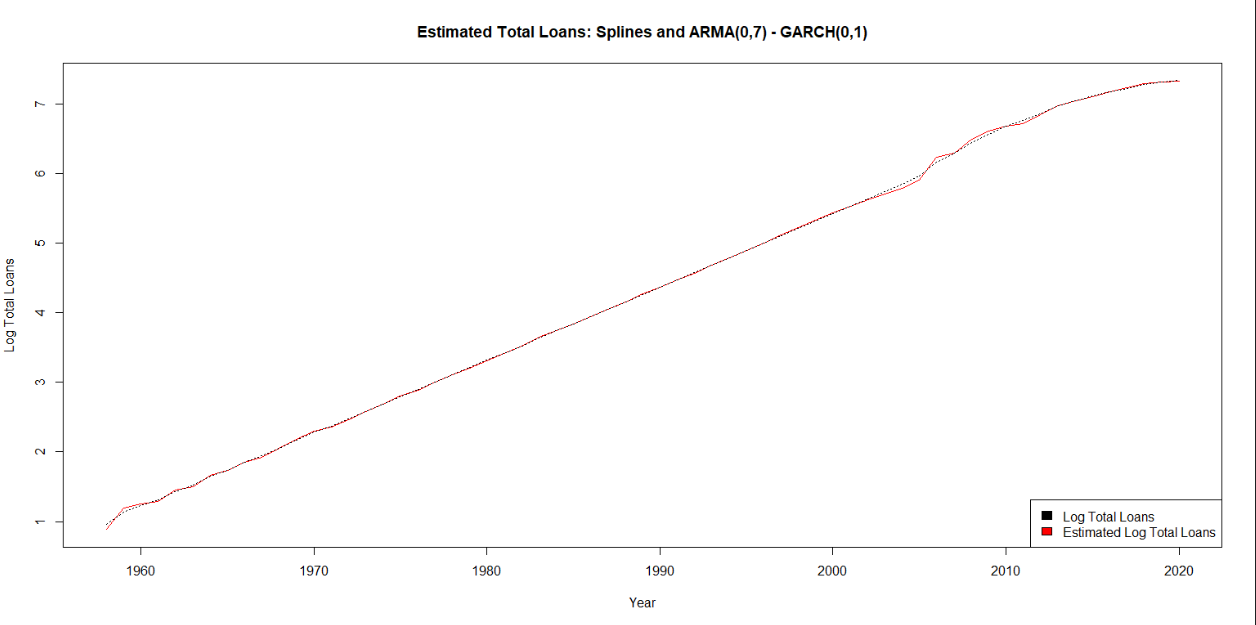
**Fig.2.** Spline Regression Analysis

Lastly, this study also analyzes the residual process after the trend removal. We fit an ARMA model to the post-trend-removal residuals (Fig. 3.). The resulting process was one of order (3,0,3). Analysis of this process’ residuals yielded signs of stationarity via an ADF test, signs of a lack of serial correlation via both the Box-Ljung and Box-Pierce tests, a lack of normality in the residuals, and signs of heteroskedasticity (Appendix: Total Loans ARMA Hypothesis Testing)



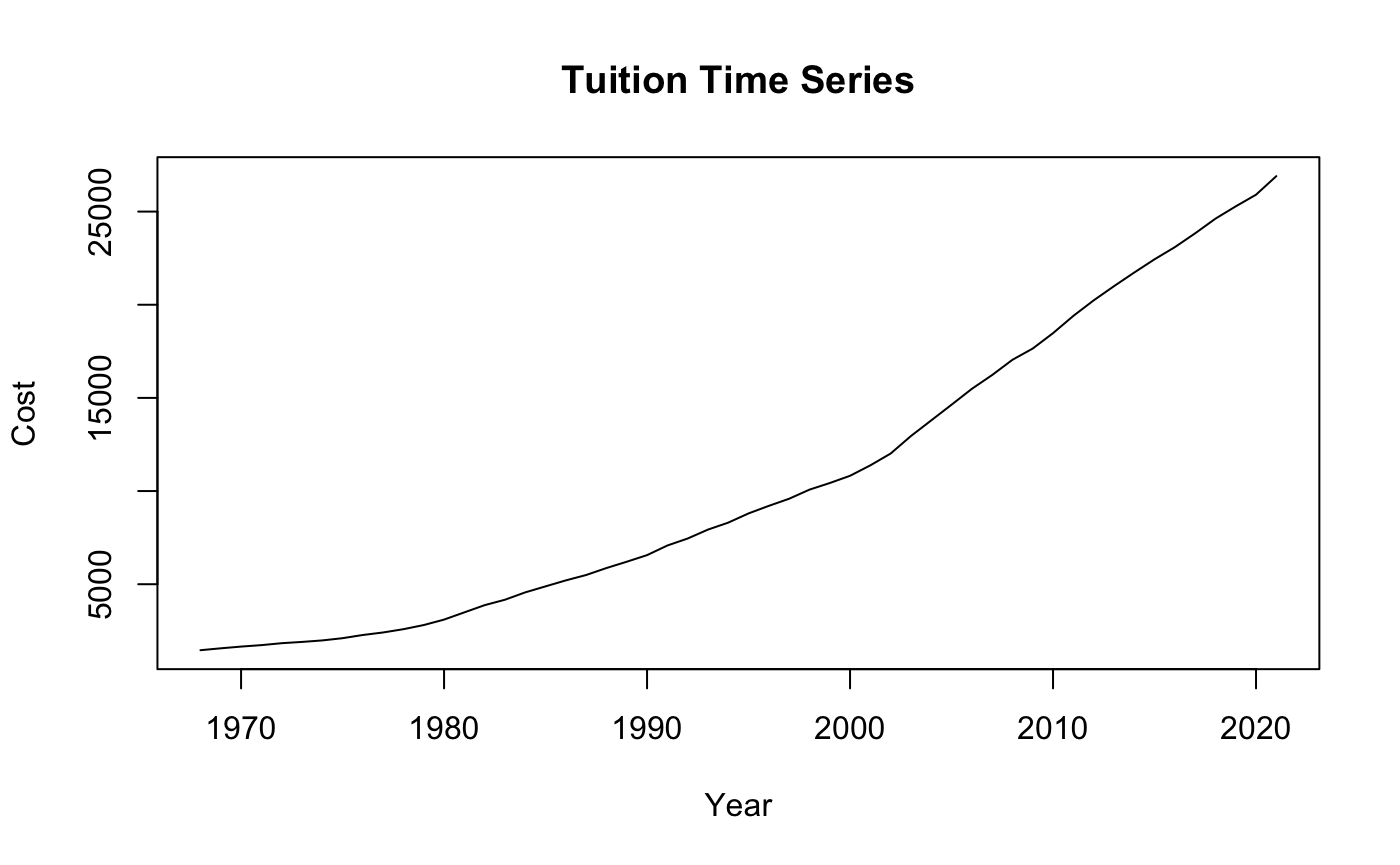
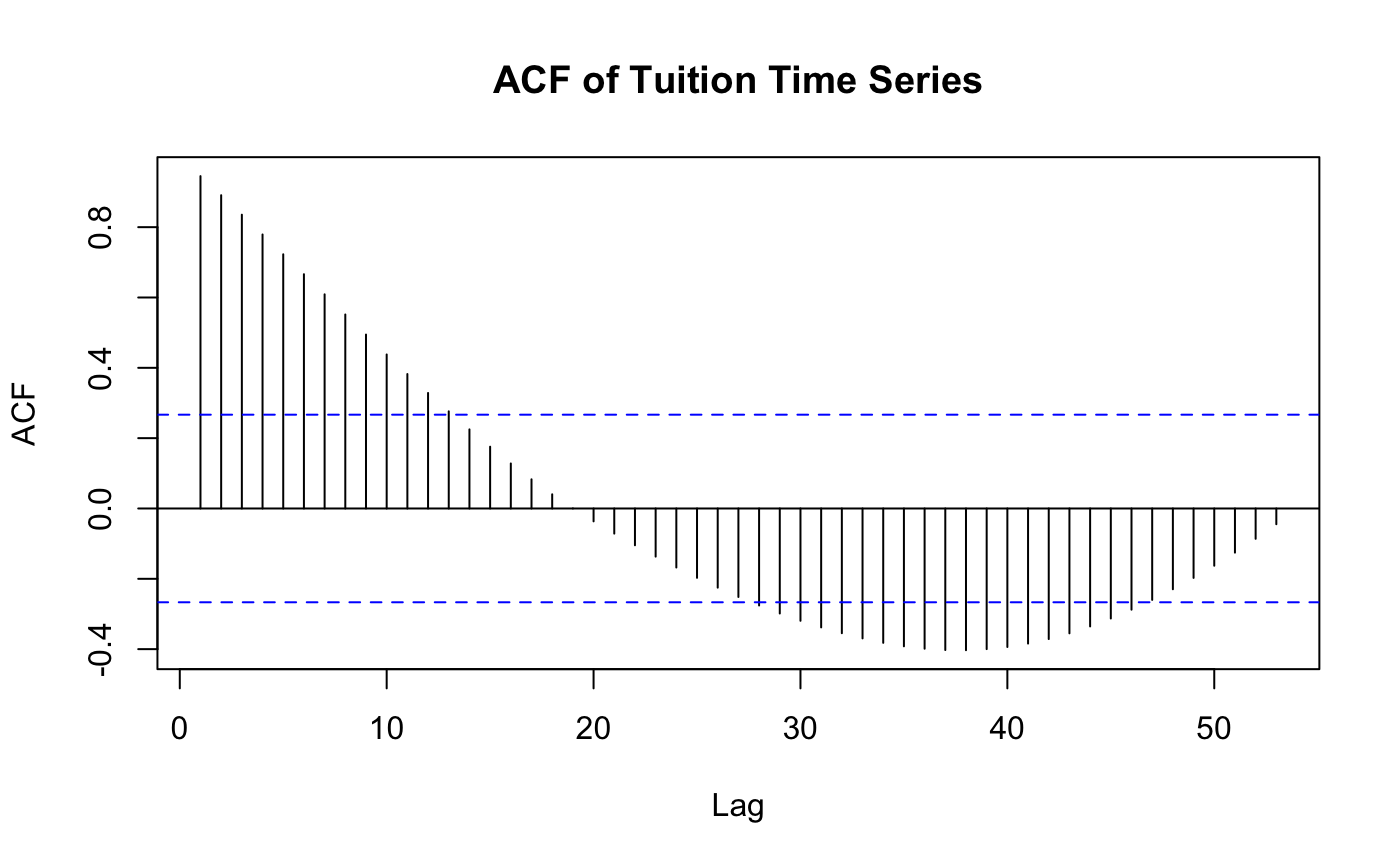
**Fig.3.** ACF and PACF Trend of Residual Process

Given these findings, we looked at an ARMA-GARCH model of order (7,0,7)x(1,0,1). This produced similar results to the above ARMA model (Fig. 4.), but with the added benefit of showing signs of normality in the residuals. Given these results, we believe that the ARMA-GARCH model is slightly better than the ARMA model if one wants to perform predictions.

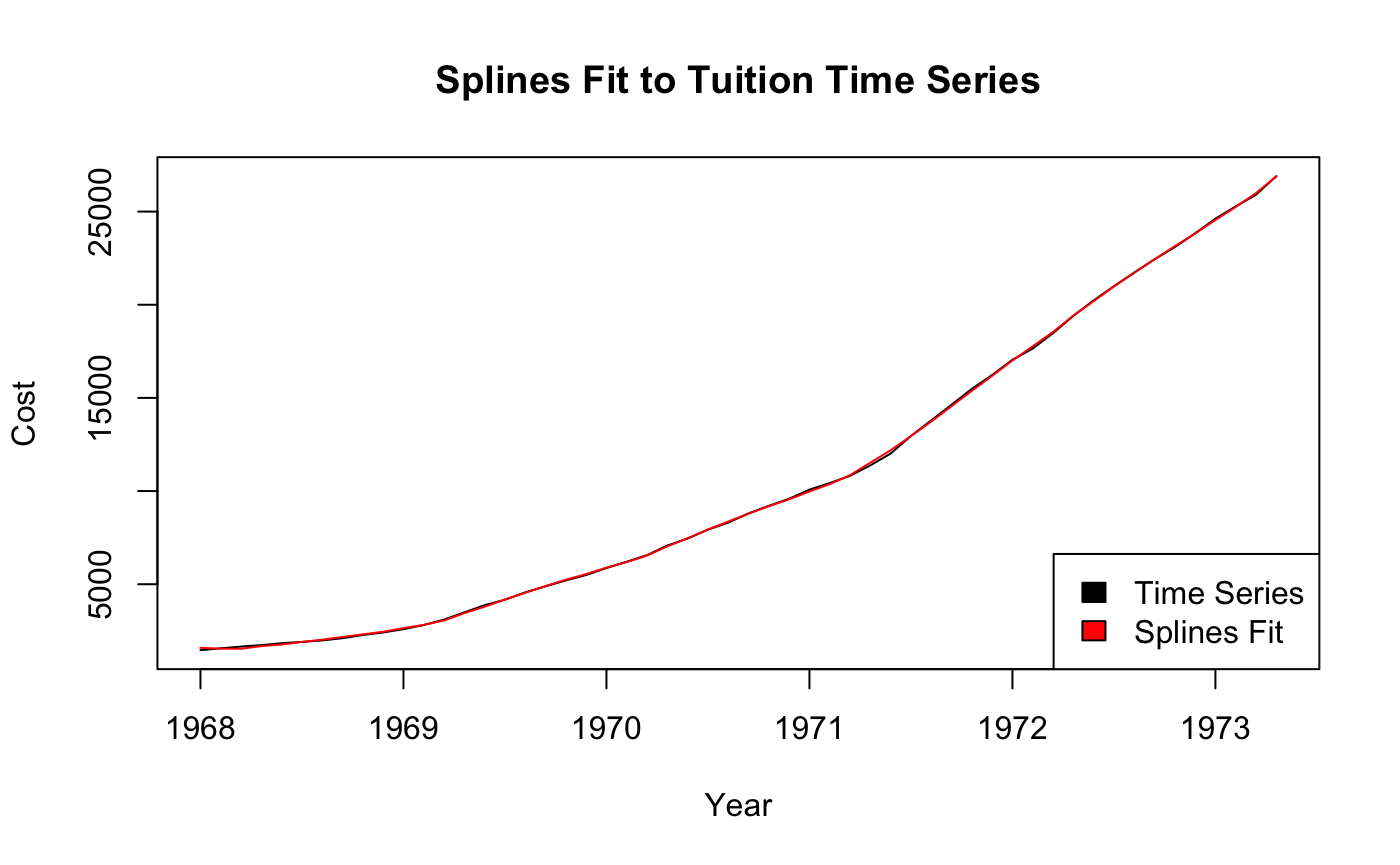


**Fig.4.** Time Series and ACF Plot for College Costs

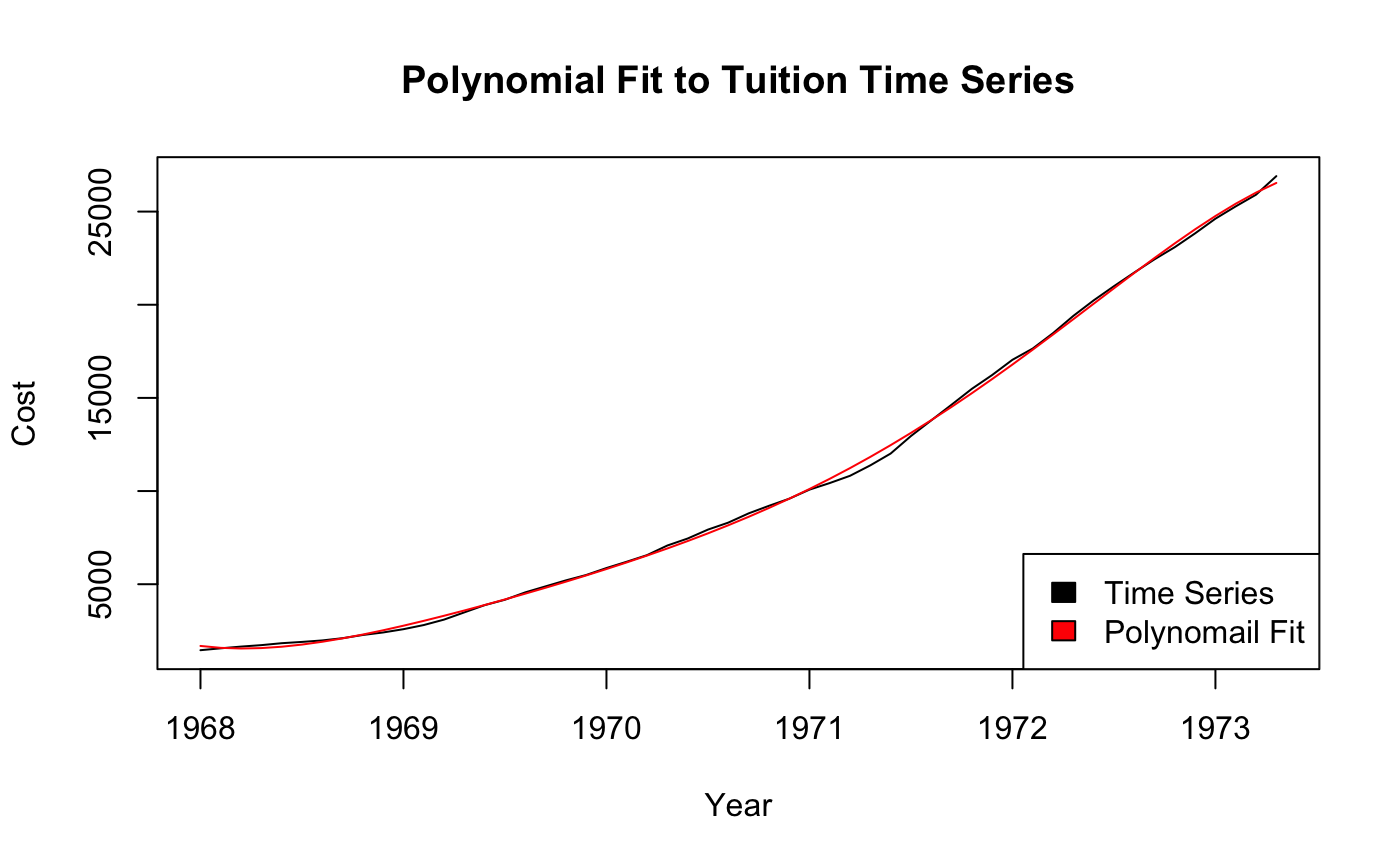
1. **Tracing College Costs from the 1960s to the 2010s**

The non-adjusted college cost (including tuition and fees, room, and board) is a strictly increasing time series from 1968 to 2021 as seen in figure 1. There is an obvious trend in the data and the autocorrelation function (ACF) plot of the time series confirms as much.

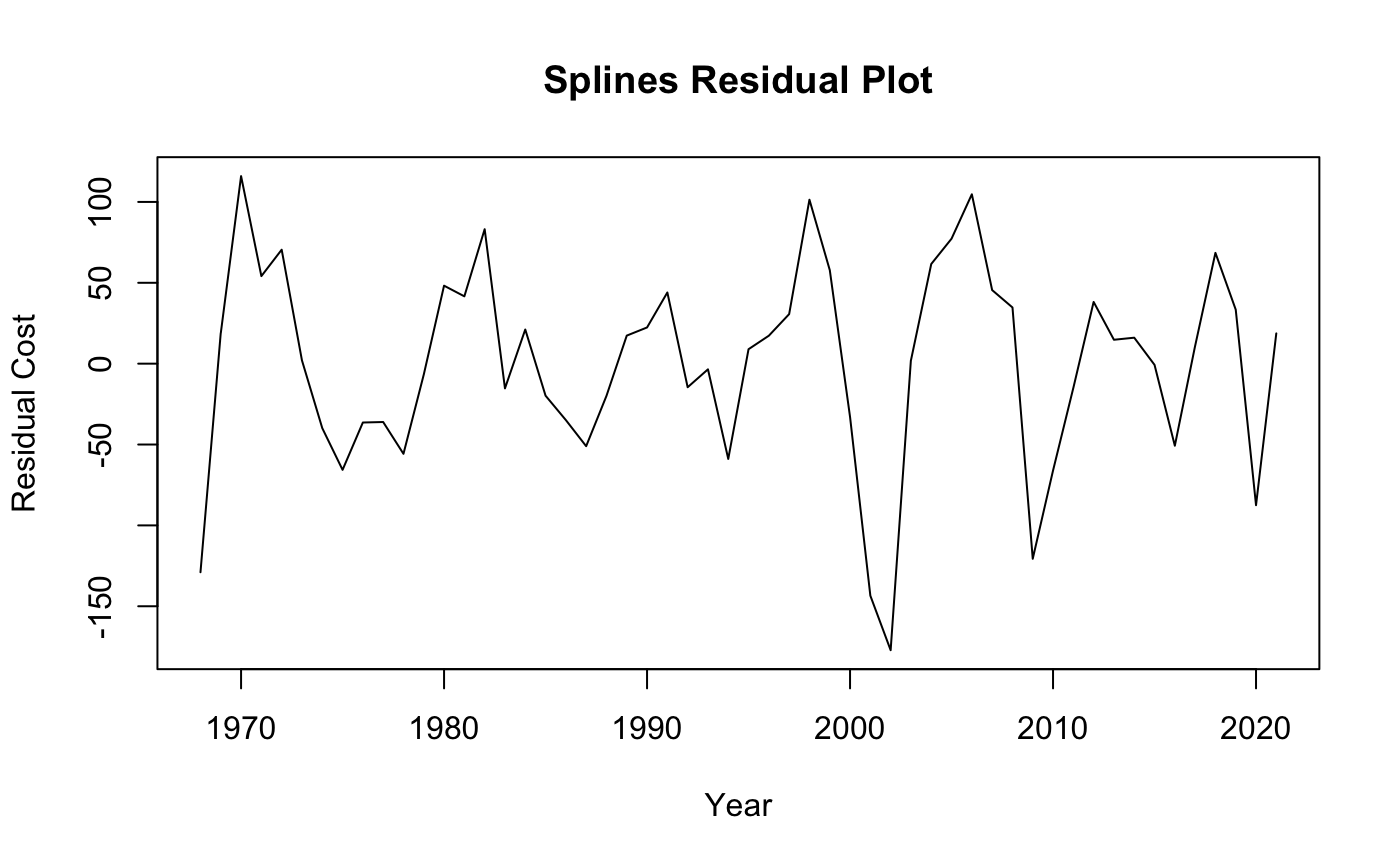
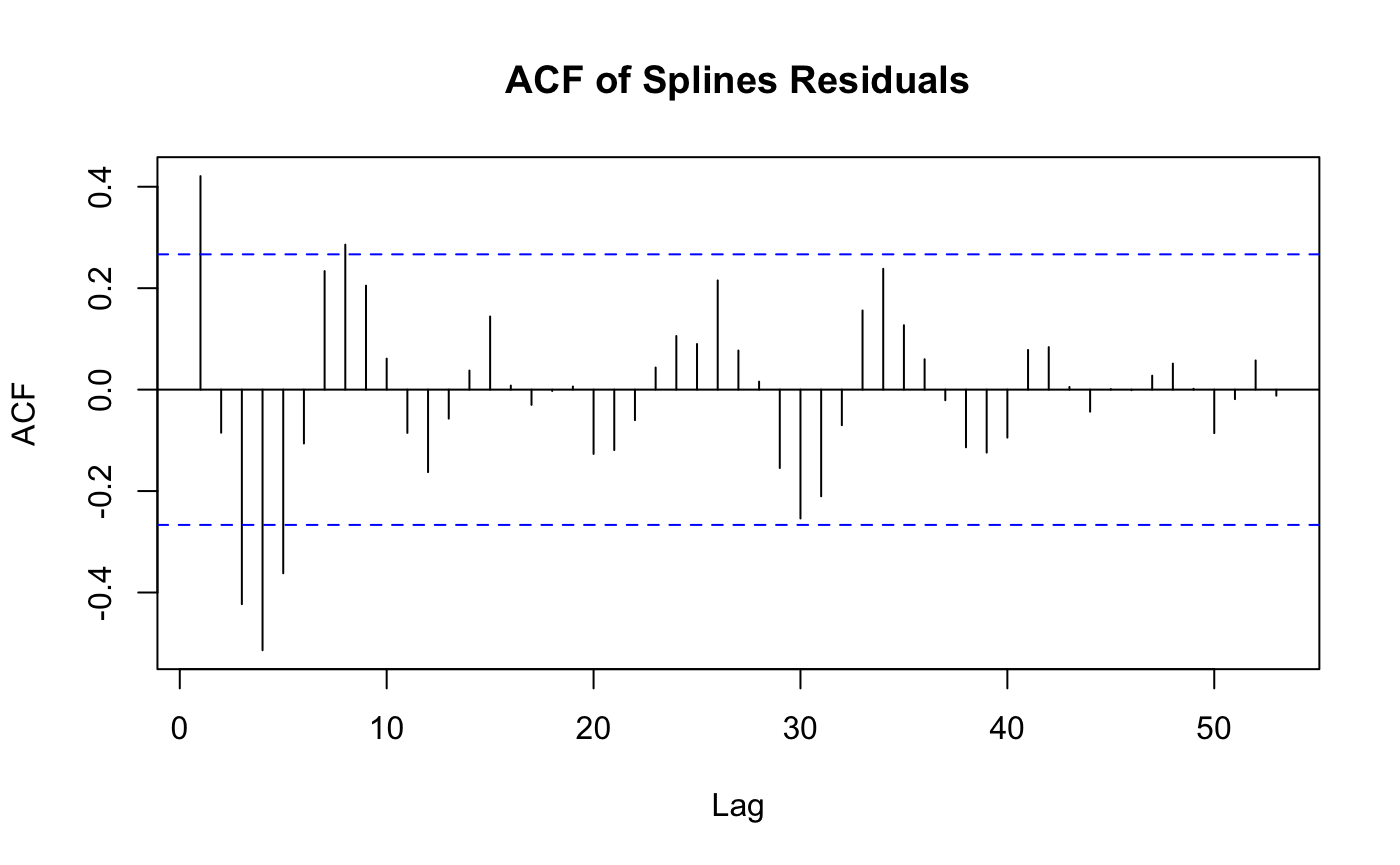
**Fig.1.** Time Series and ACF Plot for College Costs

In pursuit of discovering an appropriate trend estimation method, we also discovered a significant 10-year seasonality in the time series. We suspect that this seasonality, which was also observed in the starting salary time series, is related to economic fluctuations. This motivates the inclusion of exogenous factors in our multivariate analysis. As a result we will discuss two trend and seasonality functions: splines with 10-year ANOVA seasonality and a fifth degree polynomial fit. The splines model, detailed in figure 2, proved to be a good fit where the smooth terms were significant as well as every 3rd and 8th year proving to be significantly different from the intercept. 

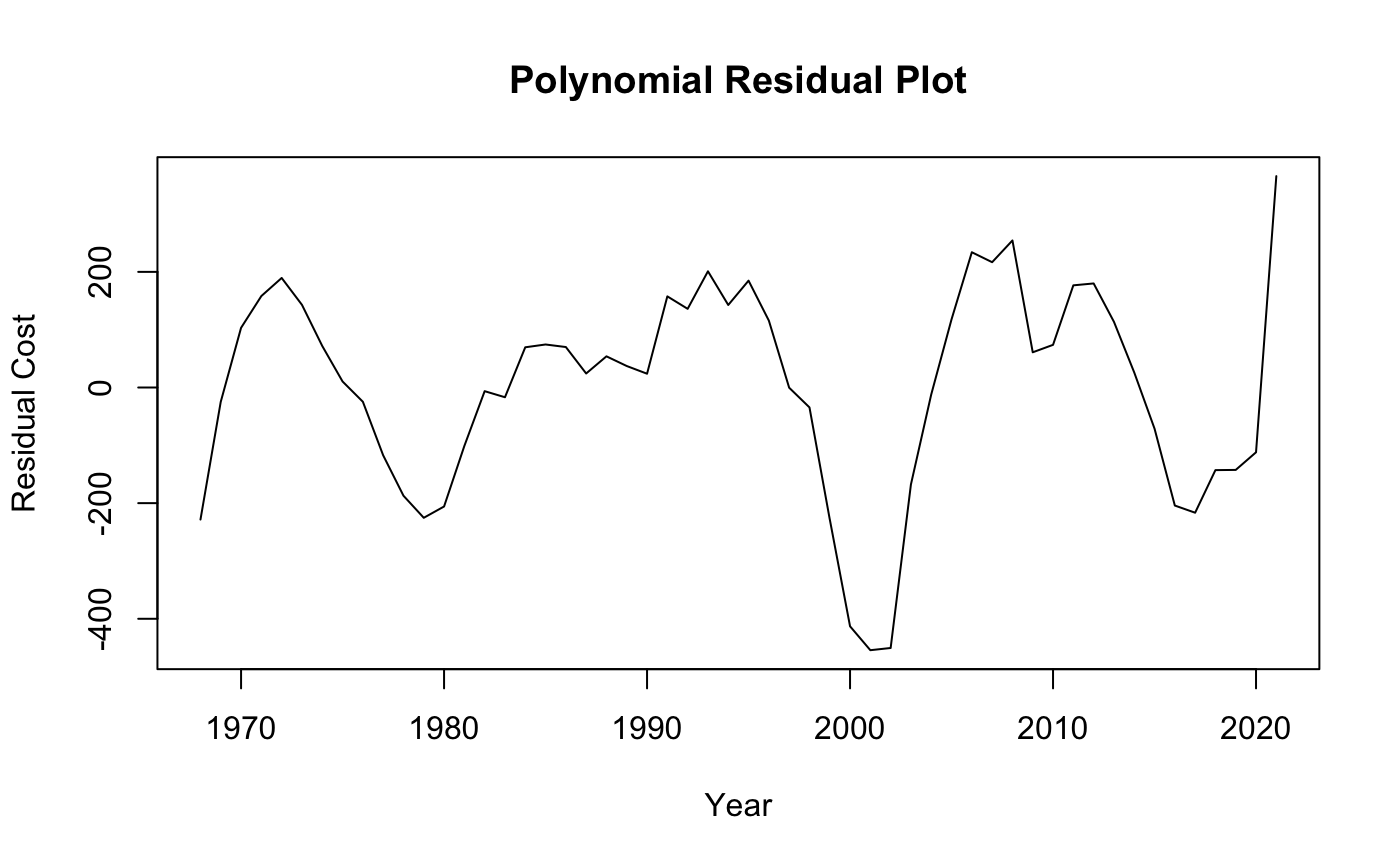
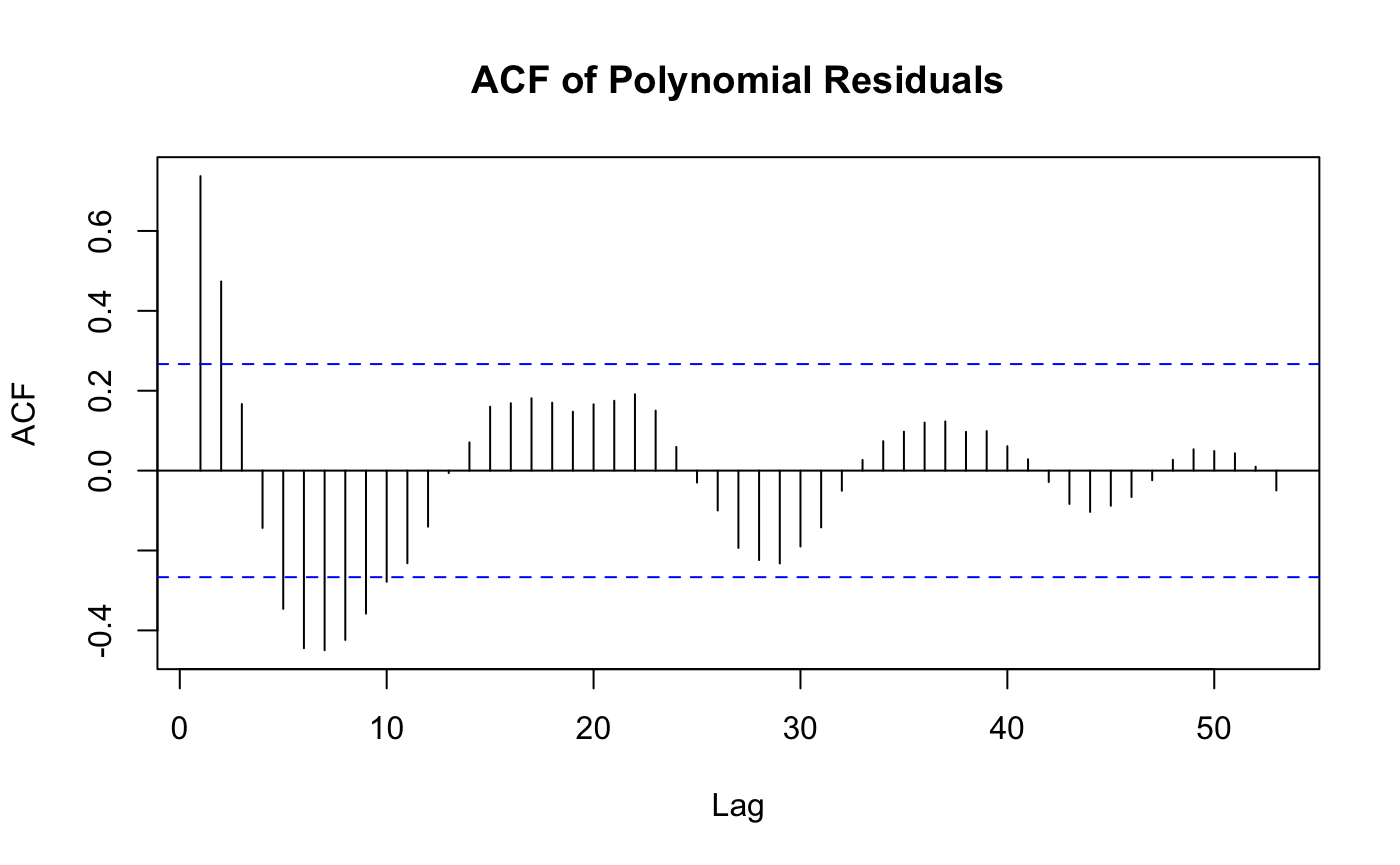
**Fig.2.** Fitted Splines Regression vs. College Costs

The polynomial fit provided similar results to the splines model, but did not include a seasonality term because the fifth degree polynomial already includes many terms. All factors were deemed to be significant, but there is concern about this model overfitting the data.

**Fig.3.** Fitted Polynomial Regression vs. College Costs

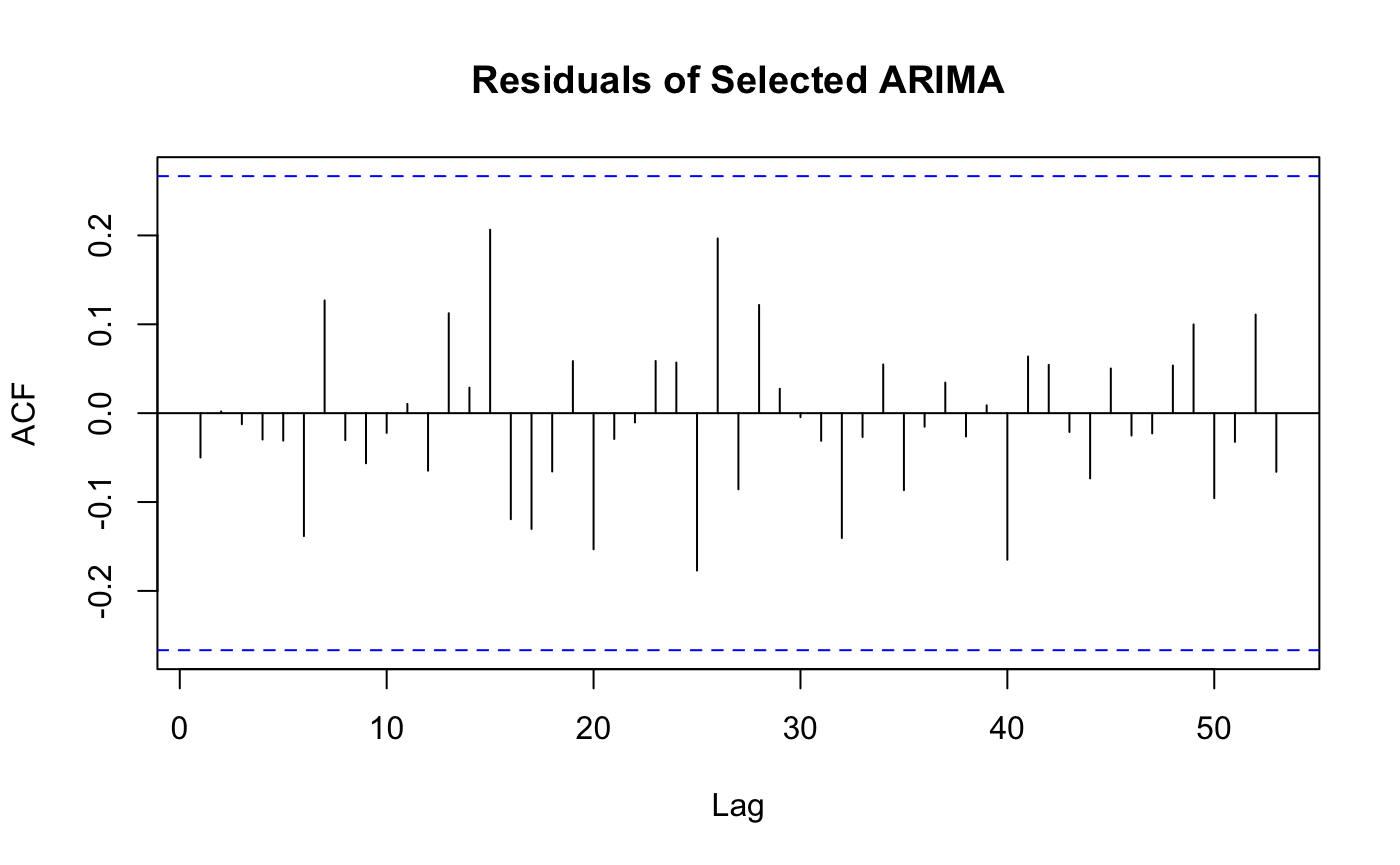
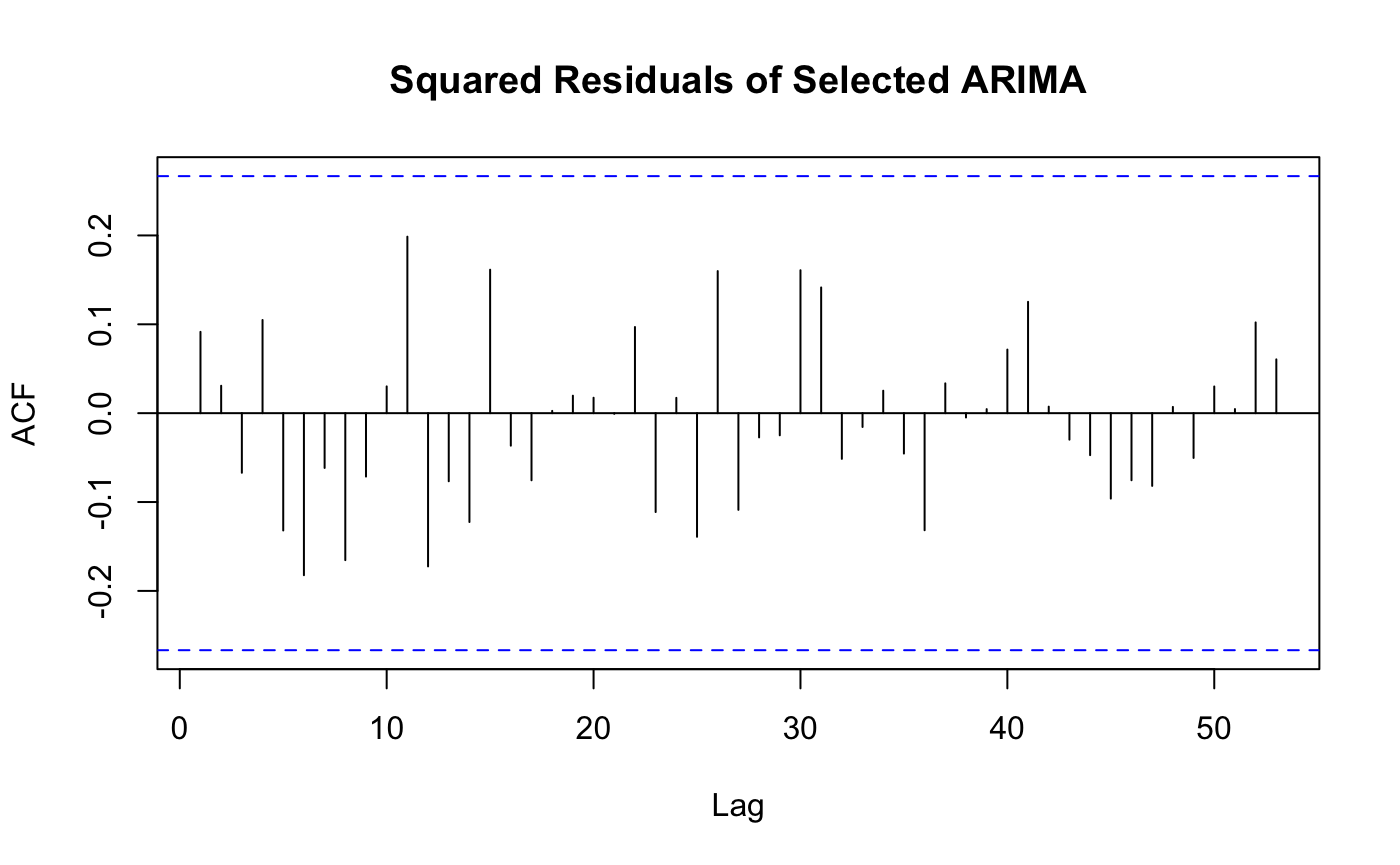
The residual plot and ACF plot of the splines regression and ANOVA seasonality residuals in figure 4 show plausible stationarity in the residuals, which is confirmed by an Augmented Dickey-Fuller (ADF) Test.

**Fig.4.** Residual and ACF Plot for Splines Trend and ANOVA Seasonality Estimation

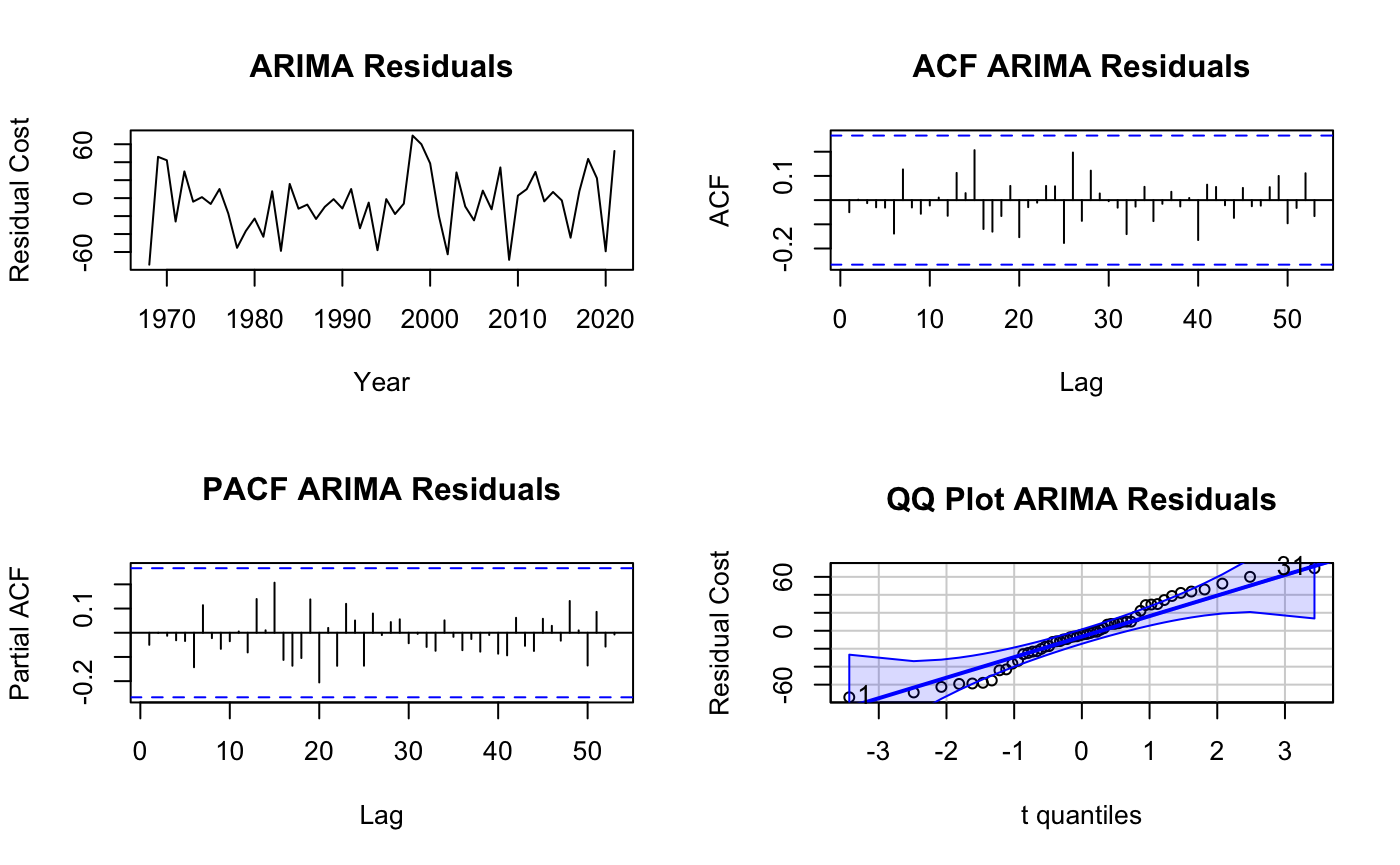
Based on the residual and ACF plots of the polynomial fit in figure 5 and an ADF test, we again conclude that the residuals are stationary.

**Fig.5.** Residual and ACF Plot for Fifth Degree Polynomial Trend Estimation

Due to our concerns about the polynomial model overfitting, we chose to proceed with the splines model residuals for further analysis.

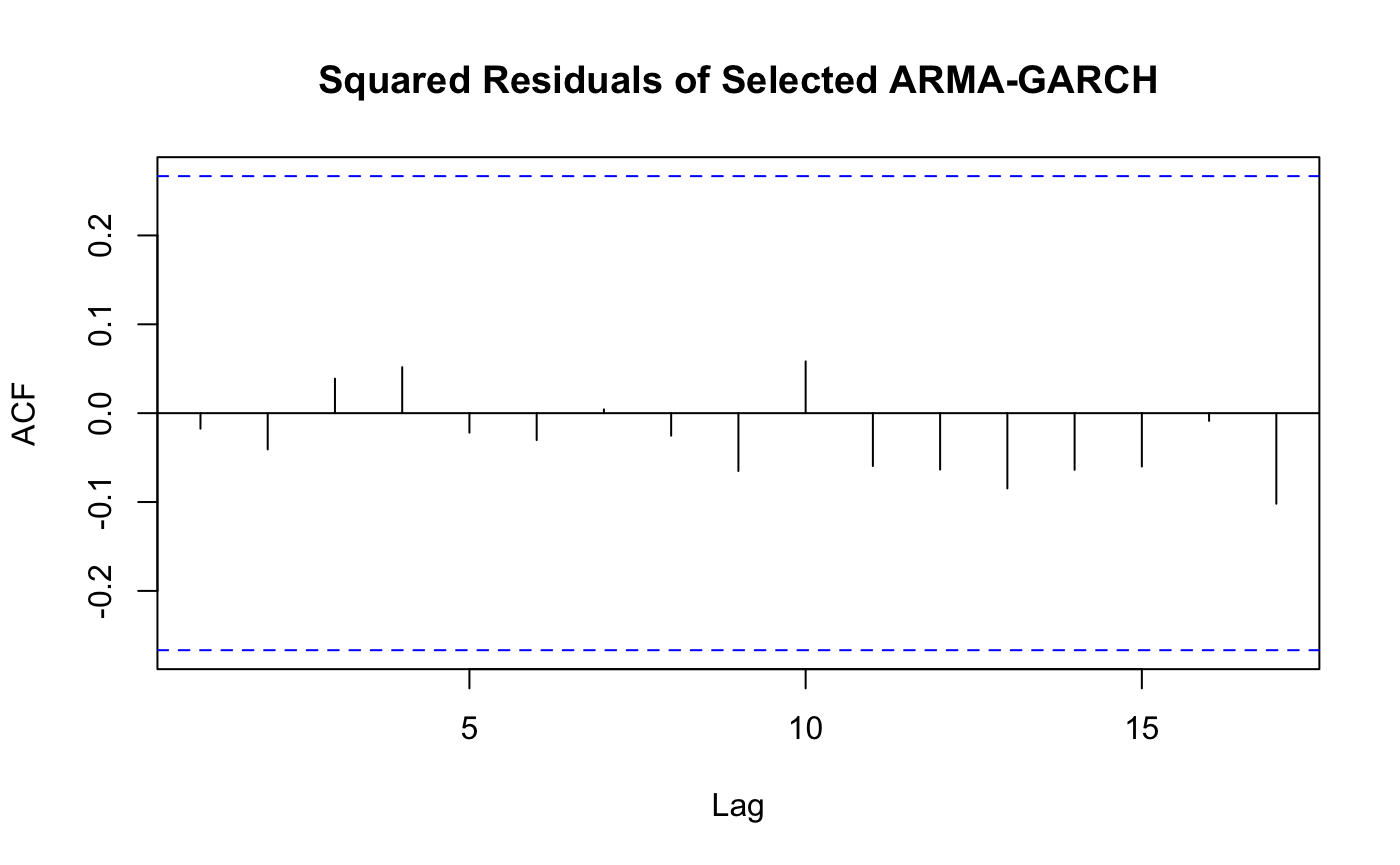
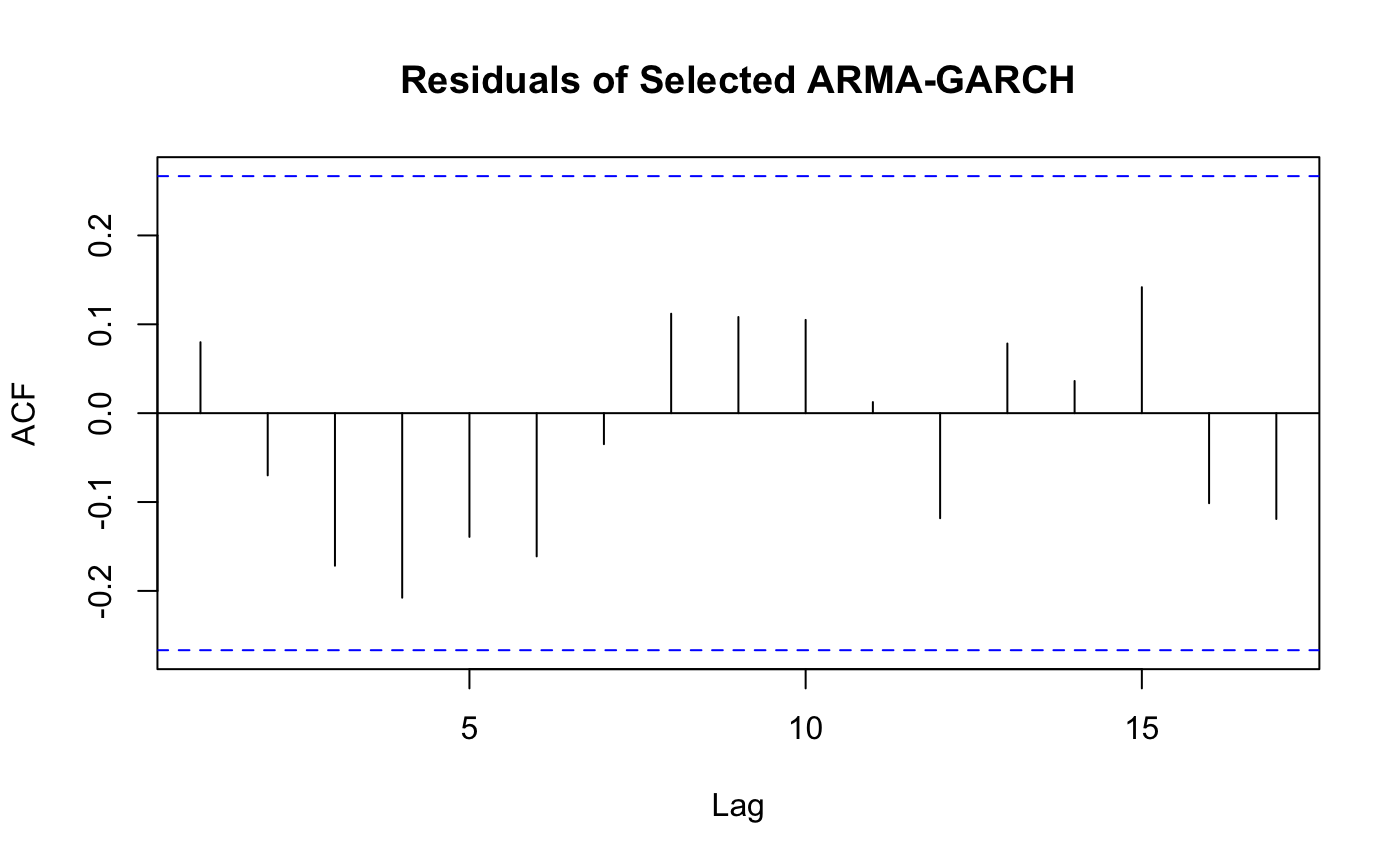
With these residuals, order selection deemed an ARIMA(2,0,3) to be an appropriate fit, however, the AR3 coefficient is not significant. Based on a Shapiro-Wilk test of the ARIMA residuals, the model residual normality assumption is satisfied. Further, a Box-Ljung test confirms the lack of serial correlation that the ACF plot shown in figure 6 shows. However, a Box-Ljung test of the squared residuals contradicts the ACF of the squared residuals, suggesting that there may be heteroskedasticity in the residuals.

**Fig.6.** ACF Plot of ARIMA(2,0,3) Residuals and Squared Residuals

In figure 7, we see that the residual plot and PACF agree with the assertion that there is no serial correlation, but the QQ Plot against a t-distribution with 5 degrees of freedom does raise some concerns regarding the normality assumption.

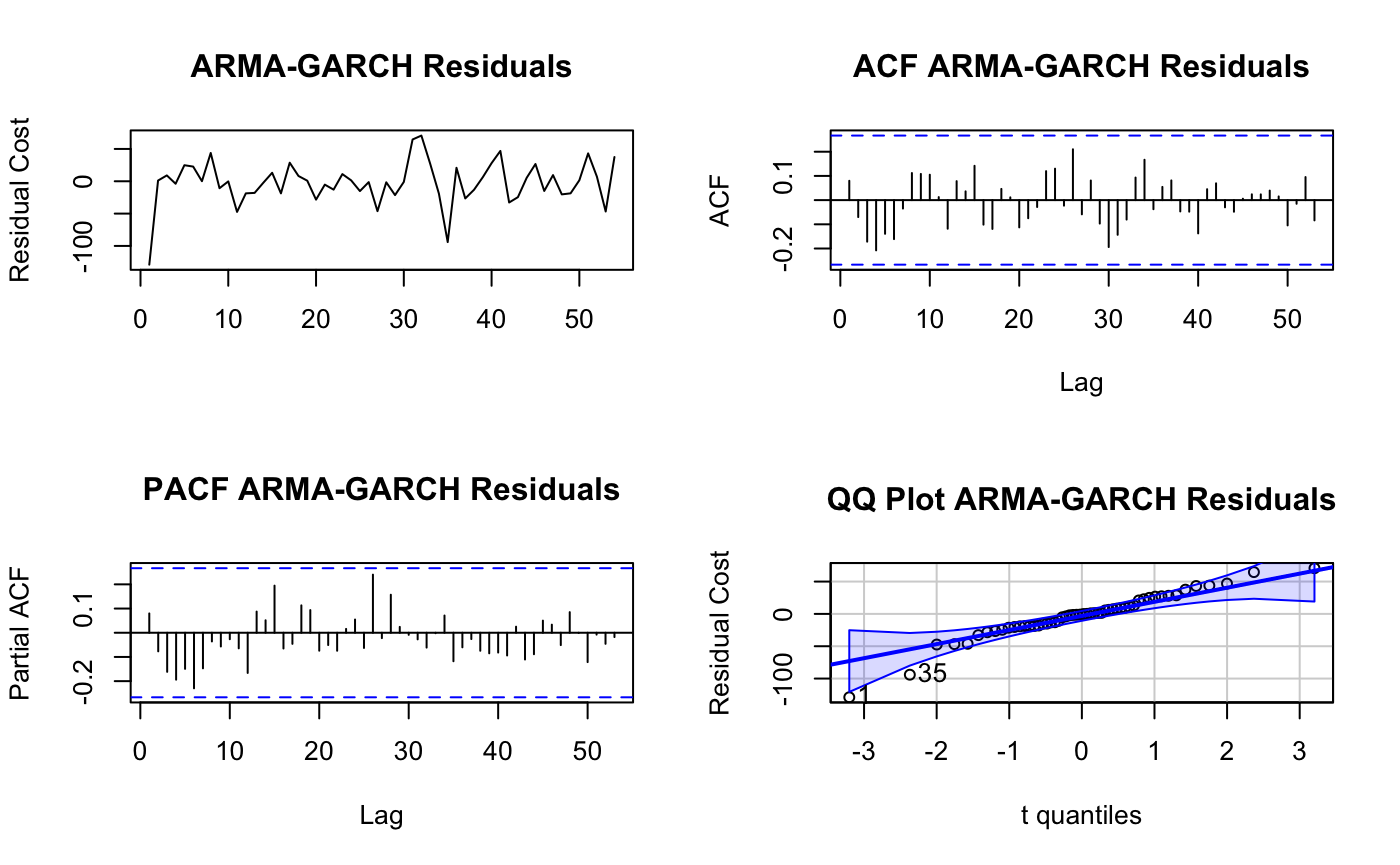
**Fig.7.** ARIMA(2,0,3) Residual Plot, ACF, PACF, and QQ Plot

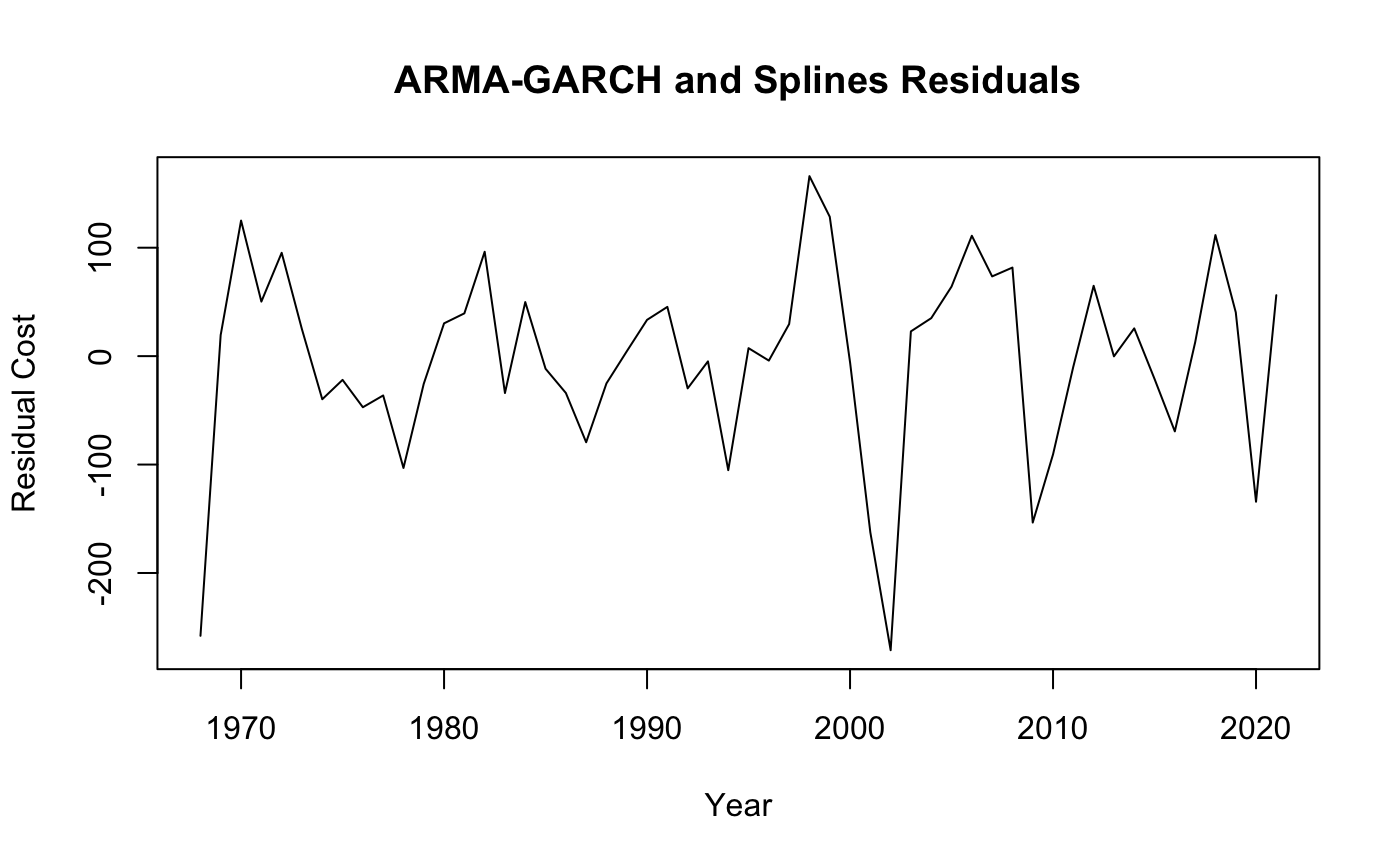
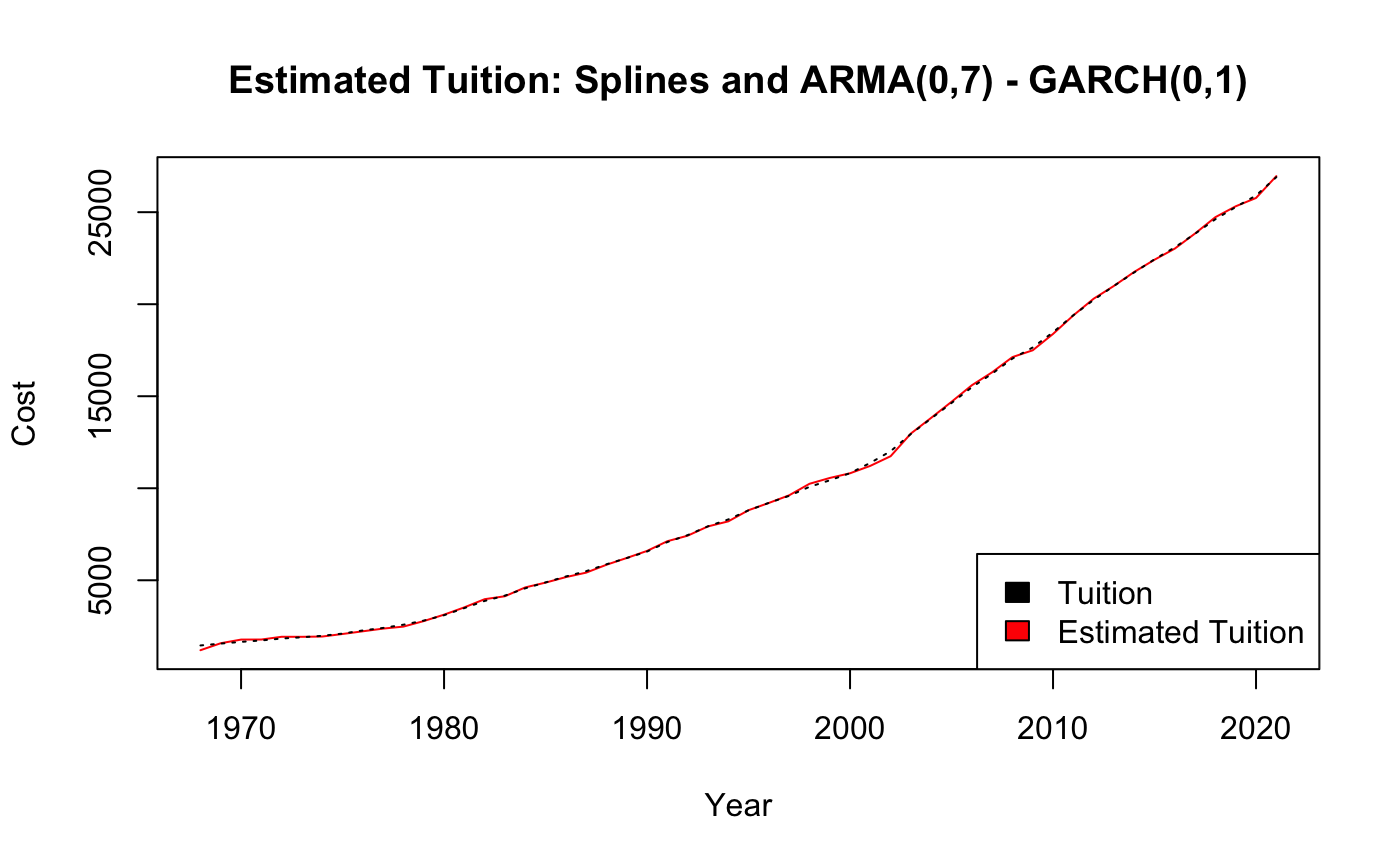
The potential heteroskedasticity from the Box-Ljung test on the squared residuals is motivation for an ARMA-GARCH fit. For the order selection of our ARMA-GARCH model, we initially fixed ARMA(2,3) and found that the best GARCH order based on AIC was 0, 1. Now fixing GARCH(0,1), ARMA(0,7) produced the optimal AIC. To conclude this process, the best GARCH order for a fixed ARMA(0,7) starting point was again GARCH(0,1). Thus, we fit the spline residuals with an ARMA(0,7)-GARCH(0,1) model, and all coefficients were deemed significant.

Based on figure 8, we observe that the ACF plot of the residuals shows potential serial correlation, but the squared residuals reveal little to no heteroskedasticity, as expected when using GARCH.

**Fig.8.** ACF Plot of ARMA(0,7)-GARCH(0,1) Residuals and Squared Residuals

The residual plot and PACF from figure 9 reaffirm our earlier concern that the residuals may be serially correlated. A Box-Ljung test also confirms that the residuals are correlated. The QQ Plot in figure 9 also agrees with a Shapiro-Wilk Normality test that the residuals are not normal, but they may conform to a t-distribution that was used in the ARMA-GARCH fit. Thus, the lack of normality does not violate any assumptions of our model. Lastly, a Box-Ljung test of the squared residuals reaches the same conclusion as the squared residual ACF plot that heteroskedasticity in the data was removed.

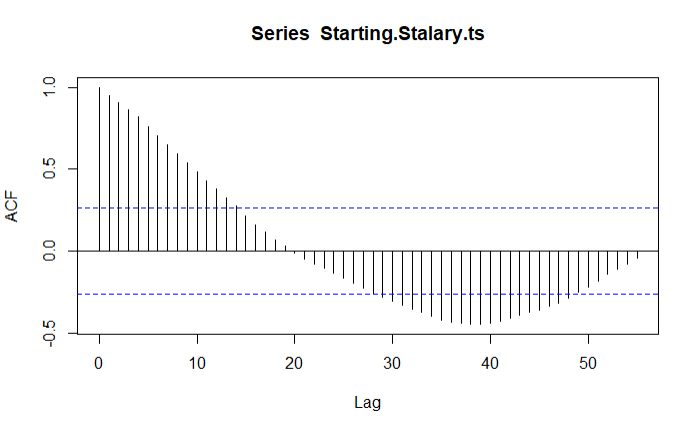
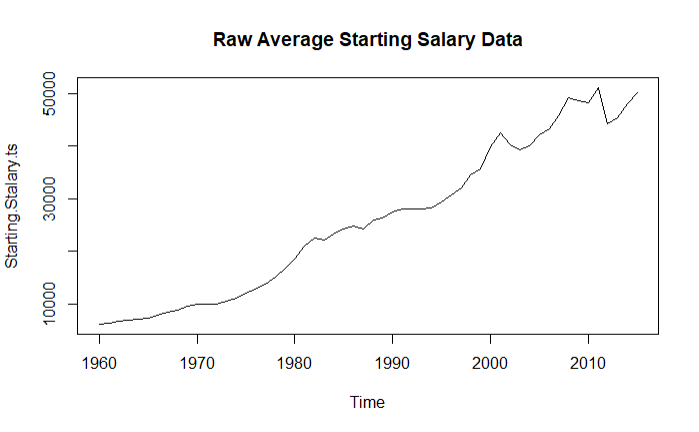
**Fig.9.** ARMA(0,7)-GARCH(0,1) Residual Plot, ACF, PACF, and QQ Plot

Finally, in figure 10 we observe how the fitted splines plus seasonality residuals combine with the ARMA-GARCH residuals to estimate the college cost time series rather well. The employed methods are able to estimate residuals to generally within $150 of actual college prices.

**Fig.10.** Combined ARMA-GARCH and Splines Residuals (left), Fit to Cost Time Series (right)

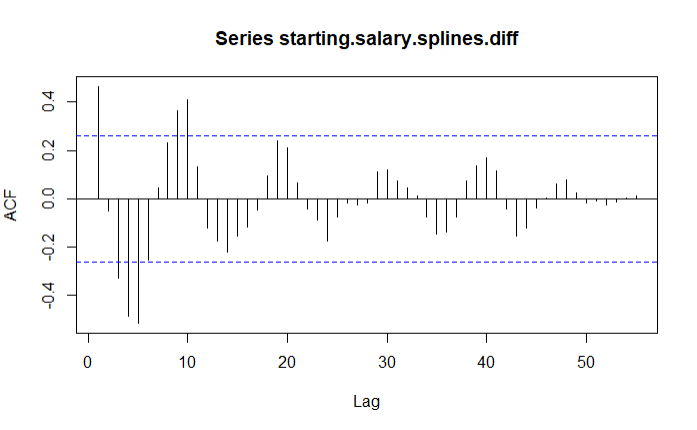
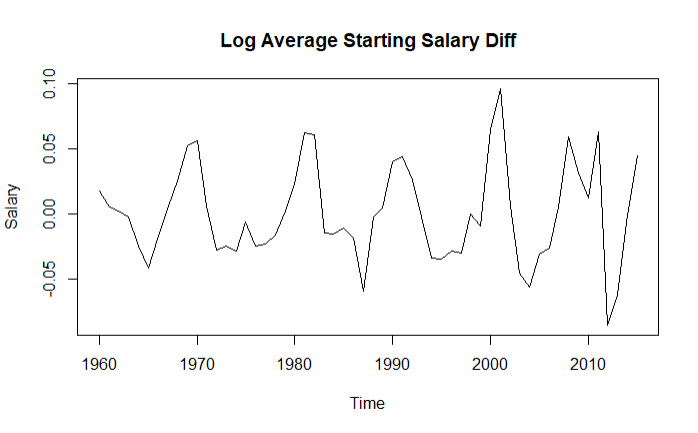
1. **Tracing Average Starting Salary from the 1960s-2010s**

The analysis of the average starting salary data showed a clear upward trend in the data and a clear trend from its ACF plot (Fig.11.).



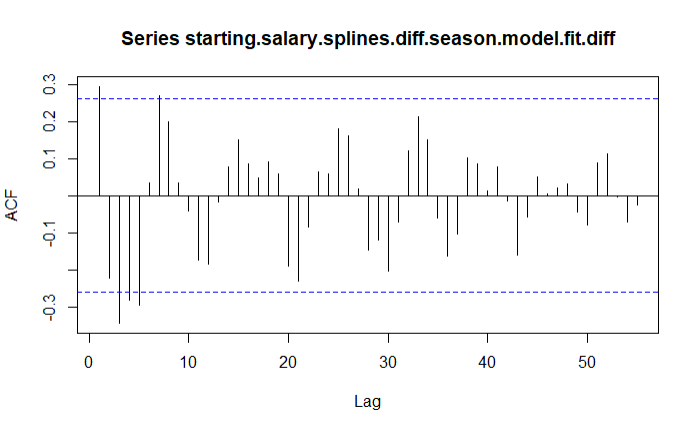
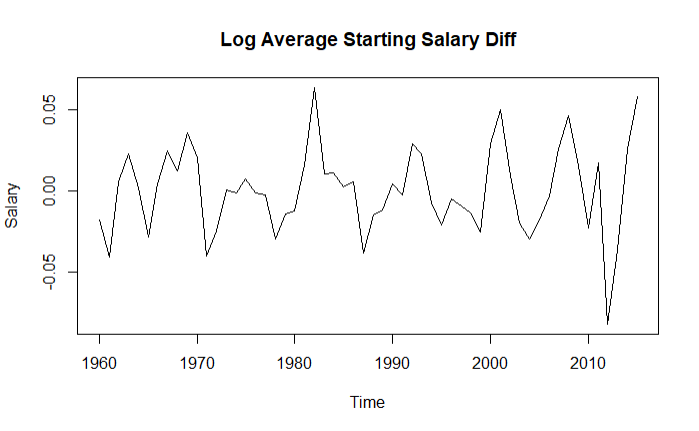
**Fig.11.** Trend Analysis of Average Starting Salary

We decided upon a splines fit for initial analysis of the residuals, this led us to take the logarithm of the data in order to see better stationarity in the residuals (Fig.12.). This splines fit was very statistically significant. When looking at the data, we determined that there must be some degree of seasonality to the data given the shape of our residuals and the sino-soidal pattern in the ACF of the residuals.

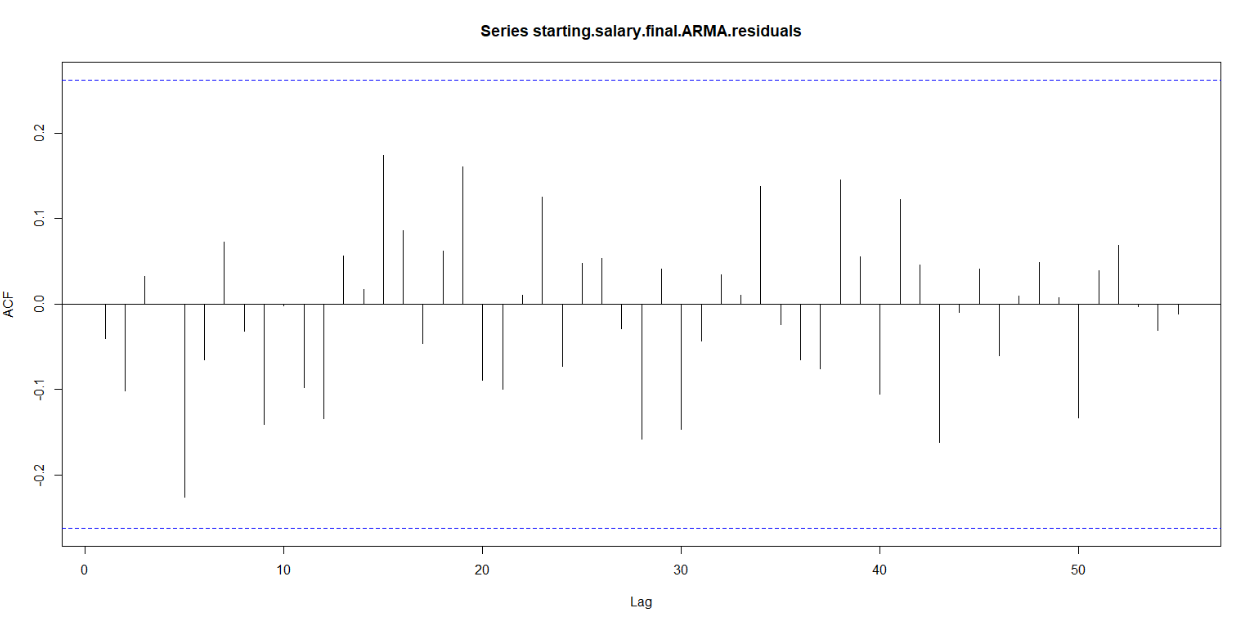


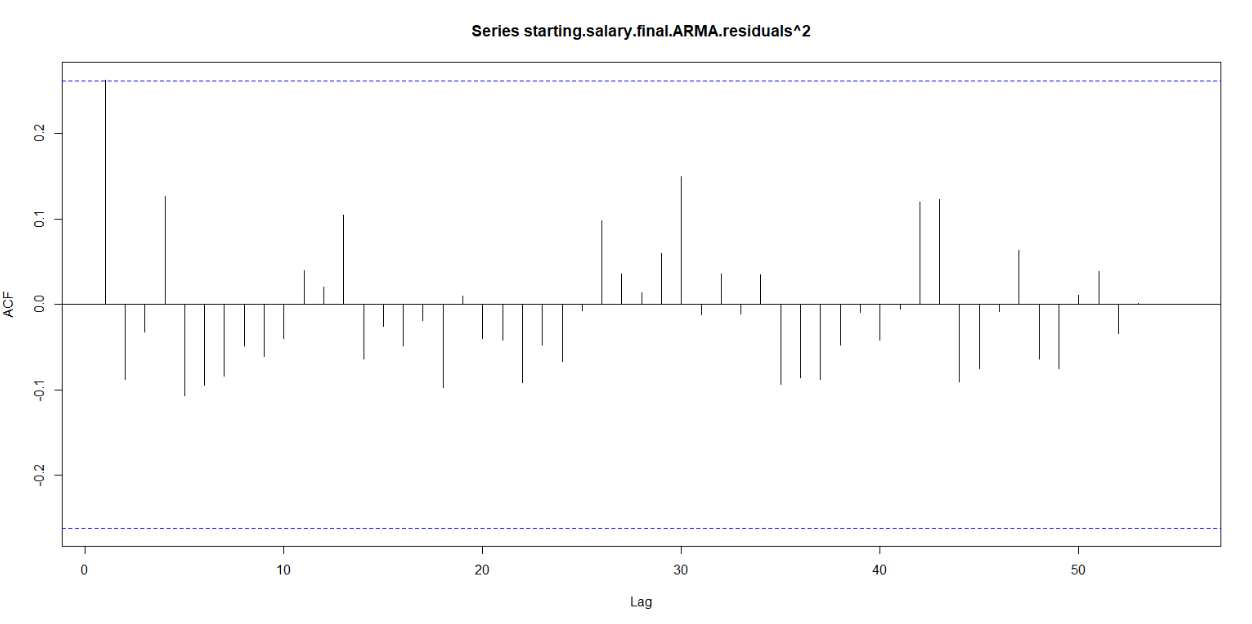
**Fig.12.** Trend Residual Analysis

Given our normal residuals plot, we thought that perhaps there is some sort of decade-based seasonality. This caused us to fit a seasonal trend based on a 10 year time period. The resulting seasonal trend showed statistically significant coefficients for 7/10 of the added coefficients (based on a p-value of 0.05). The resulting residuals also showed a higher degree of stationarity (Fig.13.), specifically when thinking about p-values of greater than 0.05.



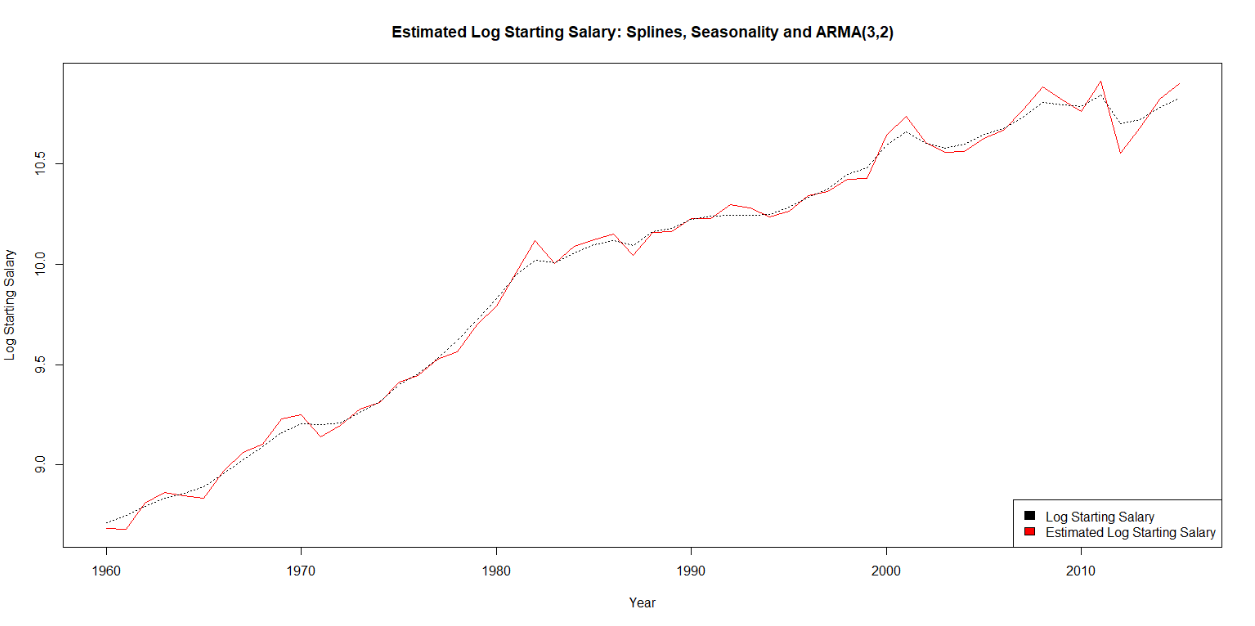
**Fig.13.** Trend and Seasonality Residual Analysis

With these residuals, order selection deemed an ARIMA(3,0,2) to be an appropriate fit. We tested our resulting residuals using the Box-Pierce and Box-Ljung tests. The resulting tests informed us that the residuals were likely to be serially correlated. The corresponding residuals were normally distributed (tested visually and using the Shapiro-Wilks normality test). The squared residuals were shown to be also serially correlated. Despite this, the corresponding ACF and PACF plots of the residuals showed signs of stationarity (Fig.14.).



**Fig.14.** ARMA Residuals (Left) and Squared Residuals (Right)

Furthermore, we had tested for an ARMA-GARCH model of order (5,0,5)x(1,0,0). This model did not provide us with any better fit than the ARMA model previously, and in-fact shows evidence of going against the ARMA-GARCH assumption of the t-distribution of the residuals. Checking the goodness of fit using our ARMA model (Fig. 15.), we see that the model is a fairly strong fit.



**Fig.15.** Splines + Seasonality + ARMA Fit to Average Starting Salary Time Series

1. **Multivariate Relationships**

When we were looking to fit a multivariate model to our data, we fit all of our data with splines and seasonal fits where necessary in order to reduce our data to a stationary state. We then tested each of our residuals with an ADF test for stationarity, rejecting the null for all cases (indicating stationarity). The resulting residuals can be seen in the Appendix: Var Model Analysis. All of the resulting VAR models were unfortunately unstable (covered in Discussion). Despite this, we decided to use the resulting restricted VAR(5) model for all of the data (no exogenous data) as it performed the best against our hypothesis tests, showing evidence of no heteroskedasticity, evidence of normality, and no evidence of serial correlation against a 0.01 p-value (Appendix: Hypothesis Tests for Chosen Restricted VAR model).

**i. Exploring the Relationship Between Total Loans and Other Factors**

We found that the log total loan data was significantly correlated with the:

1. Lag 1, 3, and 5 of the tuition cost data
2. Lag 4 and 5 of the starting salary data
3. Lag 5 of itself (the total loan data)

**ii. Exploring the Relationship Between Tuition (College) Costs and Other Factors**

We found that the tuition data was significantly correlated with the:

1. Lag 1, 2, and 3 of the starting salary data
2. Lag 1 and 2 of the inflation data
3. Lag 3 and 4 of the total loan data
4. Lag 3 and 5 of the GDP Growth data
5. A constant value (from the VAR model specification)
6. Lag 1, 4, and 5 of itself (the tuition costs)

**iii. Exploring the Relationship Between Average Starting Salary and Other Factors**

We found that the log average starting salary data was significantly correlated with the:

1. Lag 1 and 3 of the total loans data
2. Lag 1, 3, and 5 of the inflation data
3. Lag 2 of the population data
4. Lag 3 of the GDP growth data
5. Lag 2 and 4 of itself (the starting salary data)

**iv. Exploring the Relationships with Inflation and Other Factors**

We found that the inflation data was significantly correlated with the:

1. Lag 1 and 3 of the tuition data
2. Lag 2, 3, and 4 of the total loans data
3. Lag 2 of the GDP growth data
4. Lag 2, 3, and 5 of the population data
5. Lag 3 and 4 of the starting salary data
6. A constant value (from the VAR model specification)
7. Lag 1, 2, and 4 of itself (the inflation data)

**v. Exploring the Relationships with GDP Growth and Other Factors**

We found that the GDP growth data was significantly correlated with the:

1. Lag 1 and 3 of the tuition data
2. Lag 1, 2, 3, and 4 of the inflation data
3. Lag 2, 3, and 4 of the starting salary data
4. Lag 2 of the total loans data
5. Lag 2, 3, and 5 of the population data
6. A constant value (from the VAR model specification)
7. Lag 4 of itself (the GDP growth data)

**vi. Exploring the Relationships with Population and Other Factors**

We found that the log starting salary data was significantly correlated with the:

1. Lag 2 of the inflation data
2. Lag 2 and 4 of the GDP growth data
3. Lag 3 of the total loans data
4. Lag 3 and 4 of the tuition cost data
5. Lag 5 of the starting salary data
6. Lag 1 and 3 of itself (the population data)

Note that the above information can be found in Appendix: Var Model Analysis, all of the significance (and the formulation of the restricted VAR model) was using a 0.05 significance level.

***V. Discussion***

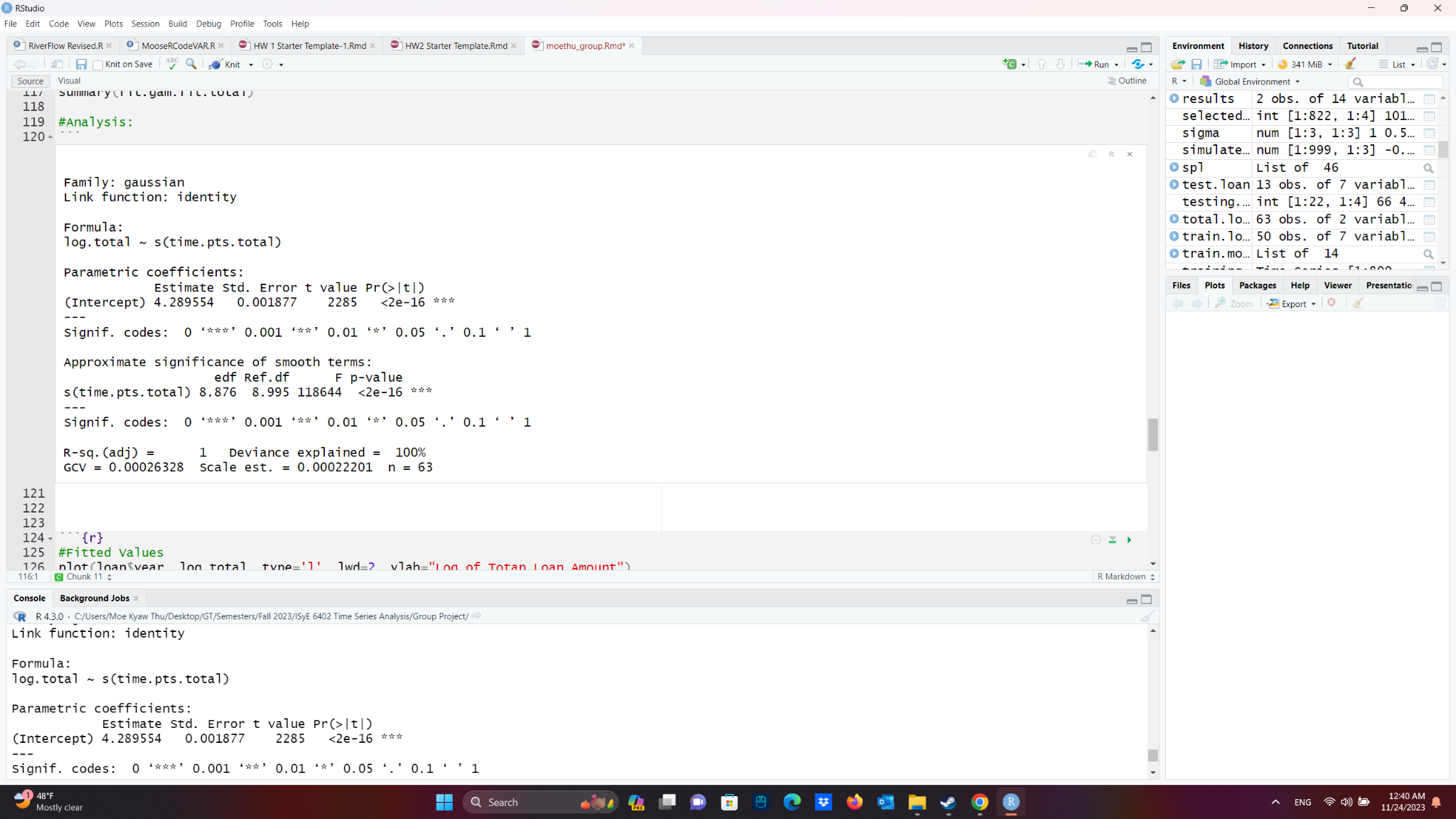
By exploring these datasets, we were able to compare our prior understanding of the data with our new understanding. Initially, we suspected all of our data to be simply trend-stationary. This was proven incorrect with our findings that the Tuition data, Starting Salary data, and the Population data contained a 10 period seasonality. In addition, the population data contained even more seasonality signs that were difficult to remove (a seasonal trend that seemed to reduce its period over time). Related to the VAR model, we expected the best model to be a VARX model with the GDP growth, Inflation, and Population data as exogenous variables. Rather than this, we found that a general VAR model had the best fit based on our hypothesis testing procedures. Briefly ignoring the fact that our VAR models were unstable, some of the relationships between the data were unexpected, such as the Population data lagging Starting Salary data, Inflation, and GDP growth data, but not lagging Total Loans and Tuition costs. This indicated that perhaps looking further at Population relationships with these factors may be an interesting area, specifically looking at nations with less growth than the US.

In the course of our work, we found that there were quite a few limitations. Our data was quite disparate, that is we had a few datasets that started almost a decade later than our other datasets (1968 rather than 1960) and ended almost a decade sooner than our other datasets (2015 rather than 2022). This resulted in us having less than 50 total data-points for our VAR model and our other models. Given that time series analysis improves with the length of the data, our models are very likely not the best, especially given the instability of our VAR models. Most of our planned exogenous variables (GDP growth, Inflation, and Population) are actually rates of change year-to-year, our inability to gather rates of change year-to-year of College Student Population for 1960s-2020s may be an important factor in the VAR model instability.

Overall, while we think our findings, specifically on the seasonal components of the individual times series, were eye-opening, we believe that further research into the relationships between our data should be explored. This could be done via further multivariate time series models, regression modeling, or advanced machine-learning modeling. Given some more clear data and further analysis, we would also want to see tests of the granger causal relationships between the data on a stable multivariate model to further understand the relationships within our data.

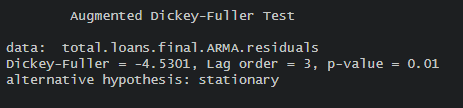
***VI. Appendix***

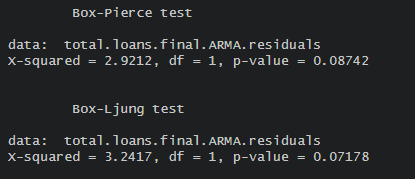
**Appendix: Student Loan Trend and Seasonal Analysis**

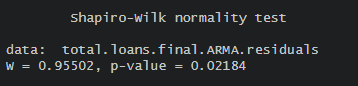


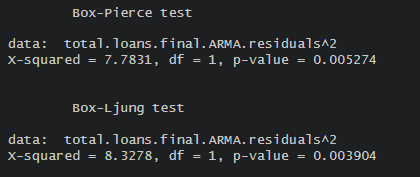
**Table. 1.** Spline Regression Output for Student Loan Data

**Appendix: Total Loans ARMA Hypothesis Testing**

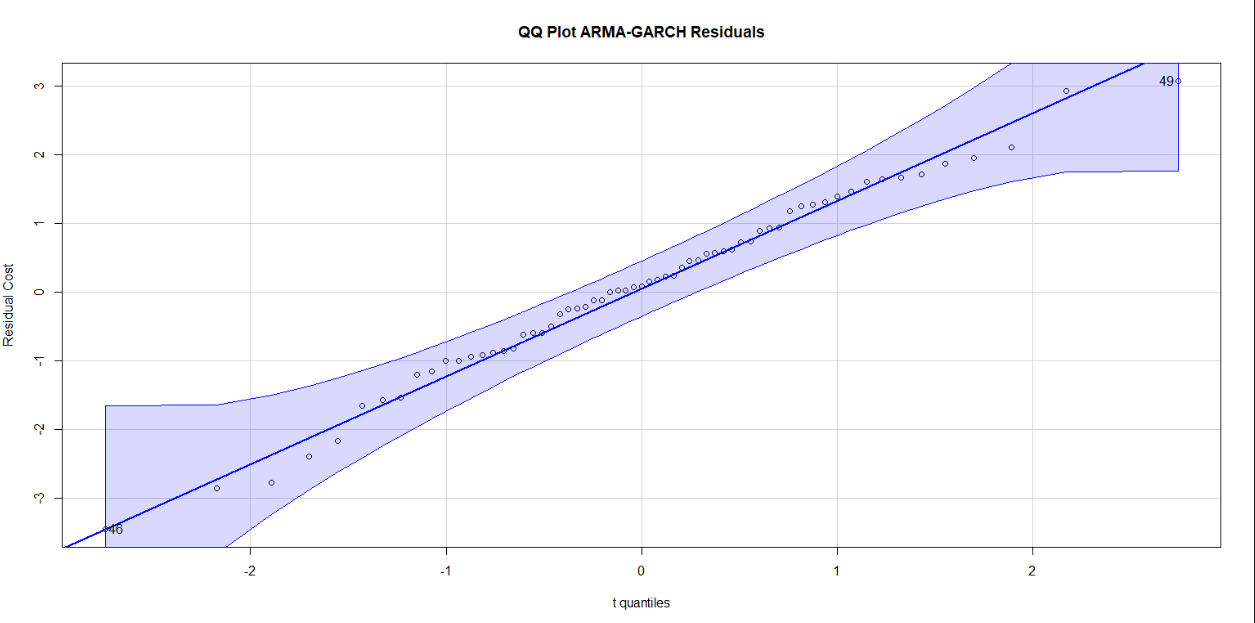




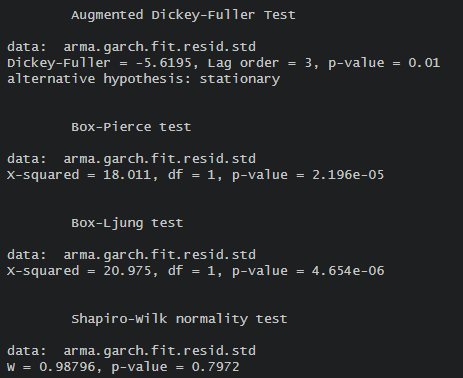




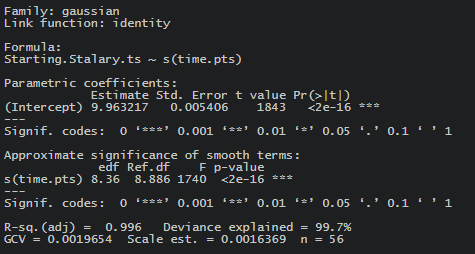
**Appendix: ARMA-GARCH of Total Loans data**



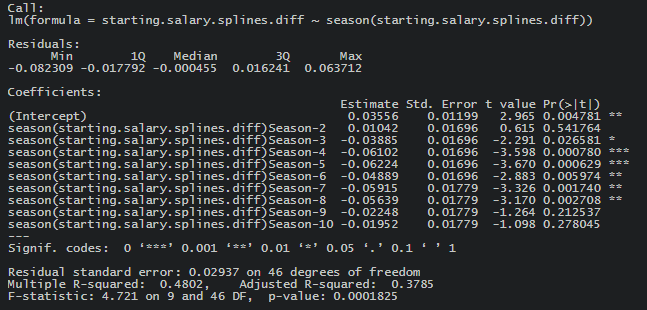
**Appendix: ARMA-GARCH of Total Loans data Hypothesis Tests**



**Appendix: Starting Salary Trend and Seasonal Analysis**

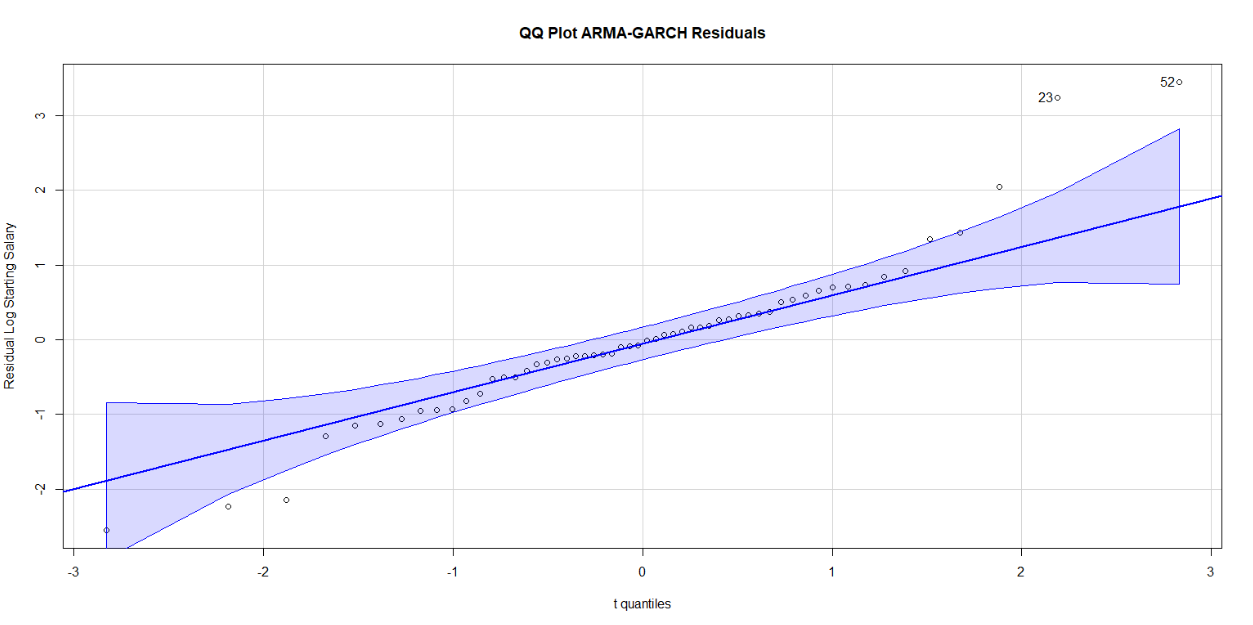


**Table. 2.** Spline Regression Output for Log Starting Salary Data



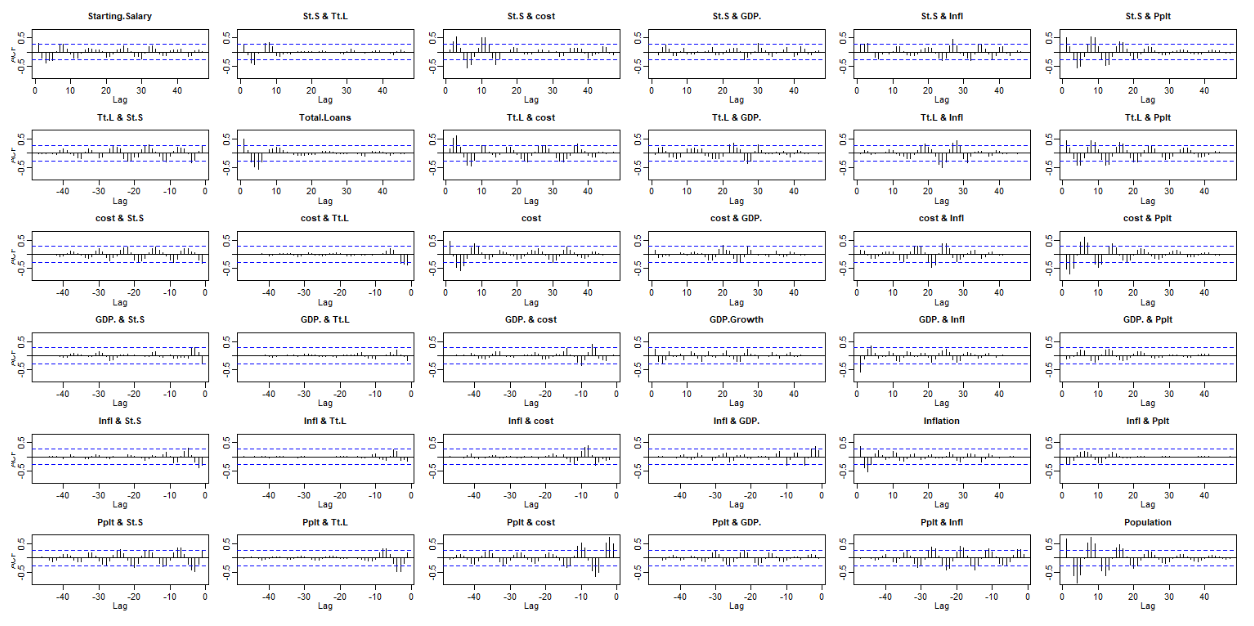
**Table. 3.** 10-Period Seasonal Regression Output for Differenced Log Starting Salary Data

**Appendix: ARMA-GARCH of Starting Salary data**

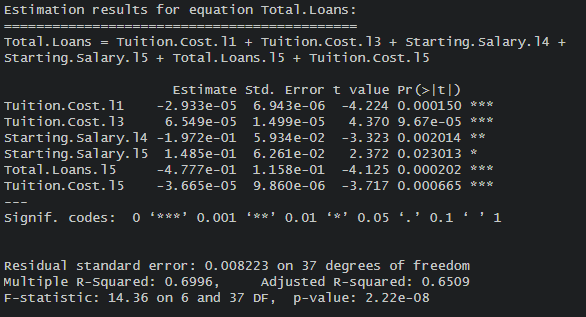


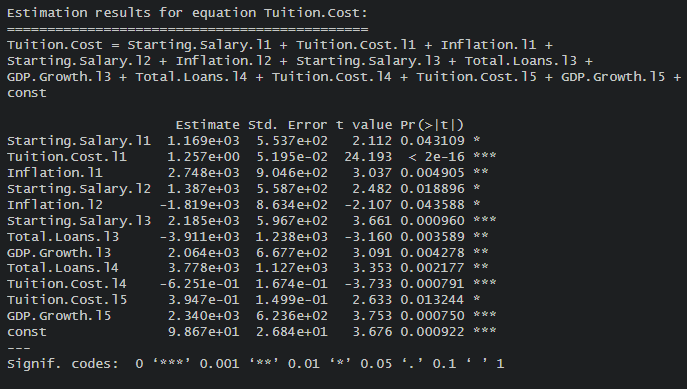
**Figure 1.** ARMA-GARCH Residuals of the Log Starting Salary data

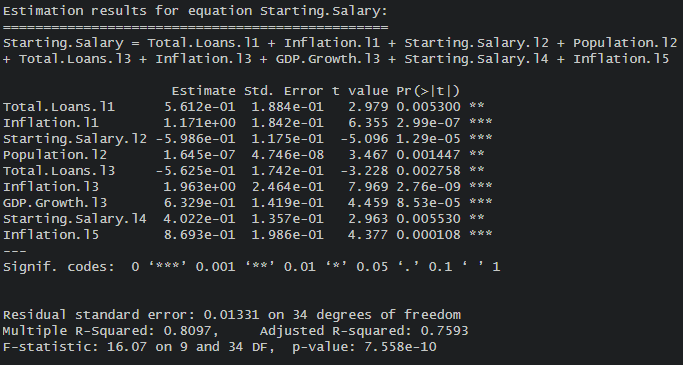
**Appendix: VAR Model Analysis**

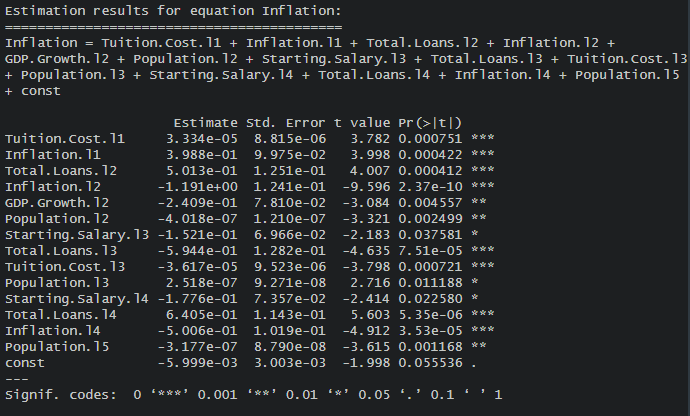


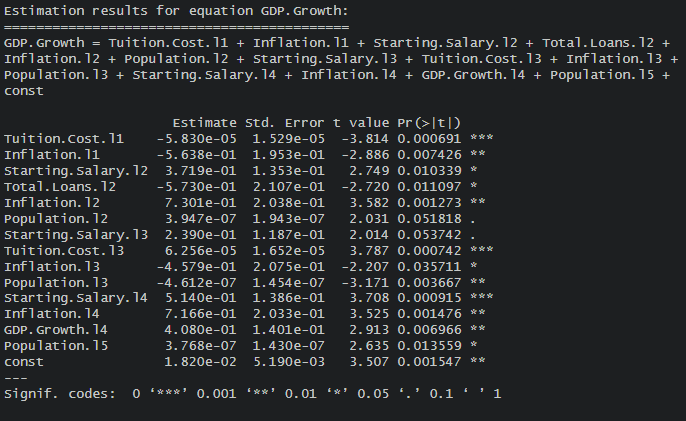
**Figure 2.** Residual process of all data (Before VAR)

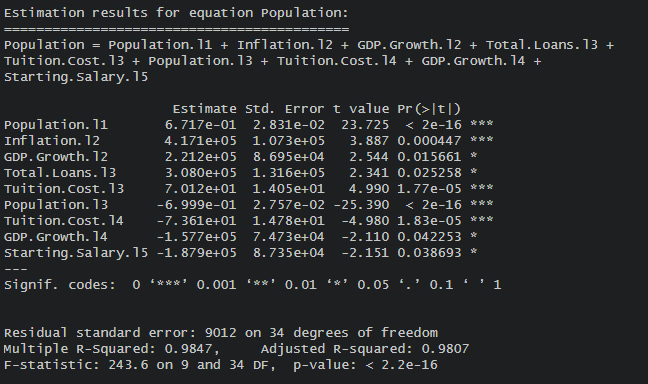
****

****

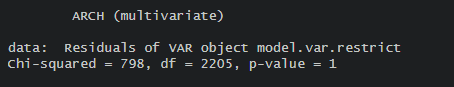
****

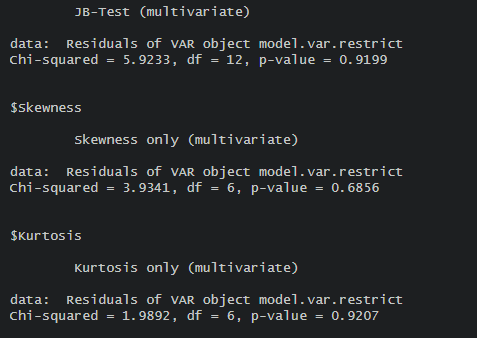
****

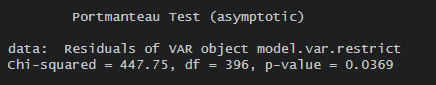
****

****

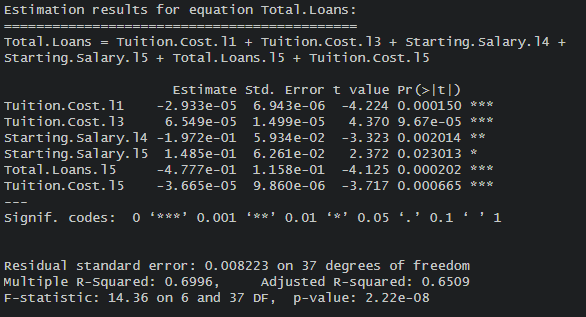
**Hypothesis Tests for Chosen Restricted VAR model**

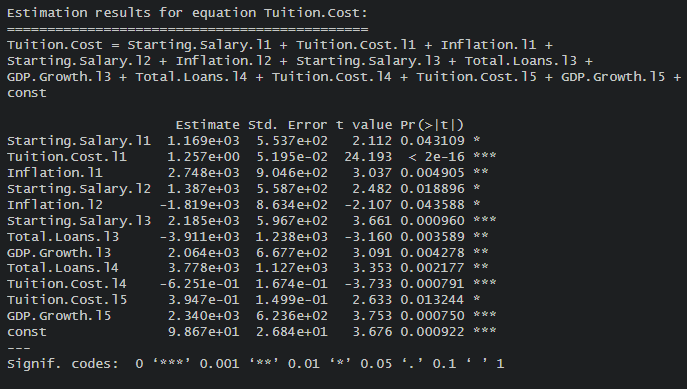


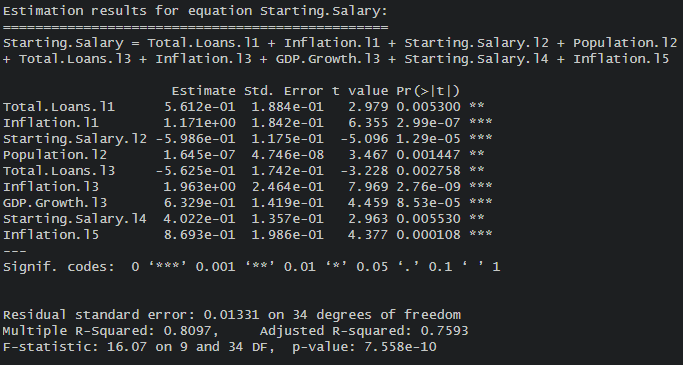


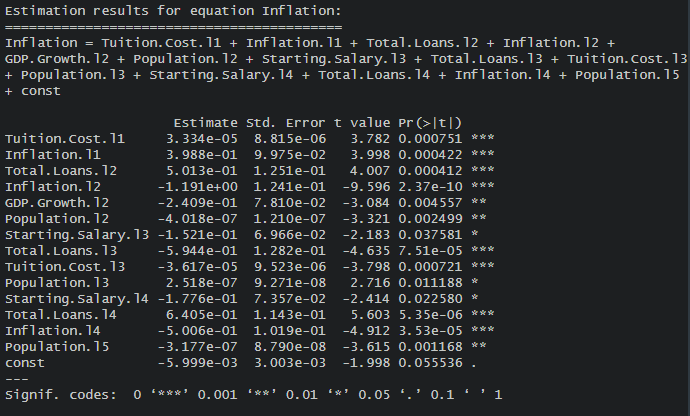


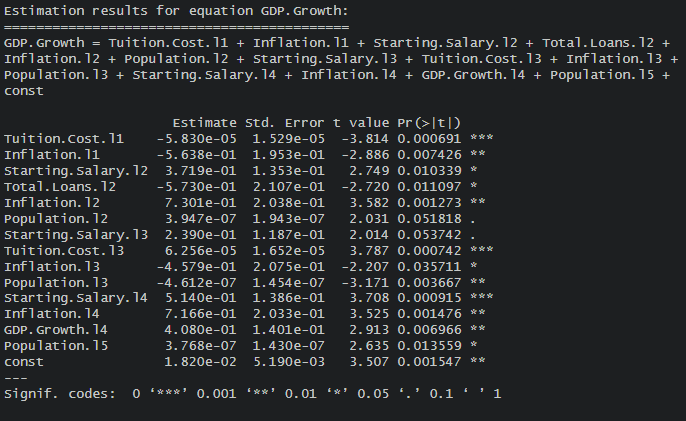
**Parameters for Chosen Restricted VAR model**

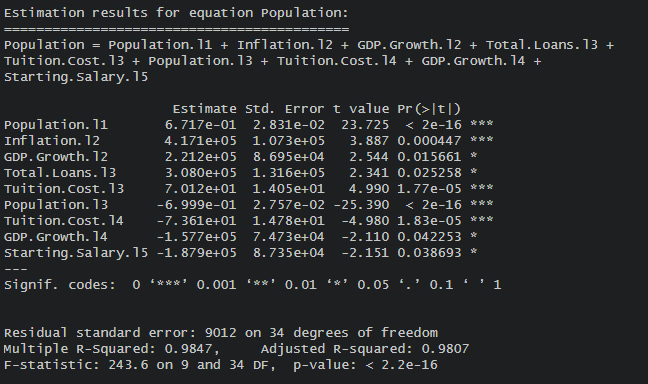
****

****

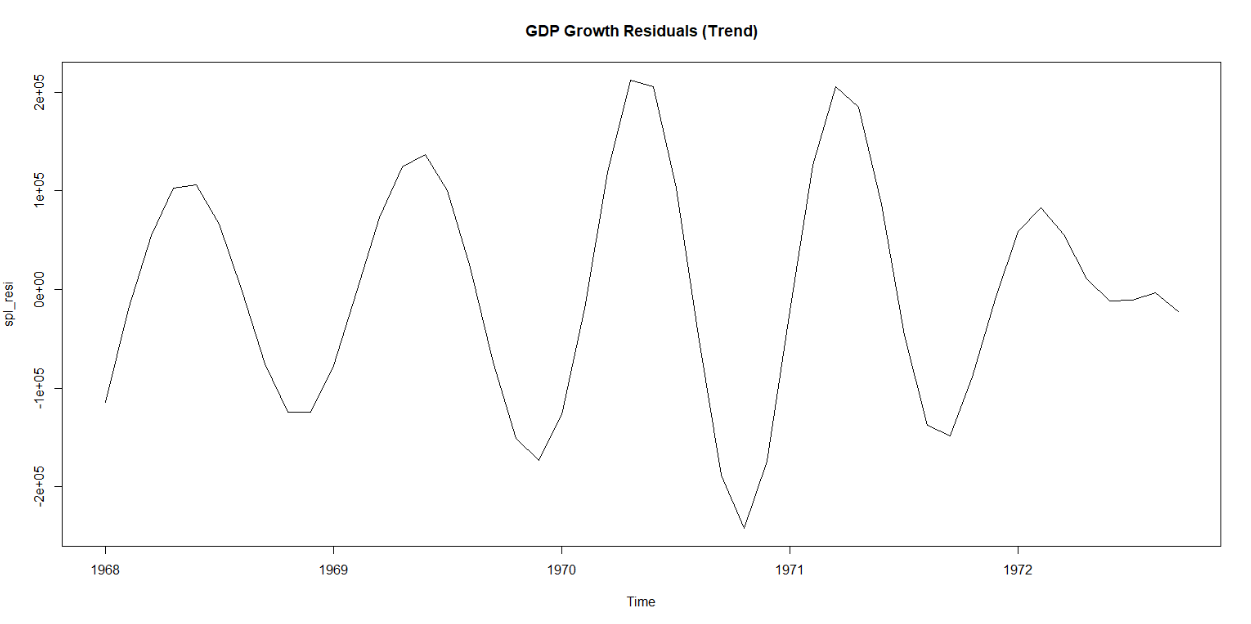
****

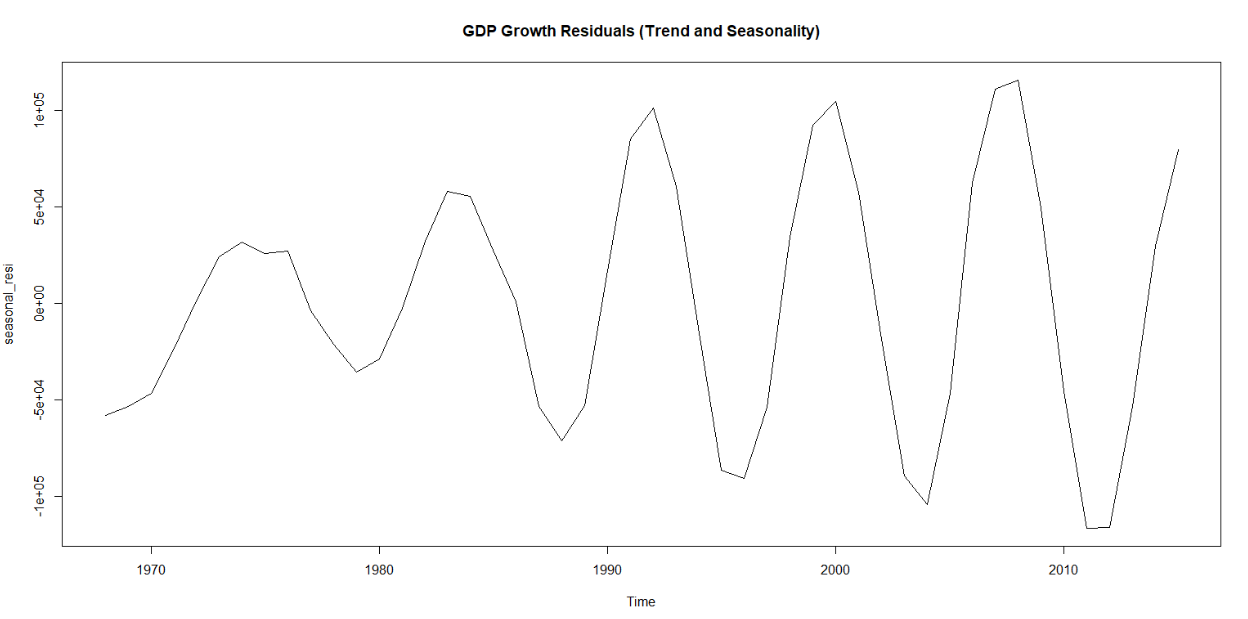
****

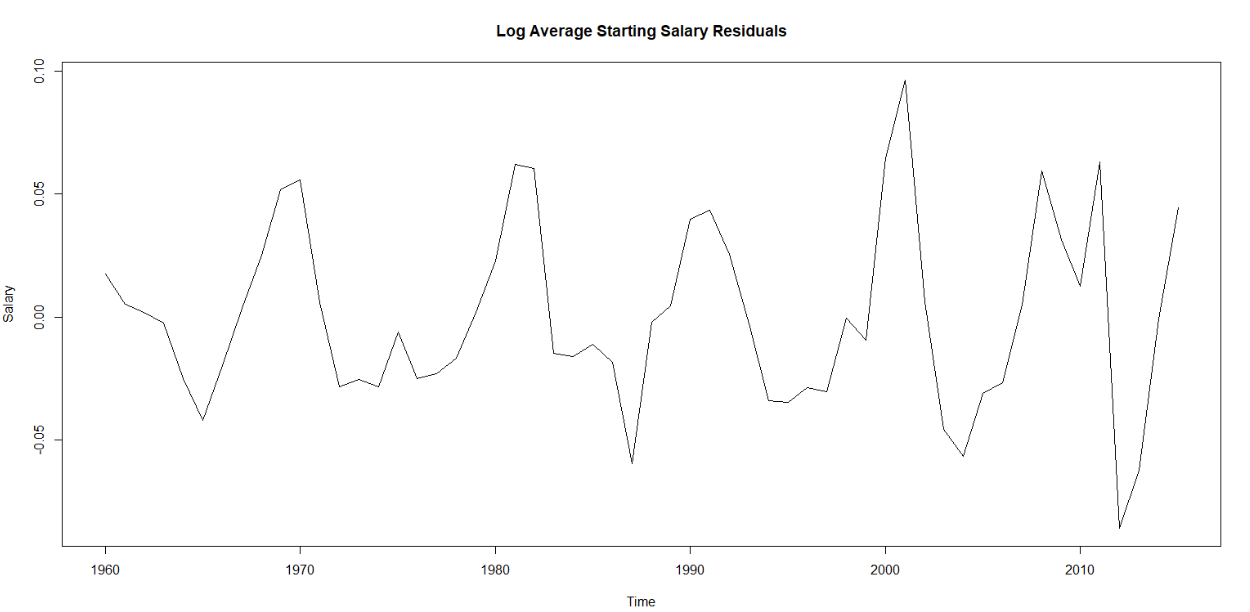
****

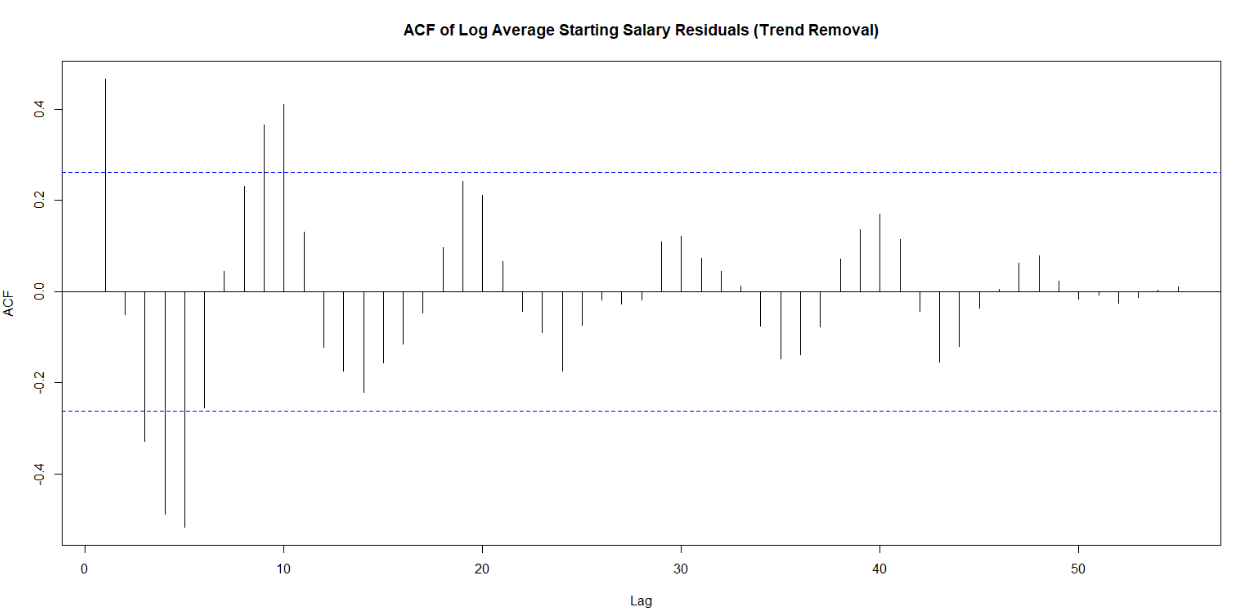
****

**Seasonal Trends**









1. This study is edited using Grammarly, an official typing assistant provided by Georgia Tech. [↑](#footnote-ref-0)
2. For full dataset on Kaggle’s Student Loan Data, please see: <https://www.kaggle.com/datasets/omarsobhy14/student-loans> [↑](#footnote-ref-1)
3. For full dataset on NY Federal Reserve Bank’s Student Loan Data, please see: <https://www.newyorkfed.org/microeconomics/databank.html> [↑](#footnote-ref-2)
4. For full dataset on NACE’s Dataset on College Costs, please see: <https://nces.ed.gov/programs/digest/d22/tables/dt22_330.10.asp> [↑](#footnote-ref-3)
5. For fulldata set on NACE’s Average Starting Salaries, please see: [https://www.naceweb.org/job-market/compensation/salary-trends-through-salary-survey-a-historical-perspactive-on-starting-salaries-for-new-college-graduates/](https://www.naceweb.org/job-market/compensation/salary-trends-through-salary-survey-a-historical-perspective-on-starting-salaries-for-new-college-graduates/) [↑](#footnote-ref-4)
6. For full dataset on Inflation, please see: <https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=US> [↑](#footnote-ref-5)
7. For full dataset on GDP, please see: <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US> [↑](#footnote-ref-6)
8. For full dataset on Population, please see: <https://www.statista.com/statistics/1067138/population-united-states-historical/> [↑](#footnote-ref-7)
9. For full model output, please refer to Appendix’s Table 1, titled “Spline Regression for Student Loan Data”. [↑](#footnote-ref-8)