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

**Avery Girsky, James Peavy & Moe Kyaw Thu (Group 7)**

**December 8th 2023  
Project Report for Time Series Analysis[[1]](#footnote-1)**

***Abstract (Summary)***

This study delves into the evolving dynamics of student loan, starting salaries, and college costs in the United States between the 1960s and the 2010s. Through a comprehensive analysis of various datasets, namely student loan, college costs, average starting salaries, inflation, GDP growth, and the population data, the study addresses key questions surrounding these dynamics. The outcomes show that there is a trend only in the student loan and college cost data and both trend and seasonality exist in the average starting salaries. This study further conducts fit between ARMA and ARMA-GARCH modelling and finds out that the latter is a stronger predictor. Lastly, in conducting multivariate analysis modelling with these three datasets with the inflation, GDP growth, and the population data, this study finds out that there is a lagged relation between variables. The outcomes of this study hope to further contribute to the existing literature on higher education in the US.

***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?

Based on these questions, this study anticipates basic time series trends prior to conducting analysis to revealing stationarity. Moreover, this study also hypothesizes that a VARX model encompassing inflation, GDP, and population data as exogenous variables will outperform any other VAR modeling. Lastly, this study also expects statistically significant relationships among lag relationships between loan and tuition data.

***II. Dataset***

As this study analyzes a variety of dataset, the table below a full summary:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Sce** | **Period** | **Key Variables** | **Purpose** |
| Student Loan | Kaggle[[2]](#footnote-2) & NY Federal Reserve Bank[[3]](#footnote-3) | 1958 to 2020 | Loan amounts (billions),  Borrower demographics, | Analysis of the evolving dynamics of student loans |
| College Costs | NACE[[4]](#footnote-4) | 1963 to 2020 | Institution type: public, private, private/non-profit, private/for-profit  Inflation adjustment: constant (adjusted to 2021-2022 prices), current (at-time cost)  Type of cost: all, tuition and fees, room, board  Program length: all, f -year degree, two-year degree. | Analysis of the evolving dynamics of college costs |
| Average Starting Salaries | NACE[[5]](#footnote-5) | 1960 to 2020 | Average starting salaries across all majors | Analysis of the economic outcomes of college major choices |
| Inflation Data | World Bank[[6]](#footnote-6) | 1960 to 2022 | US inflation rate | Analysis of the broader economic conditions that may influence both student loan and college costs |
| GDP Data | World Bank[[7]](#footnote-7) | 1960 to 2022 | US GDP | Analysis of the US GDP and its relation with other factors in the study |
| Population Data | Statista[[8]](#footnote-8) | 1960 to 2022[[9]](#footnote-9) | US population data | Analysis as an exogenous variable |

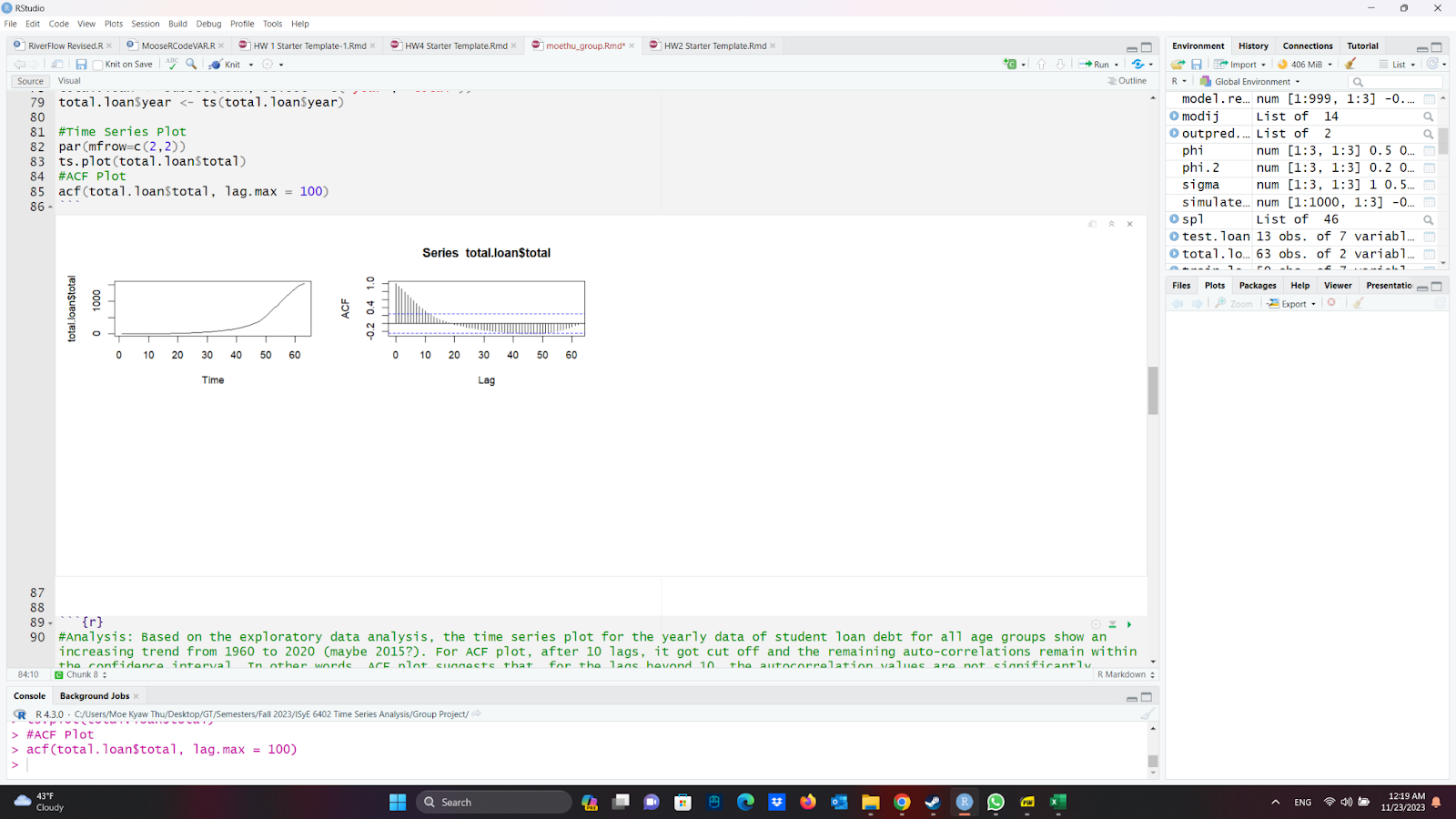
***III. Methodology***

In general, this study begins with data cleaning. For example, in 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. This is followed by trend fitting for each of the dataset, followed by the incorporation of seasonality if deemed necessary (for full process, please refer to the HTML file provided). Subsequently, the analysis proceeds to the application of Autoregressive Moving Average (ARMA) modeling. To enhance the ARMA models’ capability to capture volatility, this study also extends to Generalized Autoregressive Conditional Heteroskedasticity (GARCH) by conducting ARMA-GARCH process. The efficacy of the models is also evaluated via a variety of hypothesis testing method, providing a comprehensive comparison to ascertain their strengths and weaknesses in capturing the strengths and patters in the data. In addition to ARMA, ARMA-GARCH, and hypothesis testing, as a final step, this study also incorporates Vector Autoregressive (VAR) modeling in order to capture the multivariate relationships between datasets.

***IV. Results***

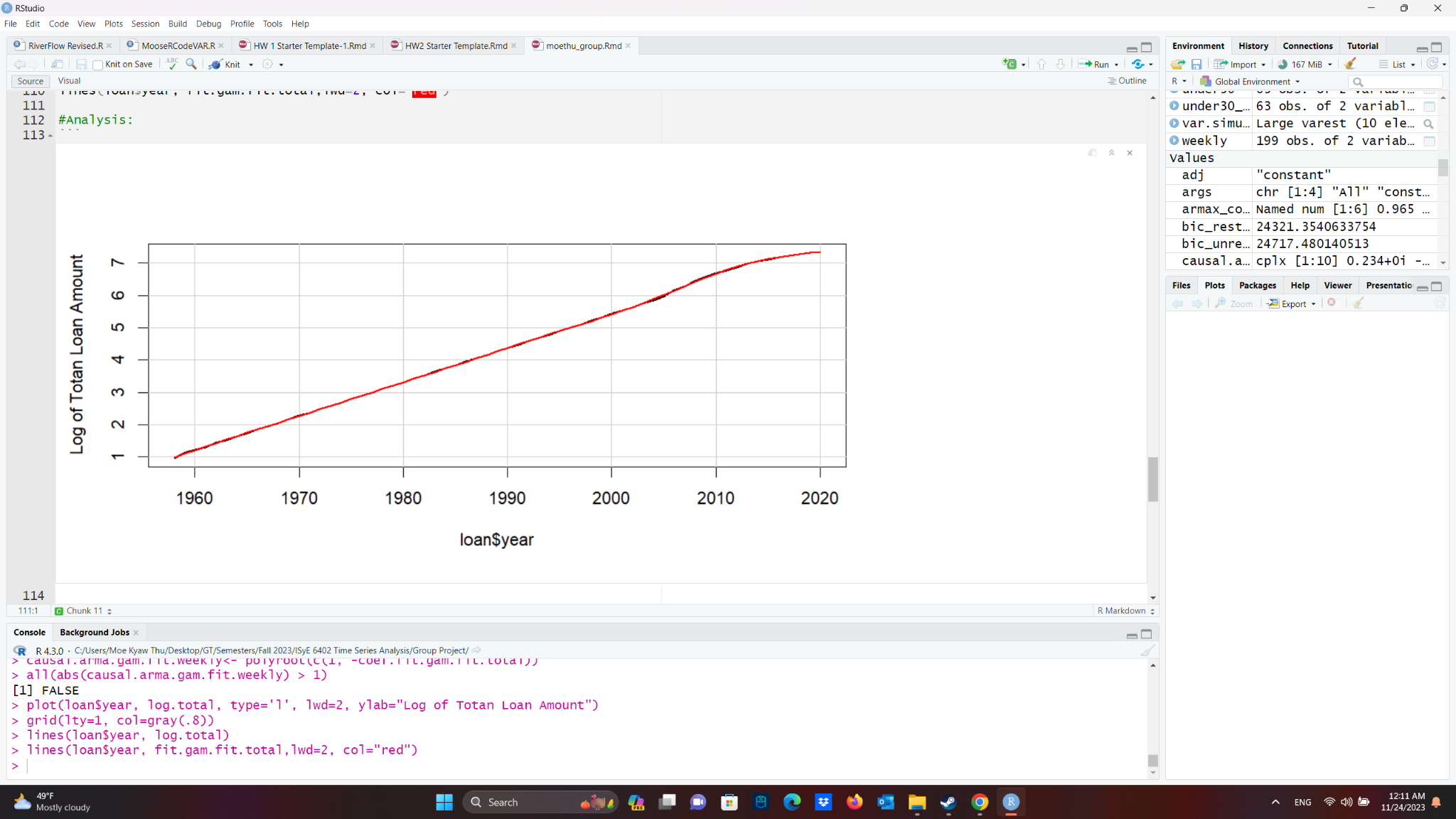
1. **Tracing the Student Loans from the 1960s to the 2020s**

Overall, the exploratory data analysis reveals an upward trend in the student loan data across all ages. With autocorrelation plot (ACF), a notable observation shows that beyond 20 lags, it exhibits a truncation, with subsequent autocorrelation values consistently falling within the confidence interval.



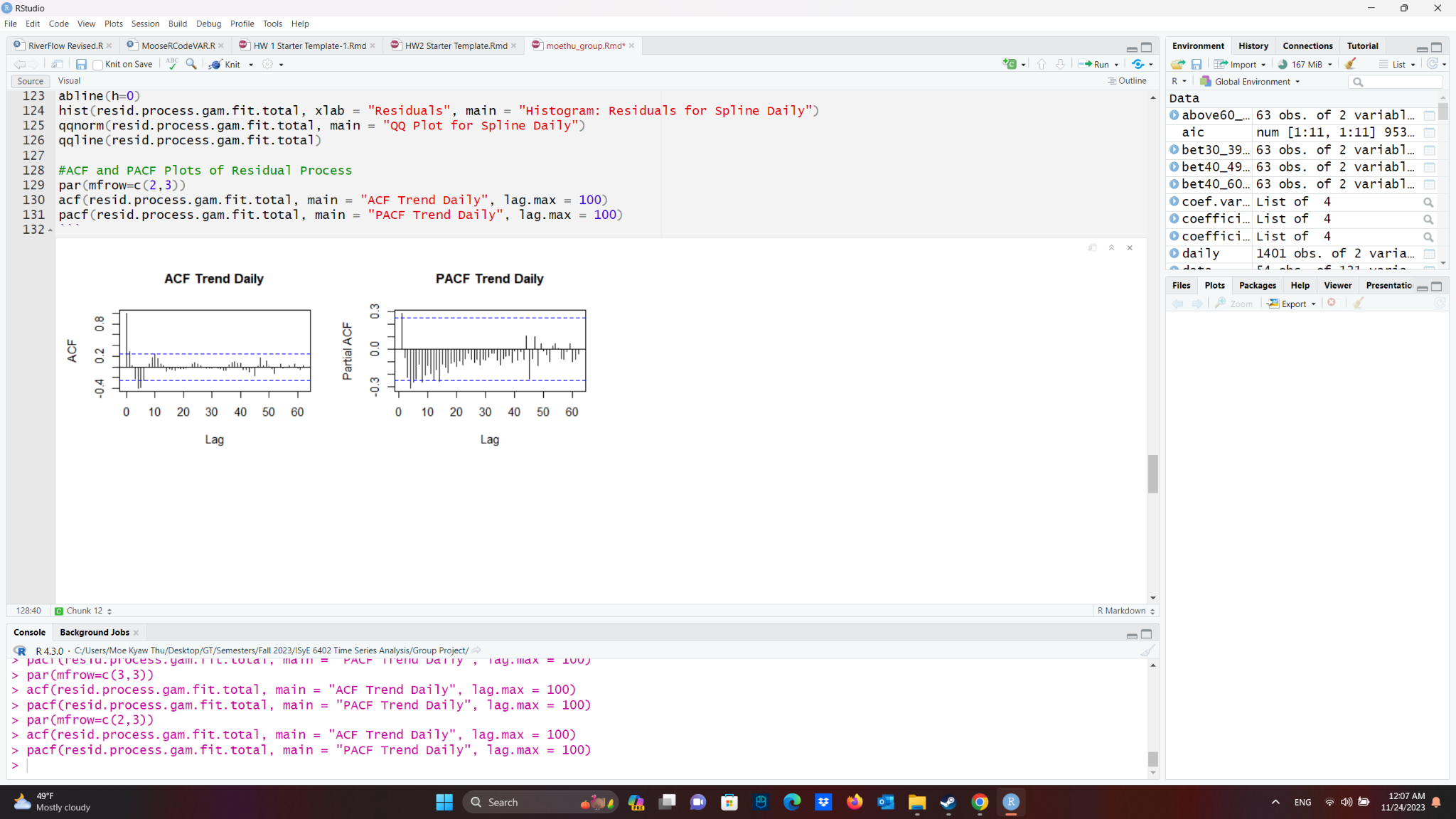
**Fig.1.** Trend and Seasonality Analysis of Total Student Loans

In order to further deepen the analysis of the student loan trend data, this study applies spline regression. The outcomes show that 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 (for full table, please see Appendix: Spline Regression for Student Loan Data). The output is further demonstrated in the spline regression output with the original time series and it shows an almost perfect fit. (see figure 2).



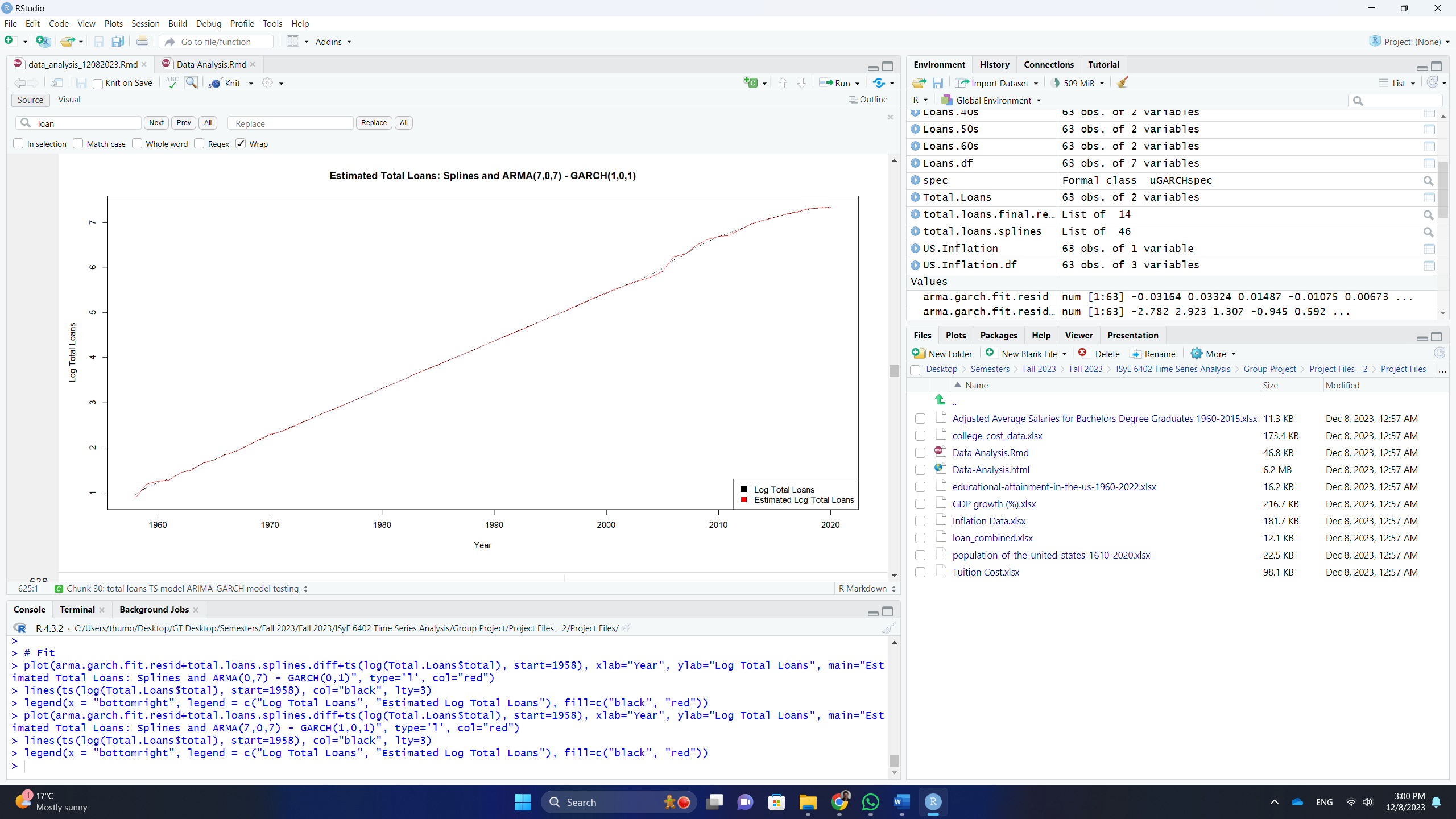
**Fig.2.** Spline Regression Analysis

This study also analyzes the residual process after the trend removal. This study fit an ARMA model to the post-trend-removal residuals (see figure 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

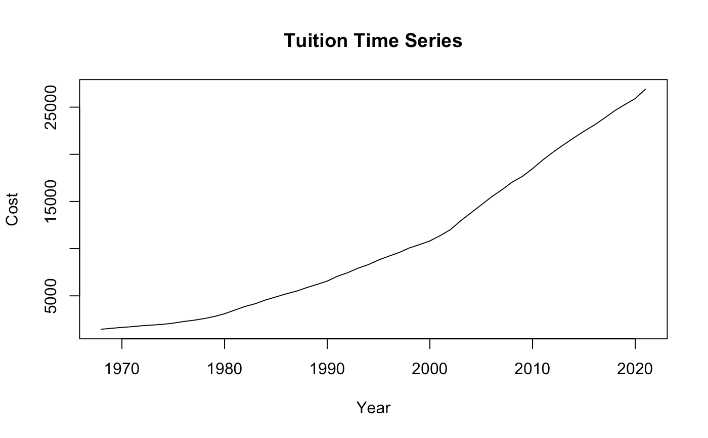
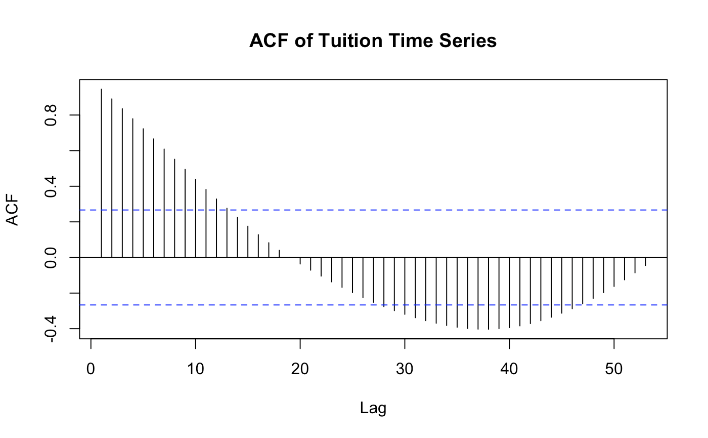
As a last step, given these findings, this study analyzes using ARMA-GARCH model of order (7,0,7) and (1,0,1) respectively (see figure 4). This produced similar results to the above ARMA model, but with the added benefit of showing signs of normality in the residuals. Due to these results, this study demonstrates that the ARMA-GARCH model is slightly better than the ARMA model for prediction purposes.



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

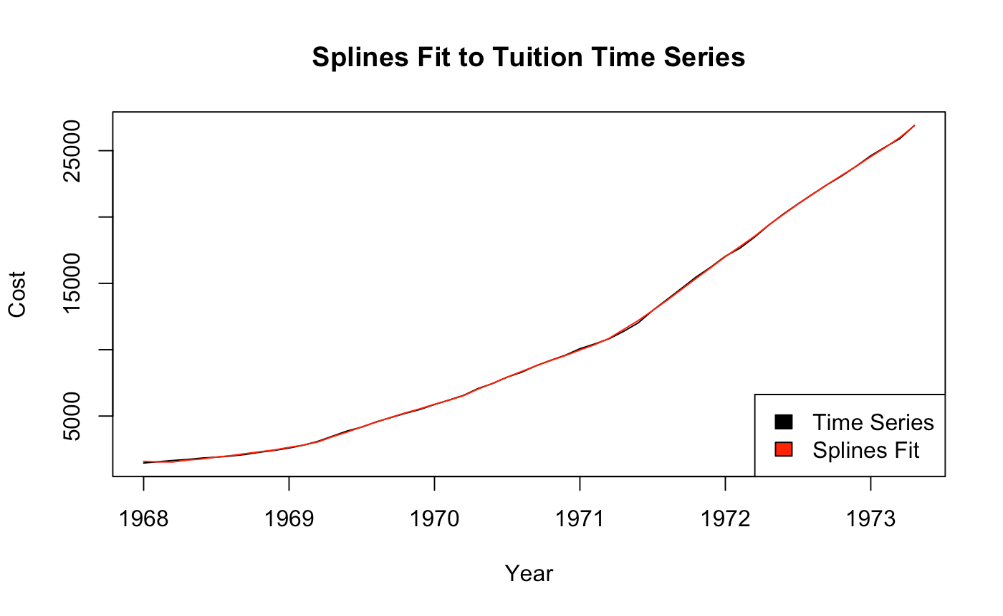
1. **Tracing College Costs from the 1960s to the 2010s**

The non-adjusted college cost, comprising tuition and fees, room, and board, is a strictly increasing upward time series from 1968 to 2021, as seen in Figure 5. The data shows an upward trend, and the ACF plot of the time series confirms it.

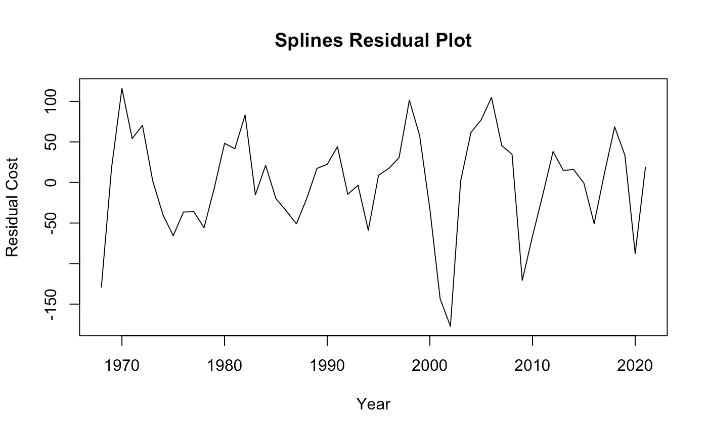
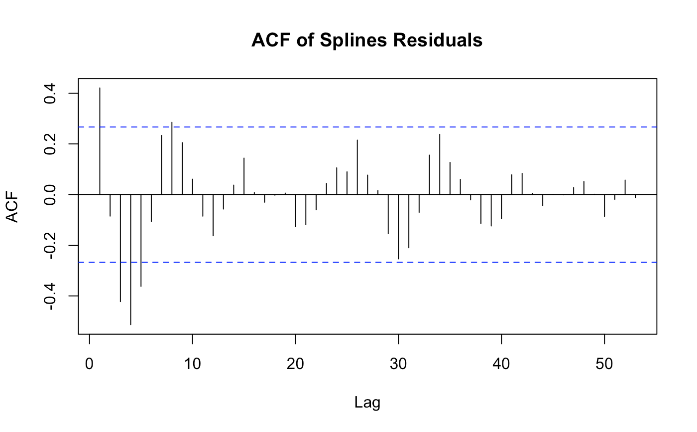
 

**Fig. 5**. Time Series and ACF Plot for College Costs

In pursuit of discovering an appropriate trend estimation method, there is a significant 10-year seasonality in the time series. This seasonality, also observed in the starting salary time series, could be related to economic fluctuations, further motivating the inclusion of exogenous factors in the multivariate analysis, which will be explored later. The splines model outcome for this seasonality, detailed in figure 6, proved to be a good fit where the smooth terms are significant, and every 3rd and 8th year proved to be significantly different from the intercept.

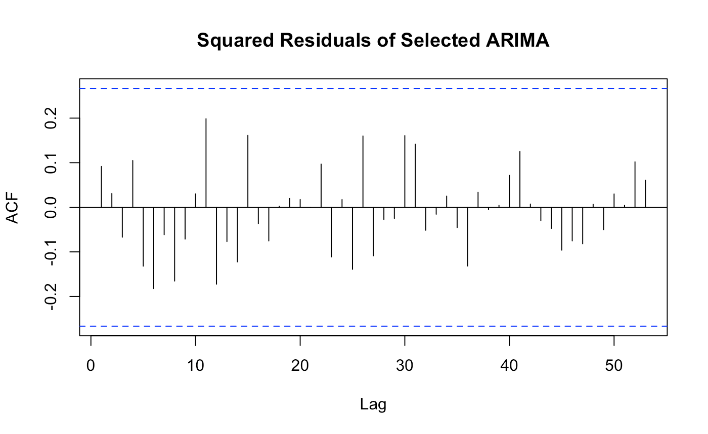
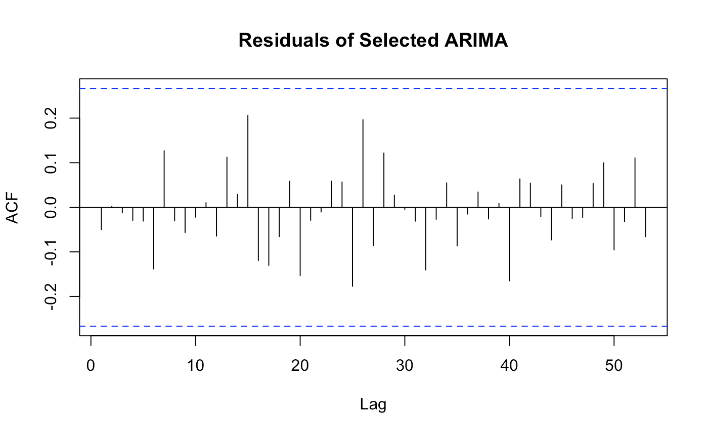


**Fig.6.** Fitted Splines Regression vs. College Costs

The residual plot and ACF plot of the spline regression and ANOVA seasonality residuals in figure 7 show plausible stationarity in the residuals.

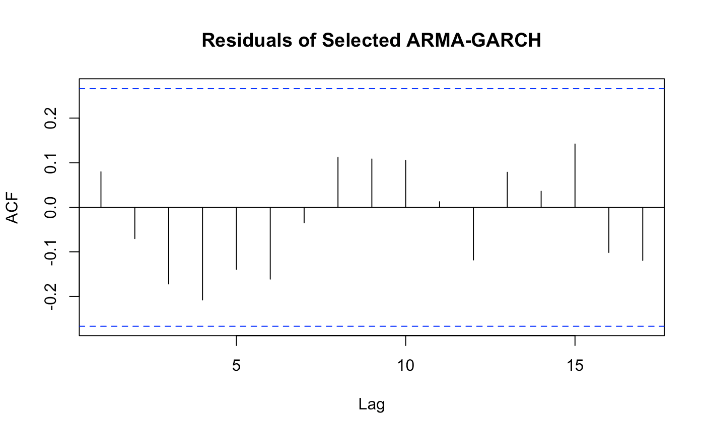
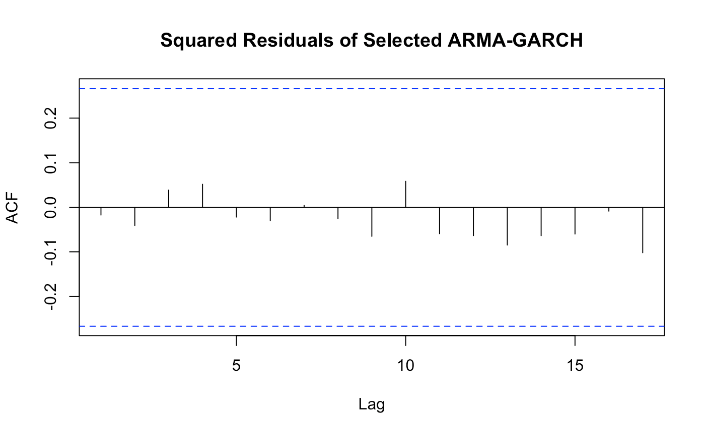
**Fig.7.** Residual and ACF Plot for Splines Trend and ANOVA Seasonality Estimation

With these residuals, order selection deemed an ARIMA (3,0,2) to be an appropriate fit. However, the AR model of 3 coefficient is not significant. The model residual normality assumption is satisfied based on a Shapiro-Wilk test of the ARIMA residuals. 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.



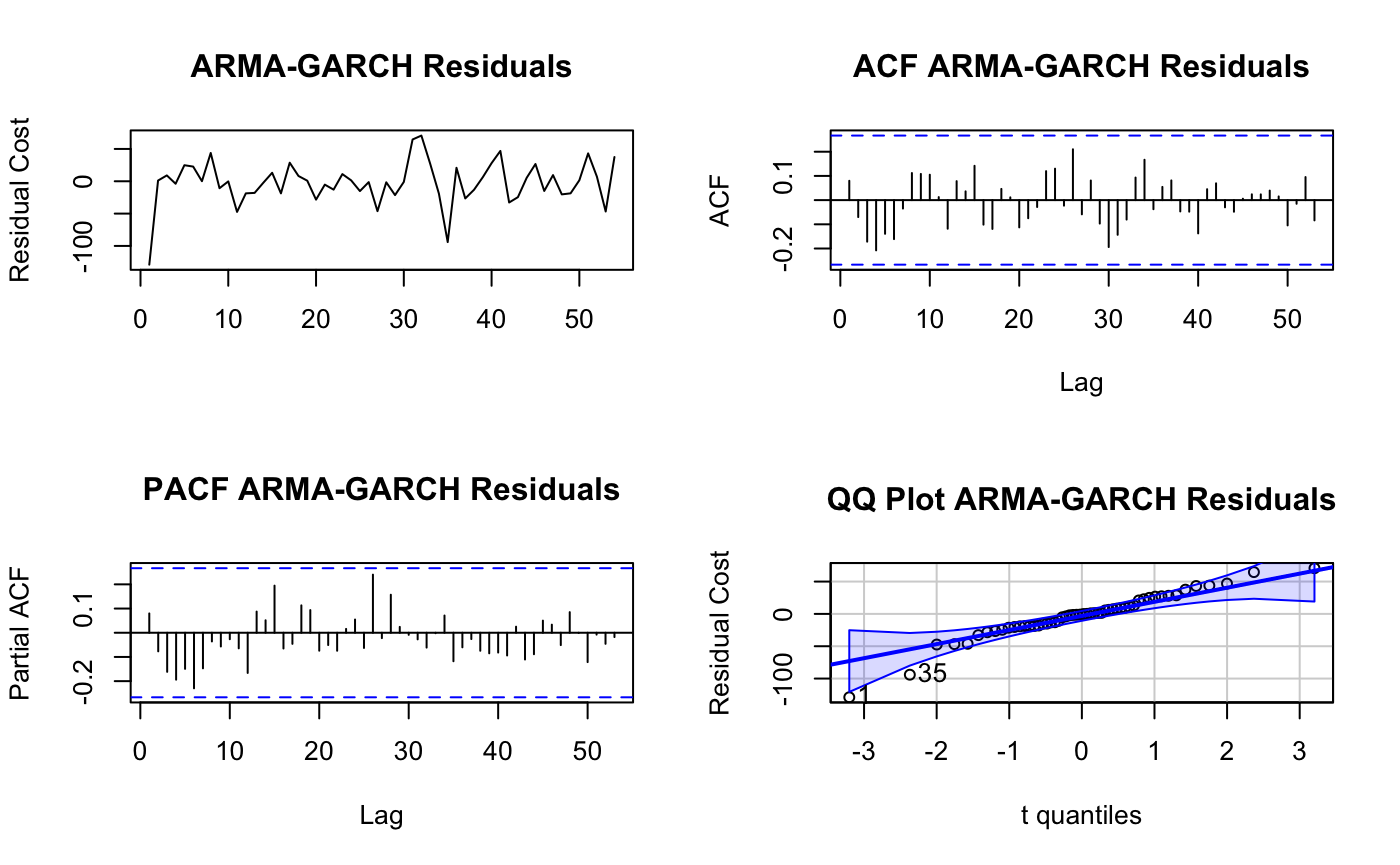
**Fig. 8.** ACF Plot, Residuals and Squared Residuals Plots for of ARIMA (3,0,2)

The potential heteroskedasticity from the Box-Ljung test on the squared residuals leads towards conducting an ARMA-GARCH fit modelling. For the order selection of our ARMA-GARCH model, this study initially fixed ARMA (2,3) and found that the best GARCH order based on AIC was 0 and 1. After 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, this study fits the spline residuals with an ARMA (0,7)- GARCH (0,1) model and all coefficients were deemed significant.

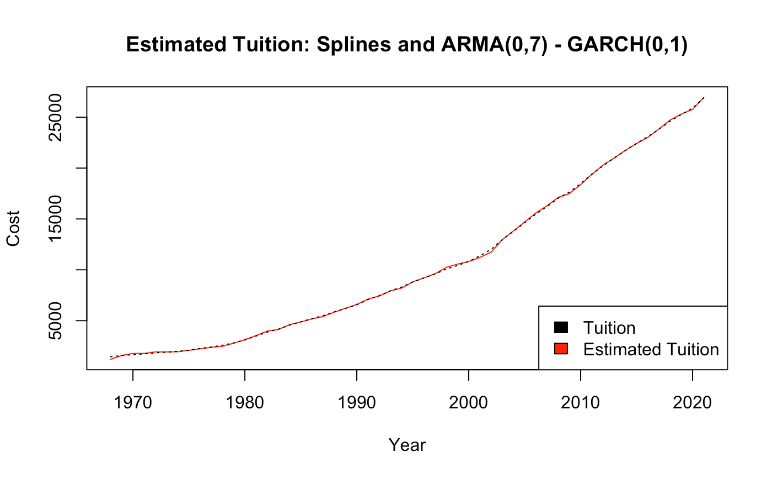
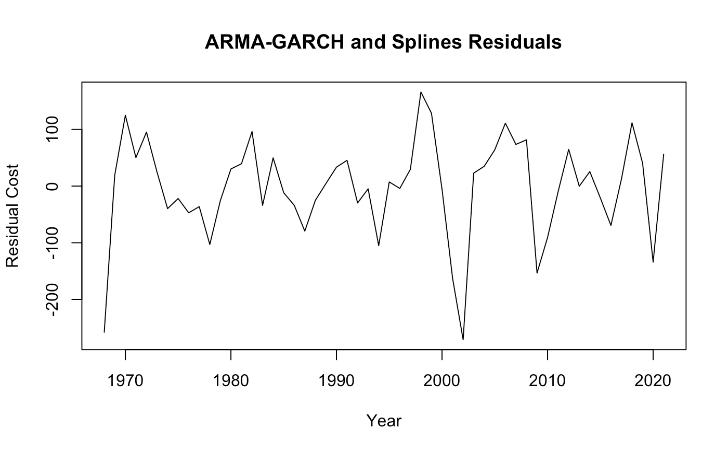
**Fig.9**. ACF Plot of ARMA (0,7)-GARCH (0,1) Residuals and Squared Residuals

The residual plot and PACF from Figure 9 reaffirm 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 abnormal but may conform to a t-distribution 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 removed heteroskedasticity in the data.



**Fig. 10.** Residual, ACF, PACF, and QQ Plots for ARMA (0,7)-GARCH (0,1)

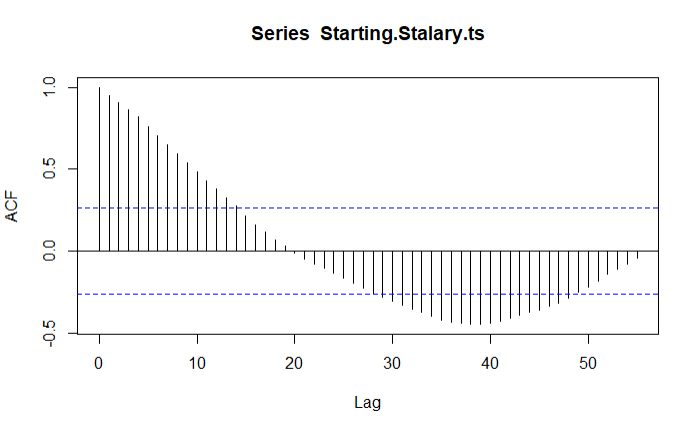
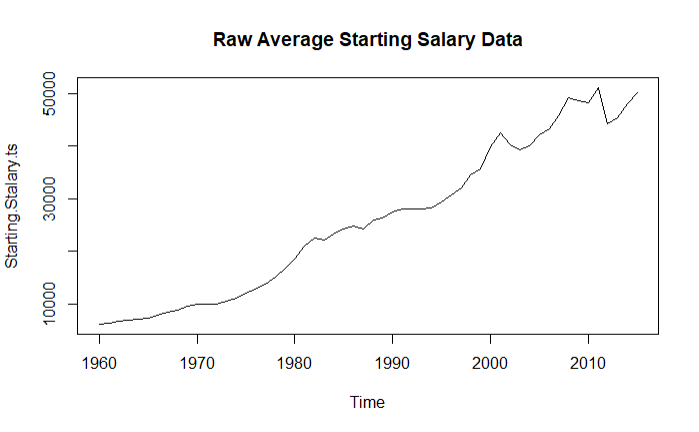
Finally, in figure 10, this study observes how the fitted splines plus seasonality residuals combine with the ARMA-GARCH residuals to estimate the college cost time series rather well. The employed methods are able to estimate residuals to generally within $150 of actual college prices.



**Fig. 11.** Combined ARMA-GARCH and Splines Residuals (left), Fit to Cost Time Series (right)

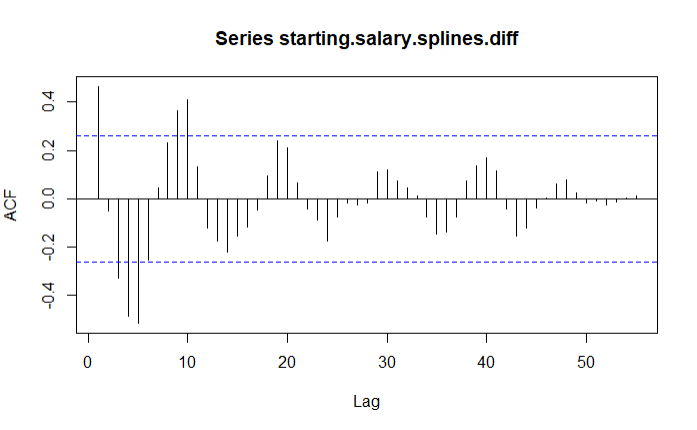
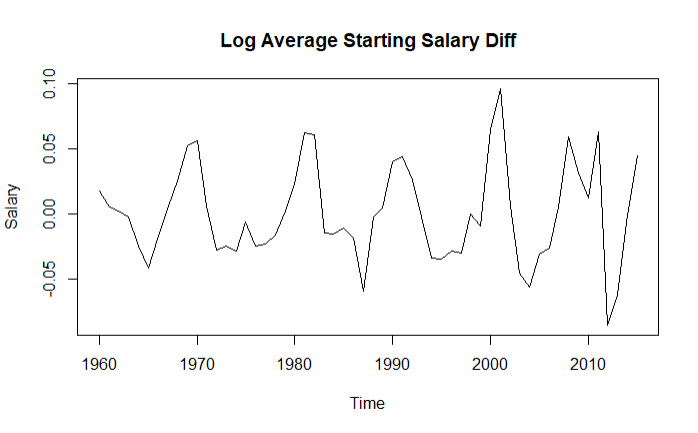
1. **Tracing Average Starting Salary from the 1960s-2010s**

The analysis of the average starting salary data showed a clear upward trend in the data and a clear trend from its ACF plot (figure 12).



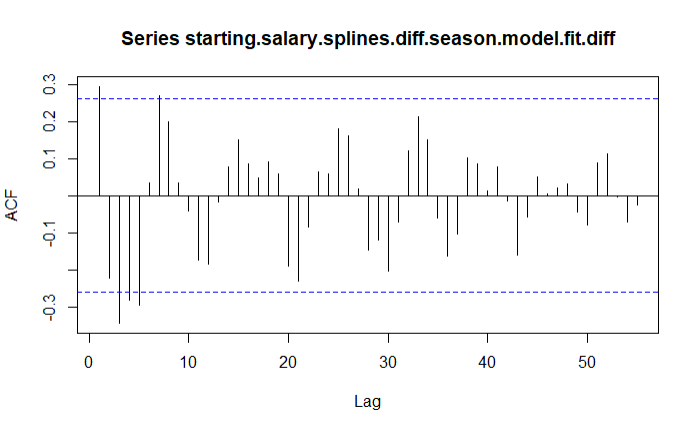
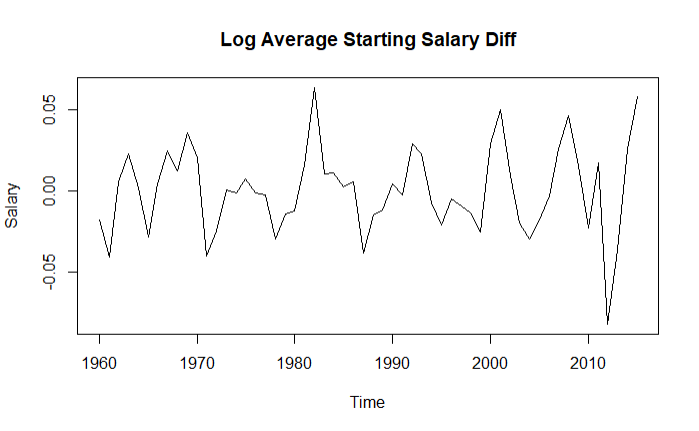
**Fig.12.** Trend Analysis of Average Starting Salary

The outcomes are further fitted using spline regression by taking account into the logarithm of the data to see clearer stationarity in the residuals. These splines fit has found to be very statistically significant. When further examining the figure 13, it became evident that there is likely some degree of seasonality present, as inferred from the residual shapes and the patterns observed in the ACF plot.



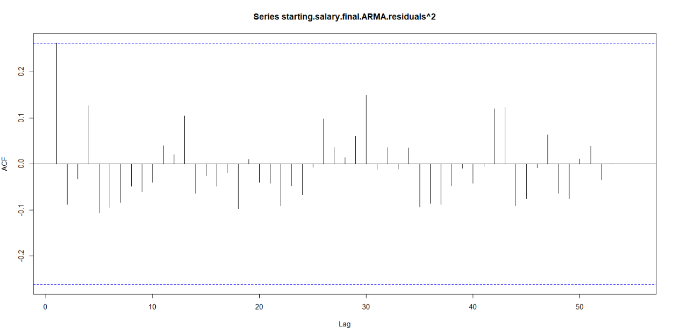
**Fig.13.** Trend Residual Analysis for Starting Salaries

Given the normal residuals plot, this study suspects that there is some sort of decade-based seasonality. This led towards fit a seasonal trend based on a 10-year time period. The resulting seasonal trend showed statistically significant coefficients with the added coefficients based on a p-value of 0.05. The resulting residuals also showed a higher degree of stationarity (figure 14), specifically when thinking about p-values of greater than 0.05.



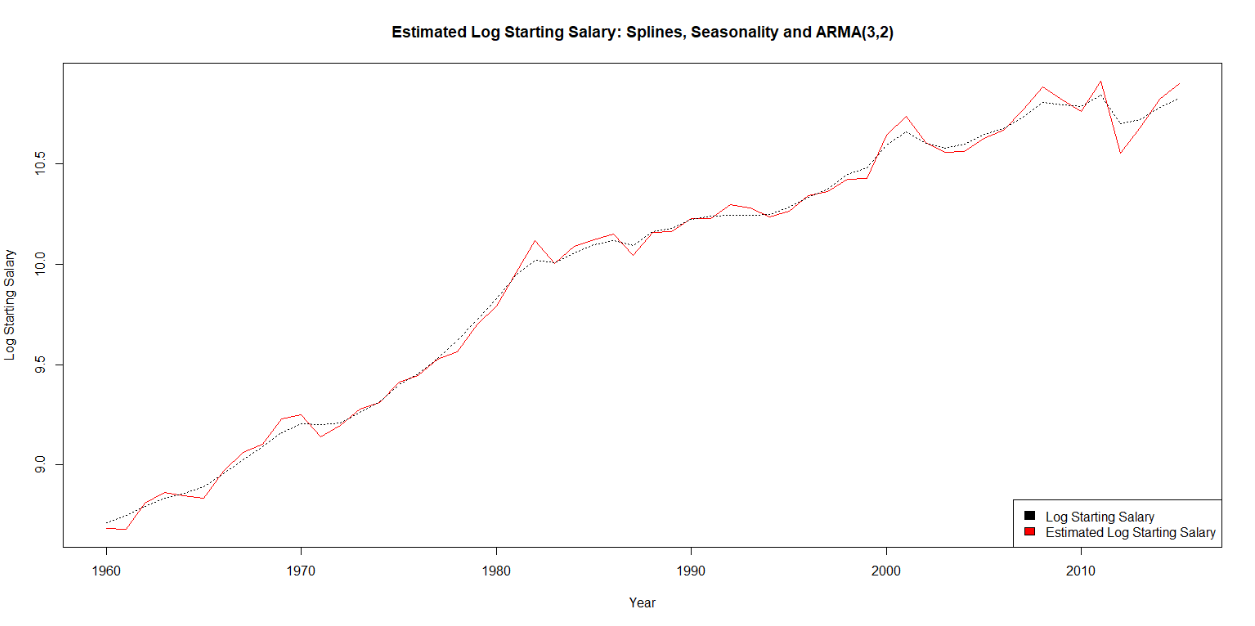
**Fig.14.** Trend and Seasonality Residual Analysis for Starting Salaries

With these residuals, order selection deemed an ARIMA (3,0,2) to be an appropriate fit. This study tested 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 (see figure 15).

**Fig.15.** ARMA Residuals (Left) and Squared Residuals (Right)

Furthermore, this study also tests for an ARMA-GARCH model of order (5,0,5) and (1,0,0) respectively. This model did not provide 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 ARMA model (Fig. 16.), the model is a fairly strong fit.



**Fig.16.** Splines, Seasonality, and ARMA Fit to Average Starting Salary Time Series

1. **Multivariate Modelling Relationships**

During the process of applying a multivariate modelling process, all datasets are fit with spline regression coupled with seasonal adjustments where necessary to achieve stationarity. Subsequently, by obtaining residuals, this study then tests each of residual outcomes with an ADF test, where the results show stationarity (for full information, please see Appendix: Var Model Analysis). Additionally, all the VAR model outcomes are also unstable, which will be further explored in the Discussion and Conclusion section. Despite this, this study utilizes the VAR (5) model as it performed the best against hypothesis tests, showing the absence of heteroskedasticity, normality, and serial correlation against a 0.01 p-value (for full information, please see Appendix: Hypothesis Tests for Chosen Restricted VAR model).

This study also conducts multivariate lagged relationships between datasets. The outcomes are further summarized in the table below[[10]](#footnote-10):

|  |  |
| --- | --- |
| Factor | Correlated Relationships |
| Total Loan | * Lag 1, 3, and 5 of the tuition cost data * Lag 4 and 5 of the starting salary data * Lag 5 of itself (the total loan data) |
| College Costs | * Constant value from VAR specification * Lag 1, 2, and 3 of the starting salary data * Lag 1 and 2 of the inflation data * Lag 3 and 4 of the total loan data * Lag 3 and 5 of the GDP Growth data * Lag 1, 4, and 5 of itself |
| Average Starting Salaries | * Lag 1 and 3 of the total loans data * Lag 1, 3, and 5 of the inflation data * Lag 2 of the population data * Lag 3 of the GDP growth data * Lag 2 and 4 of itself |
| Inflation | * Lag 1 and 3 of the tuition data * Lag 2, 3, and 4 of the total loans data * Lag 2 of the GDP growth data * Lag 2, 3, and 5 of the population data * Lag 3 and 4 of the starting salary data * Constant value from VAR specification * Lag 1, 2, and 4 of itself |
| GDP growth data | * Lag 1 and 3 of the tuition data * Lag 1, 2, 3, and 4 of the inflation data * Lag 2, 3, and 4 of the starting salary data * Lag 2 of the total loans data * Lag 2, 3, and 5 of the population data * Constant value from VAR specification * Lag 4 of itself |
| Population | * Lag 2 of the inflation data * Lag 2 and 4 of the GDP growth data * Lag 3 of the total loans data * Lag 3 and 4 of the tuition cost data * Lag 5 of the starting salary data * Lag 1 and 3 of itself |

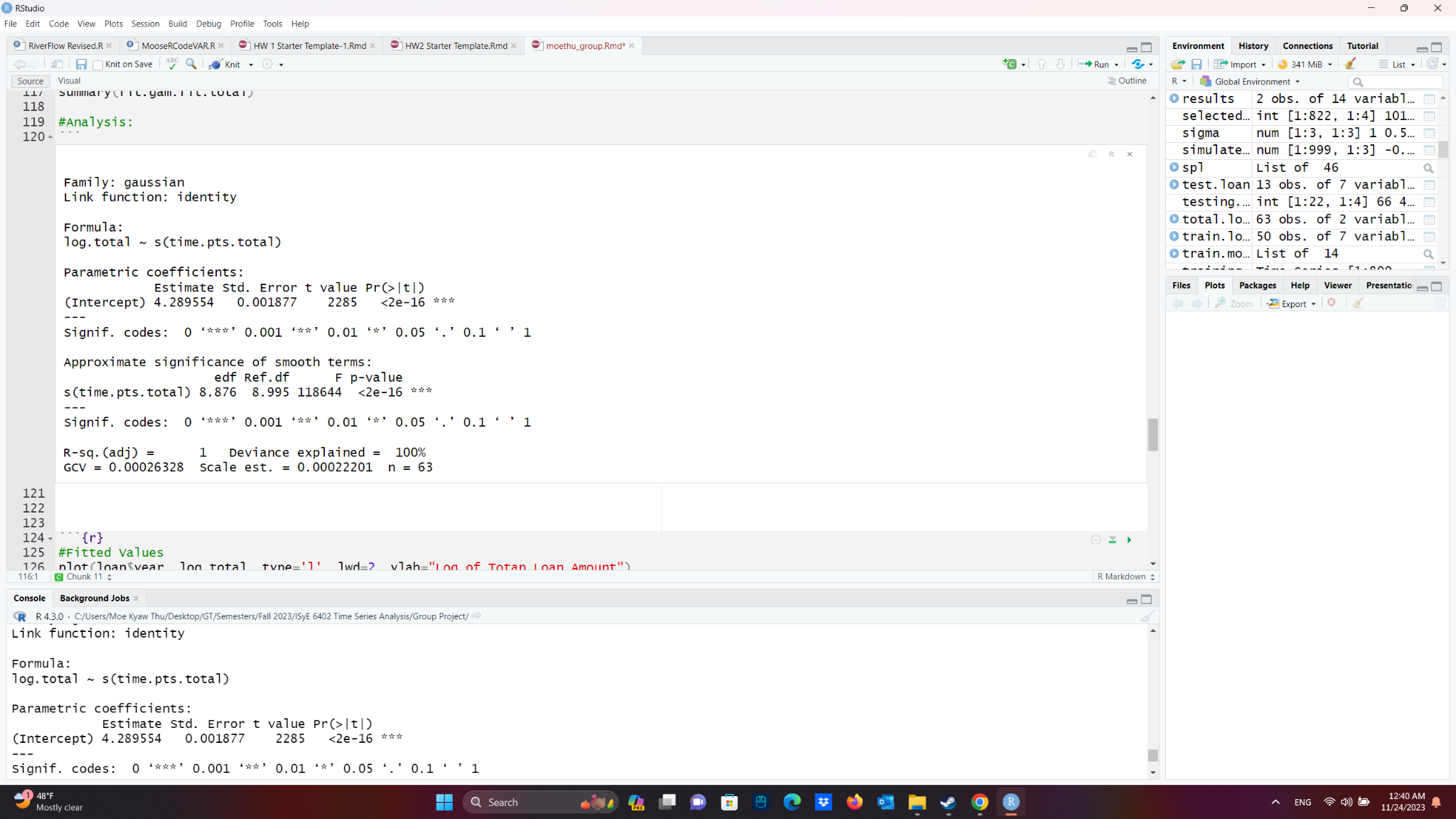
***V. Discussion and Conclusion***

By exploring these datasets, this study compares prior understanding of the data with new understanding. Initially, this study suspected all of the datasets to be simply trend-stationary. This is proven incorrect with the findings that the tuition, starting salary, and population data contains a 10-period seasonality. In addition, the population data exhibits seasonality that proved challenging to eliminate – a seasonal trend that appears to have decreased over time. For the VAR model, this study hypothesizes the best model to be a VARX model with the GDP growth, inflation, and population data as exogenous variables. Instead, the outcomes have shown that a general VAR model is a better fit based on the hypothesis testing procedures, but the outcomes are not deemed as stable. Despite this instability from the VAR models, some unexpected lag relationships also emerge. For instance, population data lags behind starting salary, inflation and GDP growth data, while not exhibiting a lagged relationship with total loan and college tuition costs.

However, there are also limitations to this study. First and foremost, as the datapoints contain yearly data from the 1960s to the 2010s, the total datapoint available stands at between 60 and 70 years. Due to this, as VAR and other modelling excel better with more time series datapoint, the outcomes of these models may not be optimal; for example, in the VAR model outcomes, this study has observed instabilities. While, the planned exogenous variables namely GDP growth, inflation and population data, also only represented yearly data, this study’s inability to obtain yearly enrollment college population data since the 1960s may have significantly contributed to the observed instability in the VAR model process. Furthermore, while this study has contributed to the time series analysis of each of the dataset, additional exploration into the relationships could be achieved via more advanced machine learning techniques such as neural networks and linear regressions. Lastly, based on the findings from VAR model, future studies could explore the causal factors by utilizing the Granger Causality method to understand the origins of these lagged relationships.

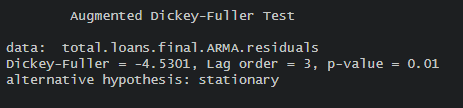
***VI. Appendix***

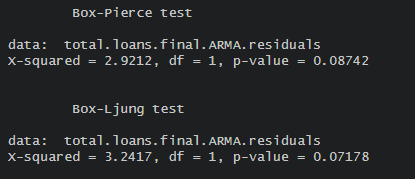
**Appendix: Student Loan Trend and Seasonal Analysis**

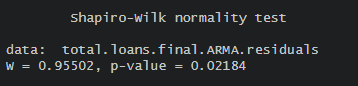


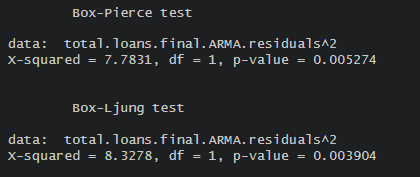
**Table. 1.** Spline Regression Output for Student Loan Data

**Appendix: Total Loans ARMA Hypothesis Testing**

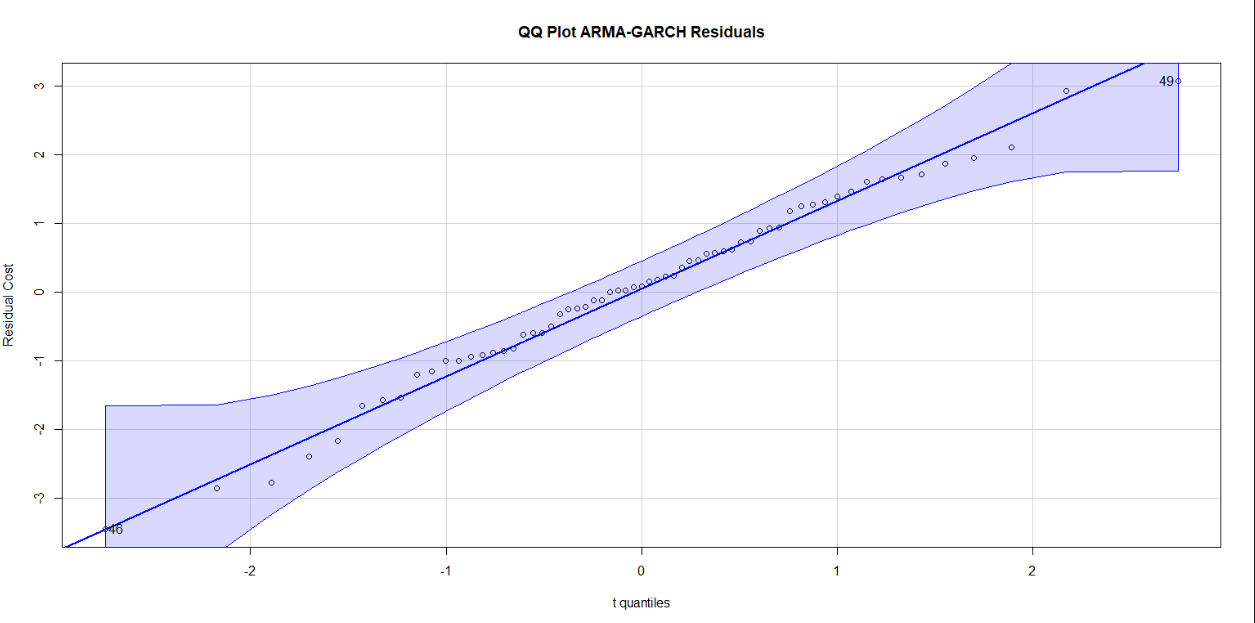




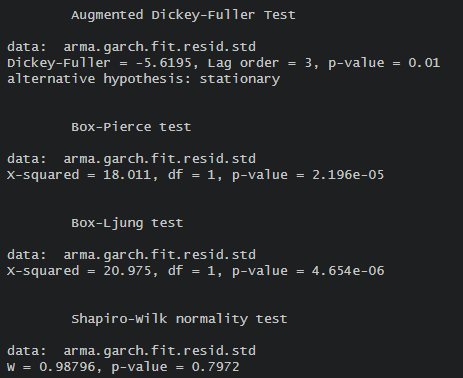




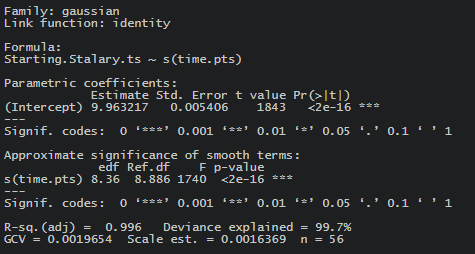
**Appendix: ARMA-GARCH of Total Loans data**



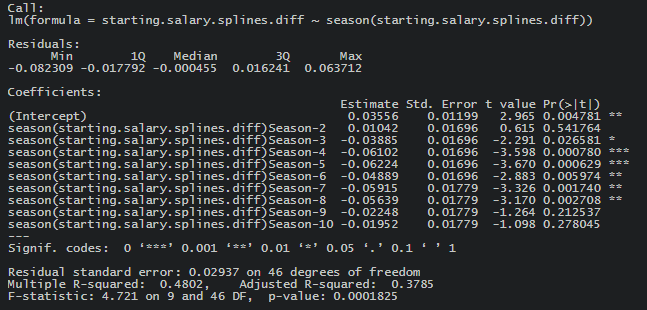
**Appendix: ARMA-GARCH of Total Loans data Hypothesis Tests**



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

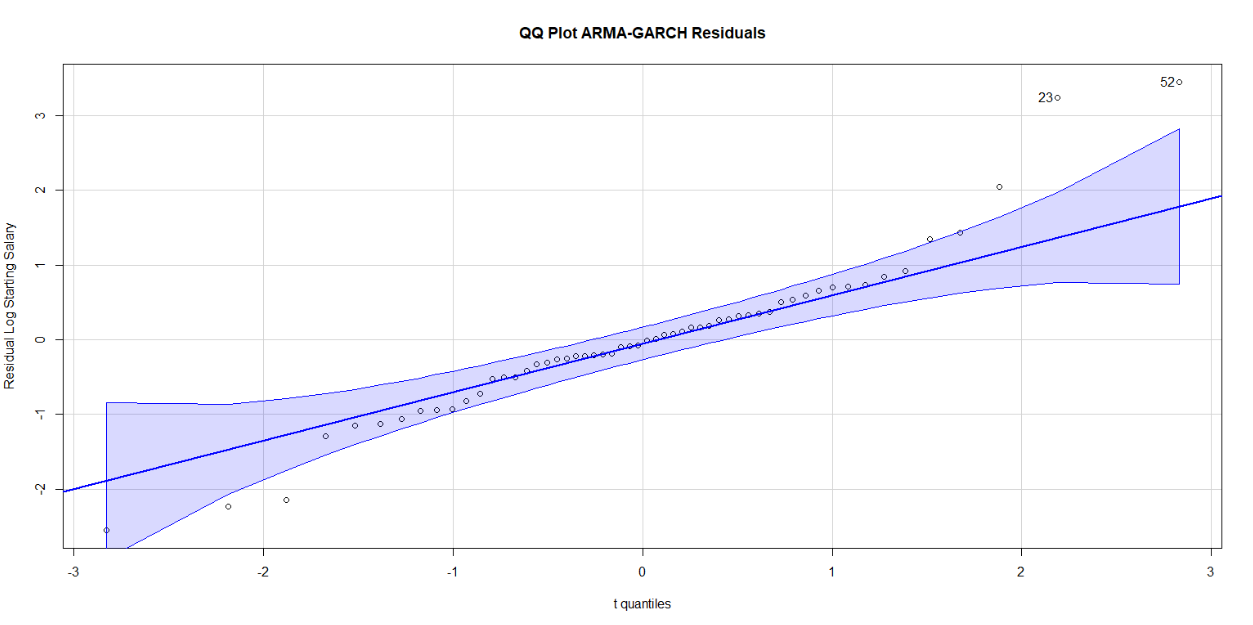


**Table. 2.** Spline Regression Output for Log Starting Salary Data

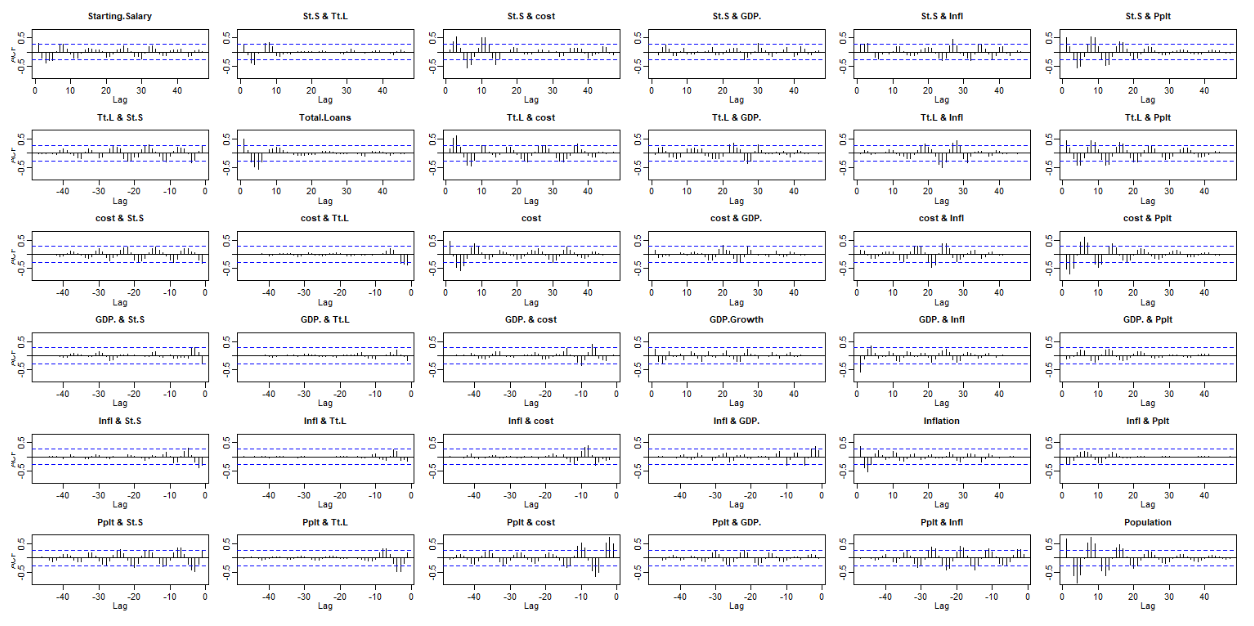


**Table. 3.** 10-Period Seasonal Regression Output for Differenced Log Starting Salary Data

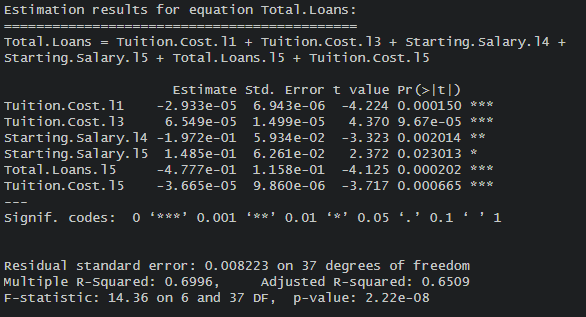
**Appendix: ARMA-GARCH of Starting Salary data**

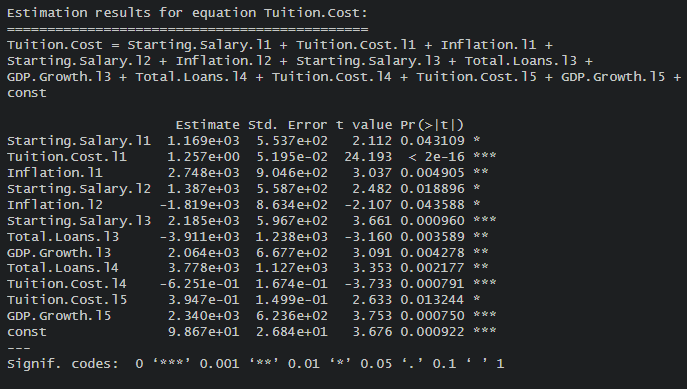


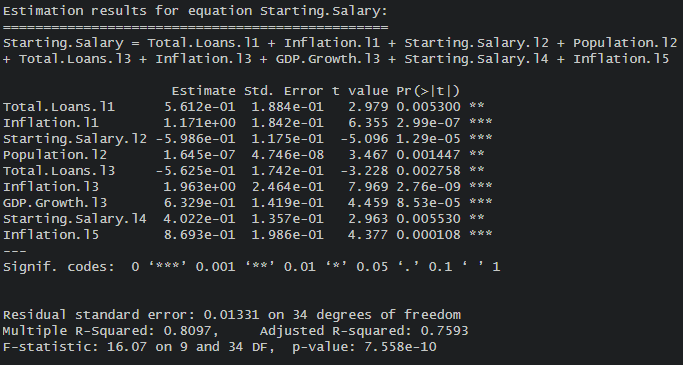
**Appendix: VAR Model Analysis**

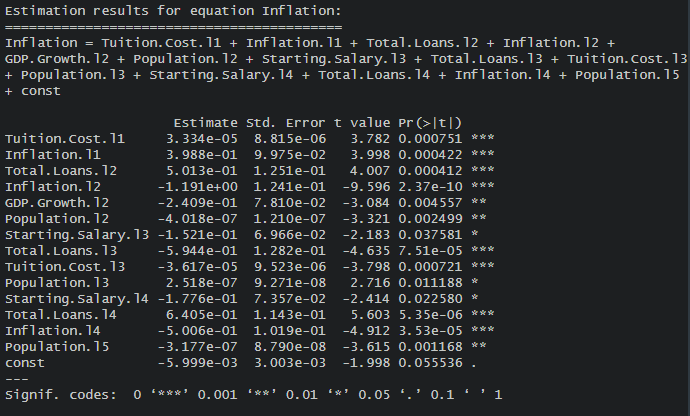


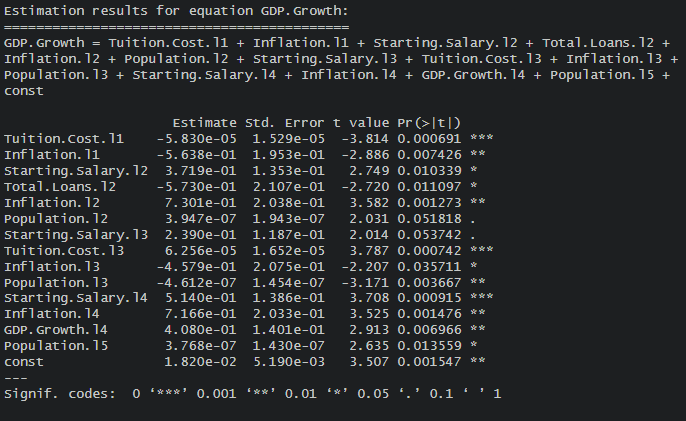
**Figure 2.** Residual process of all data (Before VAR)

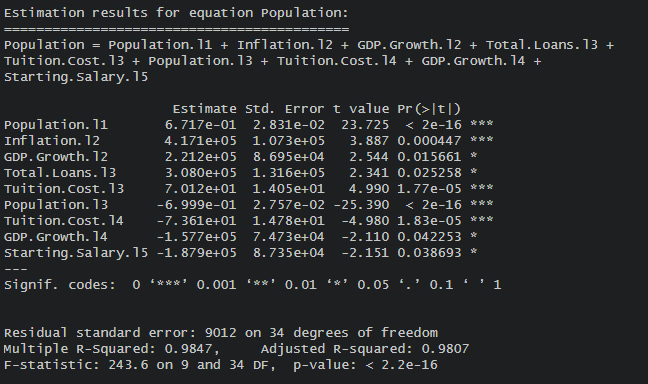
****

****

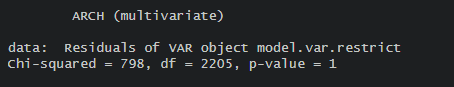
****

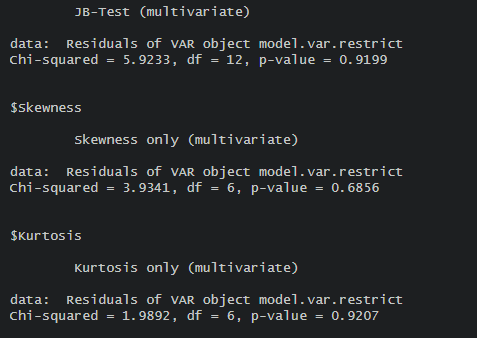
****

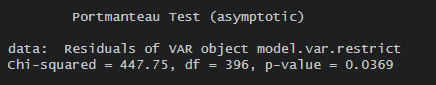
****

****

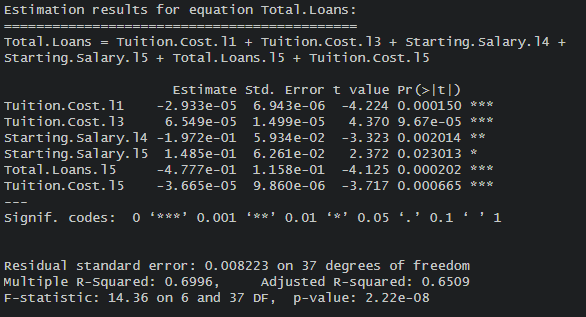
**Hypothesis Tests for Chosen Restricted VAR model**

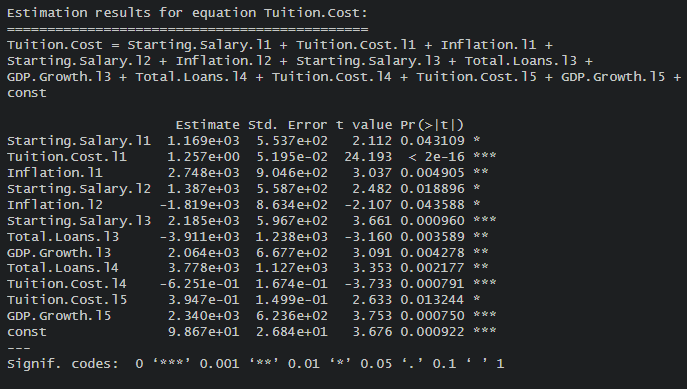


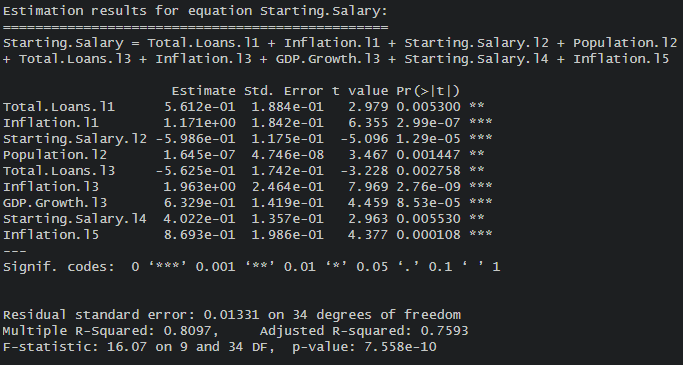


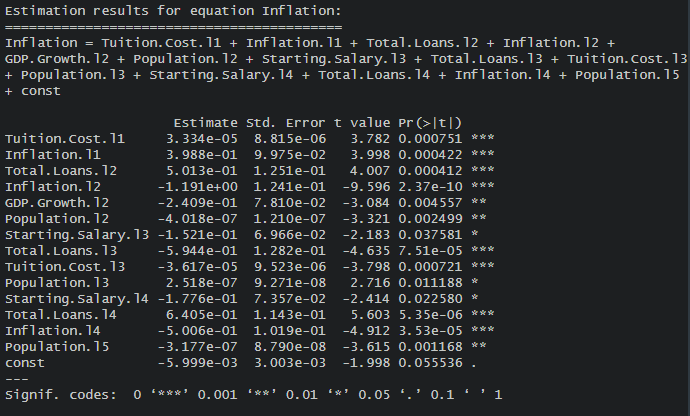


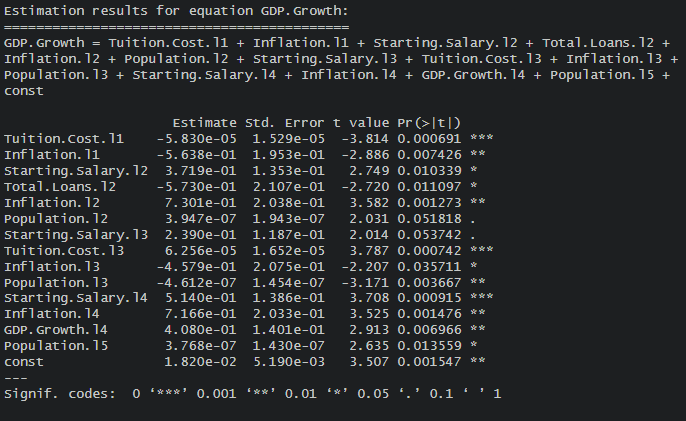
**Parameters for Chosen Restricted VAR model**

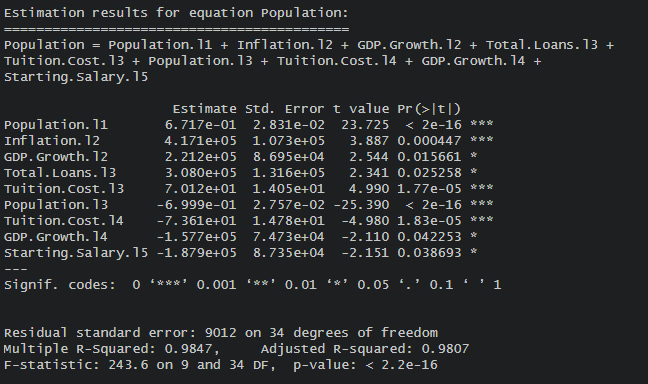
****

****

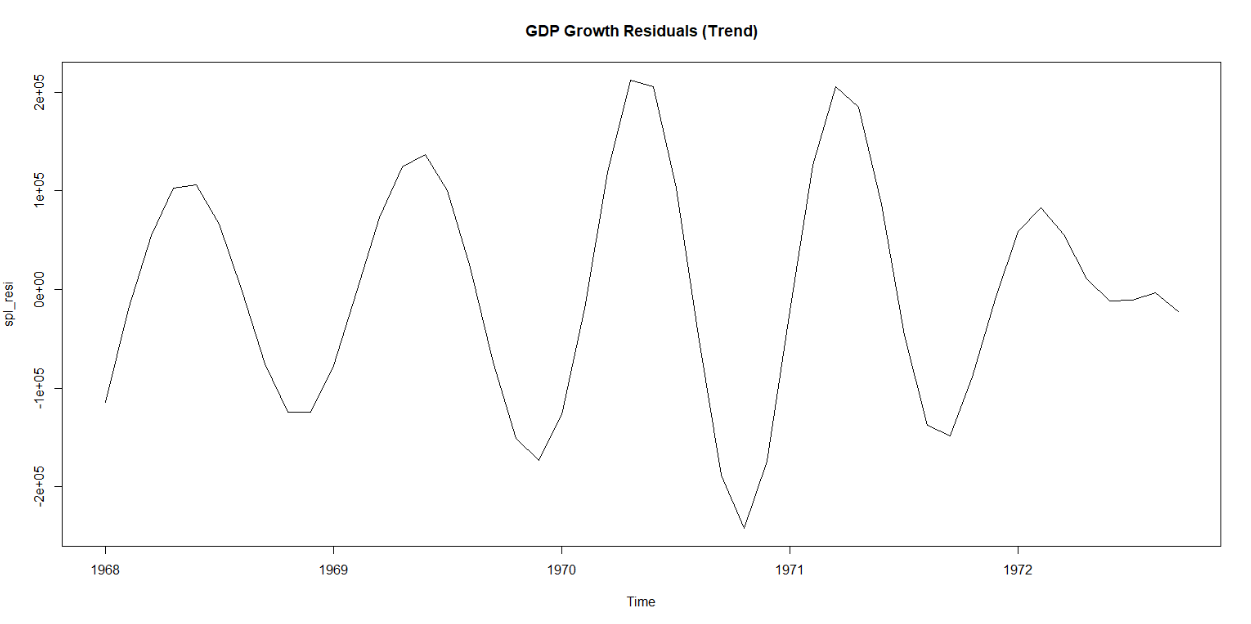
****

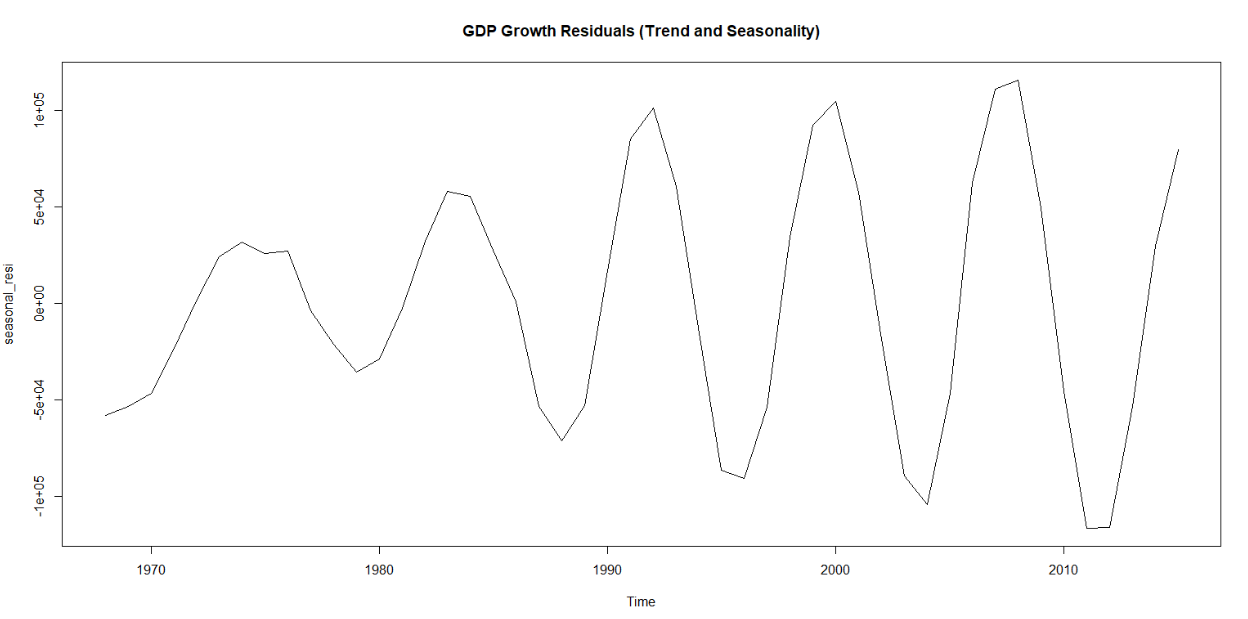
****

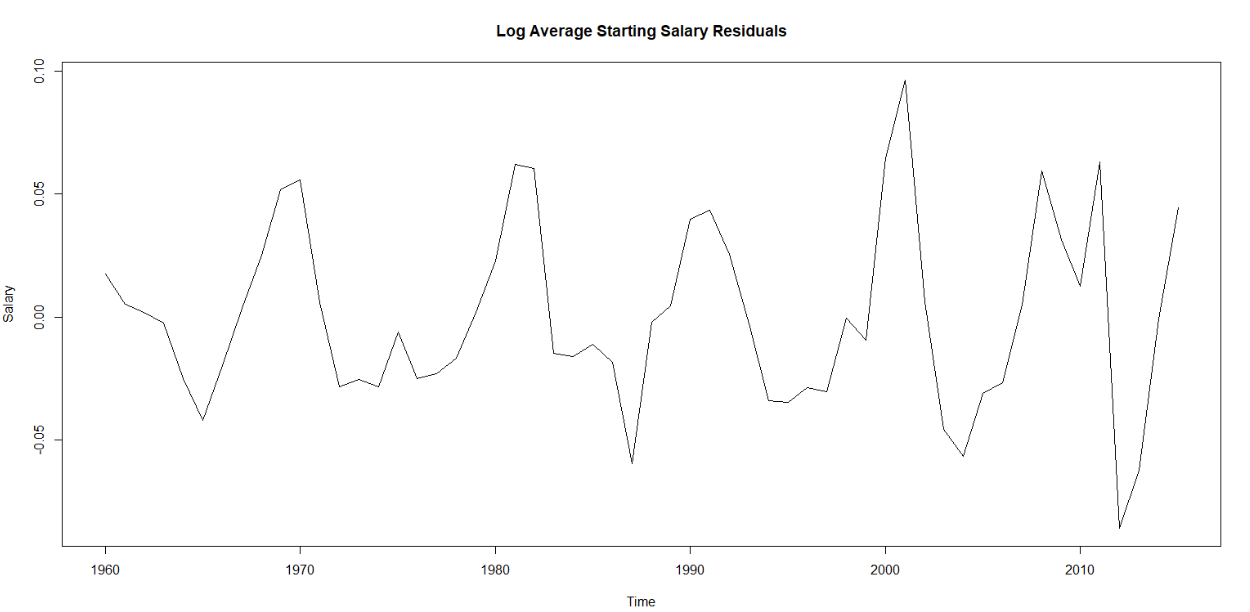
****

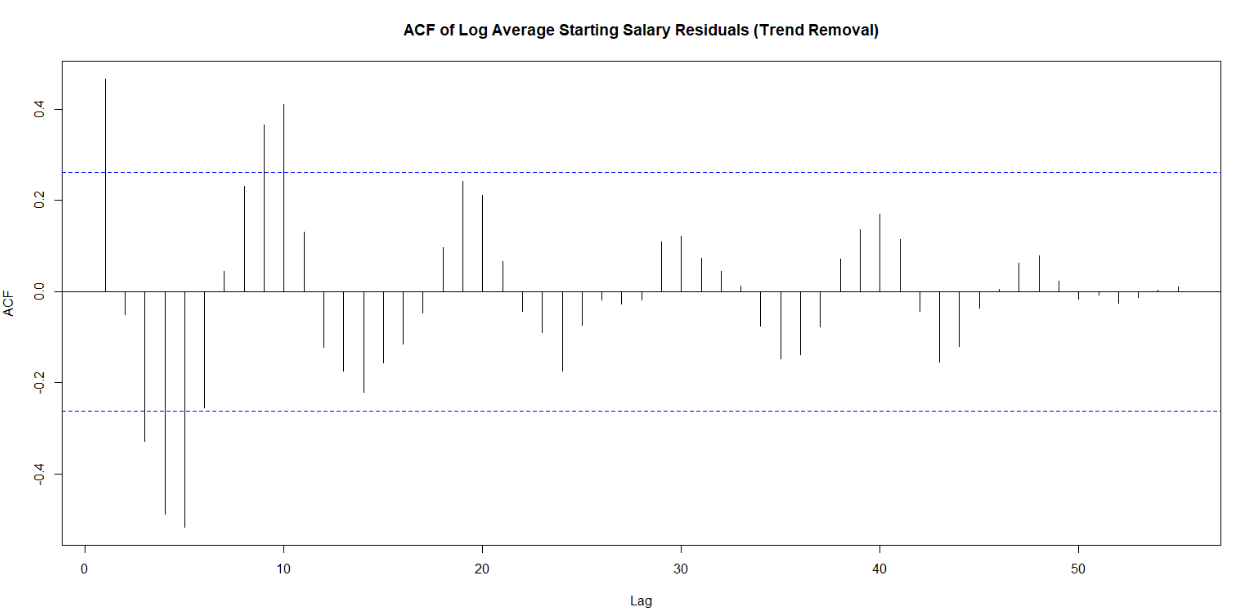
****

**Seasonal Trends**









1. This study is edited using Grammarly, an official typing assistant provided by Georgia Tech. [↑](#footnote-ref-1)
2. For full dataset on Kaggle’s Student Loan Data, please see: <https://www.kaggle.com/datasets/omarsobhy14/student-loans> [↑](#footnote-ref-2)
3. For full dataset on NY Federal Reserve Bank’s Student Loan Data, please see: <https://www.newyorkfed.org/microeconomics/databank.html> [↑](#footnote-ref-3)
4. National Association of Colleges and Employers; 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-4)
5. For full data 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-5)
6. For full dataset on Inflation, please see: <https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=US> [↑](#footnote-ref-6)
7. For full dataset on GDP, please see: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US [↑](#footnote-ref-7)
8. For full dataset on Population, please see: <https://www.statista.com/statistics/1067138/population-united-states-historical/> [↑](#footnote-ref-8)
9. The data is available since the early 18th century, but this study only focuses from 1960 in accordance with other datasets. [↑](#footnote-ref-9)
10. 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) used a 0.05 significance level. [↑](#footnote-ref-10)