Sweden, Household debt Analysis

Author: Kanchana Weerasinghe
Course Coordinator: Moudud Alam (Senior Lecturer Statistics)
April 29, 2024

Abstract

This study investigates the relationship between inflation and household debt in Sweden, a topic spurred by a recent surge in debt levels as reported in a March 2021 news article, which highlighted that Swedish household debt had reached a record high of 97.8% of the nominal GDP. Utilizing historical data extracted from the "SwedishHouseholdDebt.xls" file, this paper explores whether inflation, as measured by the Consumer Price Index (CPI), significantly influences household debt accumulation over time. A detailed analysis examines the correlation between CPI fluctuations and changes in debt ratios, providing insights into the causal impact of inflation on the economic behavior of households. Additionally, the study applies statistical modeling to predict current household debt levels, which have adjusted to approximately 87% of nominal GDP. The findings contribute to a broader understanding of economic pressures faced by households in Sweden and offer predictive insights that policymakers and economic analysts might find valuable in addressing or mitigating the impact of inflation on household indebtedness.

***Keywords**:* Real GDP per capita, Real GDP per capita growth, Average years of school of population aged 15 and over, Annual changes in consumer prices, Total dependency ratio, Banking crisis, Household and NPISH, all liabilities, Inflation

1. Introduction

Household debt is a critical indicator of financial health within a country, reflecting both consumer confidence and potential vulnerabilities within the economic system. In Sweden, the recent spike in household debt to 97.8% of the nominal GDP as of March 2021 raises significant concerns regarding the underlying factors contributing to this increase and its implications on the economy. This paper aims to dissect the interrelations among several economic indicators including Real GDP per capita, its growth rate, educational attainment, inflation, dependency ratios, and financial crises to elucidate their impact on household debt levels.

Real GDP per capita is a fundamental measure of a nation's economic output relative to its population size, providing a lens through which to view the economic prosperity and standard of living of its residents. The growth rate of Real GDP per capita further informs about the pace at which an economy is expanding or contracting, potentially influencing household debt as consumers adjust their borrowing in response to changes in economic conditions. Education, measured by the average years of schooling of the population aged 15 and over, is another pivotal factor that potentially affects household debt, as higher educational attainment can lead to better financial decisions due to increased economic awareness and literacy. The annual changes in consumer prices, typically indicated by inflation rates, directly impact household purchasing power and saving behaviors, which in turn can influence household debt levels. The total dependency ratio, which assesses the non-working population (young and old) supported by the working-age population, can affect economic pressure on productive households, thereby potentially influencing debt levels as families borrow to meet broader familial financial needs.

This study integrates these elements to provide a comprehensive analysis of the factors that drive household debt levels in Sweden, with a particular focus on the role of inflation as a possible catalyst for the recent increase in debt. By employing historical data and statistical methods, the paper aims to not only explore these relationships but also predict future trends in household debt, offering valuable insights for policymakers, economic analysts, and the broader public interested in the financial dynamics of Sweden.

2. Methodology

This section provides an overview of the process initiated to achieve the aim, which is illustrated in Figure 1.

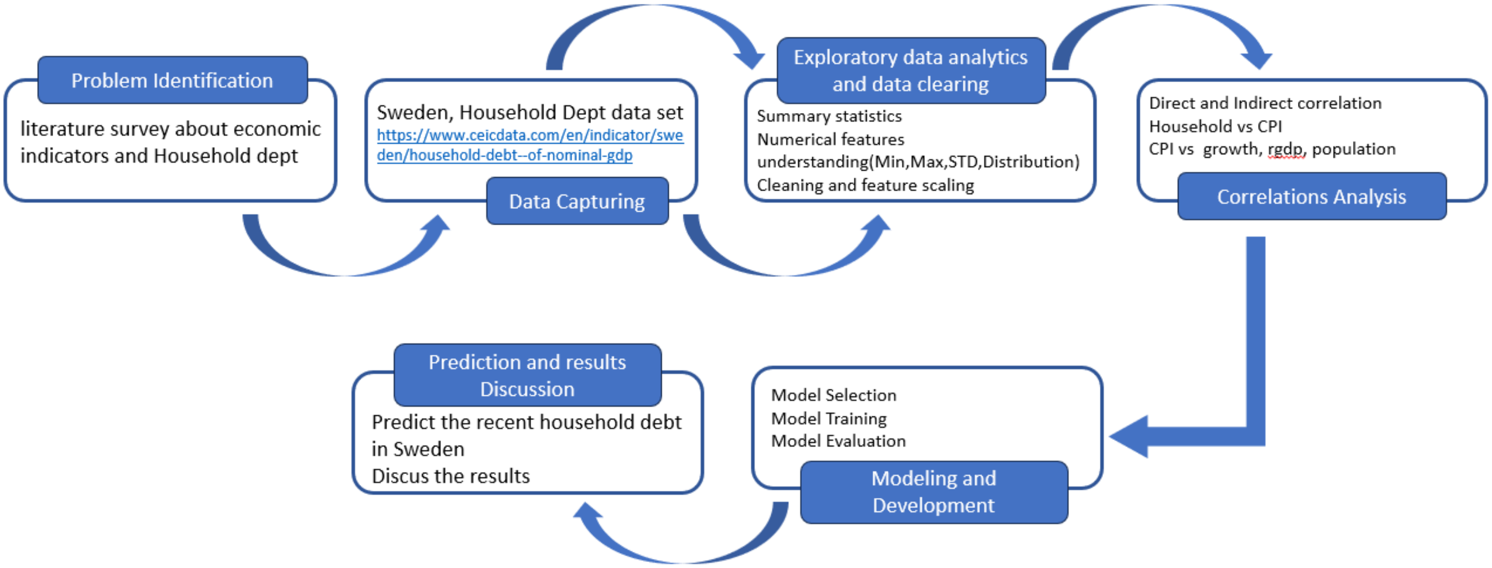
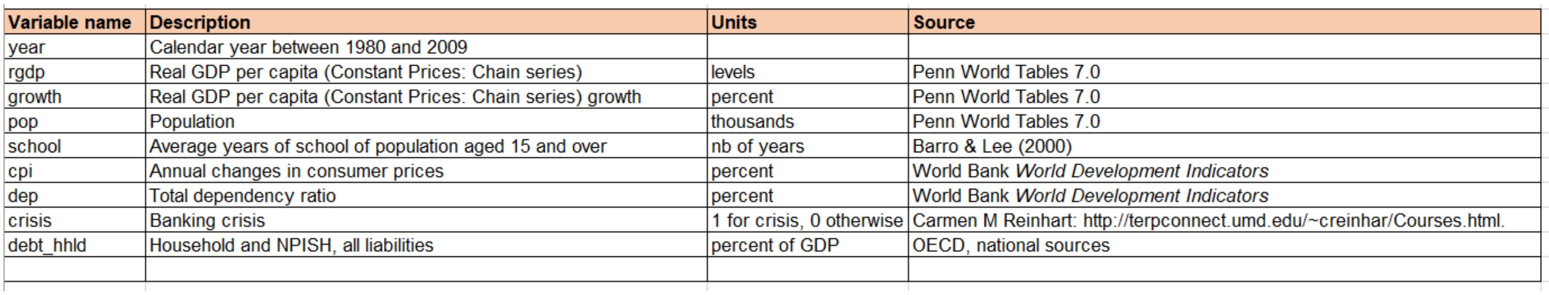

Problem Identification
literature survey about economic indicators and Household dept
Sweden, Household Dept data set
https://www.ceicdata.com/en/indicator/sweden/household-debt--of-nominal-gdp
Data Capturing
Exploratory data analytics and data clearing
Summary statistics
Numerical features
understanding(Min,Max,STD,Distribution)
Cleaning and feature scaling
Direct and Indirect correlation
Household vs CPI
CPI vs growth, rgdp, population
Correlations Analysis
Model Selection
Model Training
Model Evaluation
Modeling and Development
Prediction and results Discussion
Predict the recent household debt in Sweden
Discus the results

Figure 1- Process flow diagram

3. Data Collection

This dataset contains economic and demographic variables for the years between 1980 and 2009. It includes annual data on real GDP per capita at constant prices to reflect economic size and growth, as well as population figures in thousands. The dataset also measures the average educational attainment of the population aged 15 and over, changes in consumer prices annually, and the total dependency ratio, all of which provide insights into the economic health and social structure of the nations covered. Additionally, it tracks occurrences of banking crises and household debt as a percentage of GDP, sourced from various reputable databases and studies. Here is a brief summary of the key variables included:



Variable name	Description	Units	Source
year	Calendar year between 1980 and 2009		
rgdp	Real GDP per capita (Constant Prices: Chain series)	levels	Penn World Tables 7.0
growth	Real GDP per capita (Constant Prices: Chain series) growth	percent	Penn World Tables 7.0
pop	Population	thousands	Penn World Tables 7.0
school	Average years of school of population aged 15 and over	nb of years	Barro & Lee (2000)
cpi	Annual changes in consumer prices	percent	World Bank *World Development Indicators*
dep	Total dependency ratio	percent	World Bank *World Development Indicators*
crisis	Banking crisis	1 for crisis, 0 otherwise	Carmen M Reinhart: http://terpconnect.umd.edu/~creinhar/Courses.html.
debt_hhld	Household and NPISH, all liabilities	percent of GDP	OECD, national sources

4. Data preparation

4.1 Exploratory data analytics and data clearing

In the exploratory data analysis (EDA), I focused on examining the numerical features of the dataset related to household debt in Sweden, including economic indicators such as Real Gross Domestic Product (rgd), growth rates, cpi ,pop ,school and dep. Summary statistics , unique values, missing values, minimum, max, mean, STD, distribution (histogram) has been used to understand behaviors of the data set.

Additionally, rigorous data cleaning procedures were implemented to ensure the quality and integrity of the dataset. This involved identifying and addressing issues such as missing values, outliers, and inconsistencies that could potentially impact the accuracy of our analysis. Special attention was paid to maintaining the accuracy and reliability of the data while preparing it for further analysis, ensuring that our predictive models would be built on a solid foundation of high-quality data.

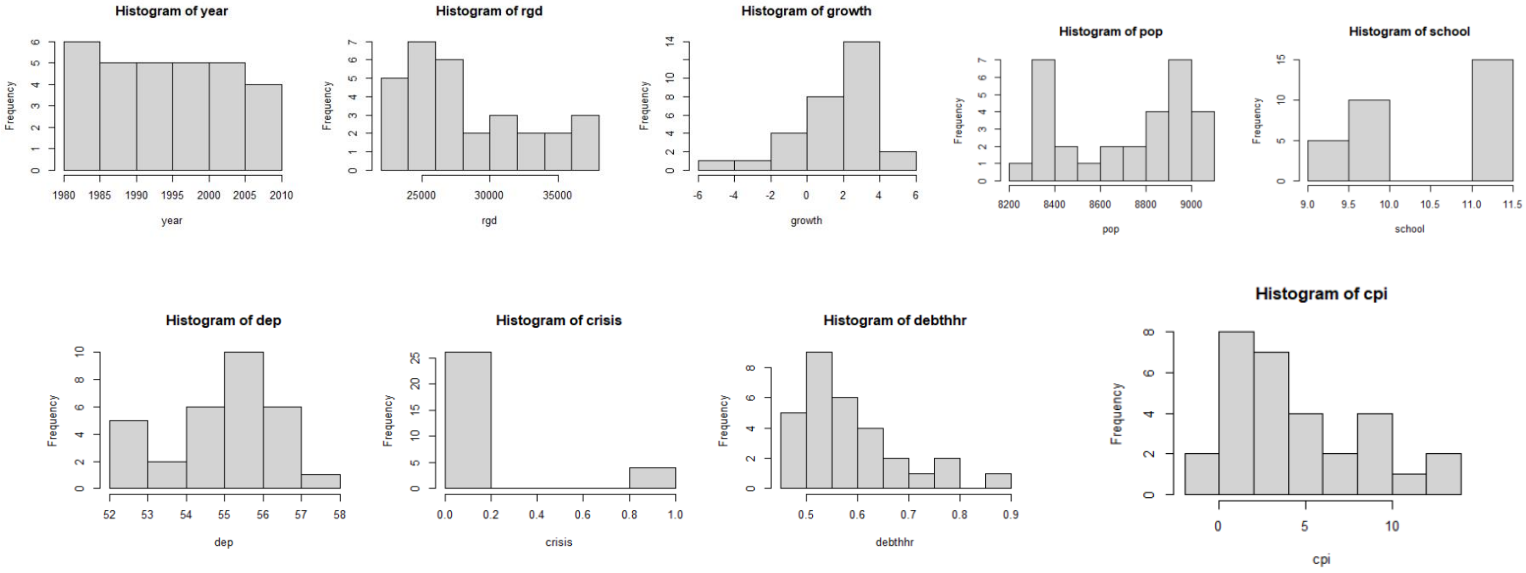
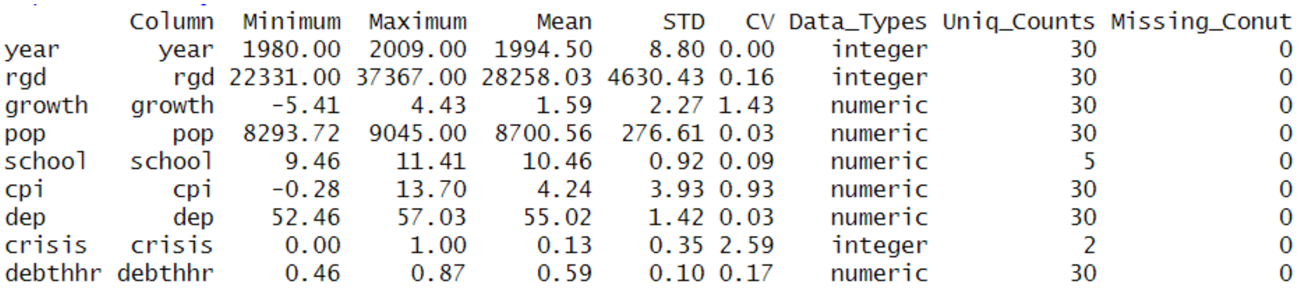


Figure 2- Numerical data distributions

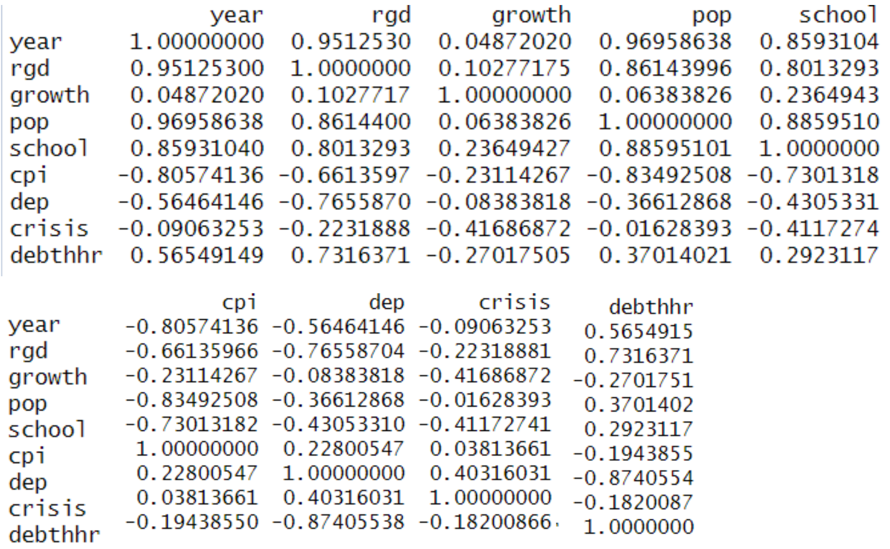


```
         Column  Minimum  Maximum     Mean     STD   CV Data_Types Uniq_Counts Missing_Conut
year       year  1980.00  2009.00  1994.50    8.80 0.00    integer          30             0
rgd         rgd 22331.00 37367.00 28258.03 4630.43 0.16    integer          30             0
growth   growth    -5.41     4.43     1.59    2.27 1.43    numeric          30             0
pop         pop  8293.72  9045.00  8700.56  276.61 0.03    numeric          30             0
school   school     9.46    11.41    10.46    0.92 0.09    numeric           5             0
cpi         cpi    -0.28    13.70     4.24    3.93 0.93    numeric          30             0
dep         dep    52.46    57.03    55.02    1.42 0.03    numeric          30             0
crisis   crisis     0.00     1.00     0.13    0.35 2.59    integer           2             0
debthhr debthhr     0.46     0.87     0.59    0.10 0.17    numeric          30             0
```

Figure 3- Summary statistics

4.2 Correlation and coefficients Analysis

This analysis was done to understand the provides insights into the relationship between two variables in a dataset under Strength of Relationship, Direction of Relationship, Linearity, Independence, Potential Causality



```
              year        rgd      growth         pop     school
year    1.00000000  0.9512530  0.04872020  0.96958638  0.8593104
rgd     0.95125300  1.0000000  0.10277175  0.86143996  0.8013293
growth  0.04872020  0.1027717  1.00000000  0.06383826  0.2364943
pop     0.96958638  0.8614400  0.06383826  1.00000000  0.8859510
school  0.85931040  0.8013293  0.23649427  0.88595101  1.0000000
cpi    -0.80574136 -0.6613597 -0.23114267 -0.83492508 -0.7301318
dep    -0.56464146 -0.7655870 -0.08383818 -0.36612868 -0.4305331
crisis -0.09063253 -0.2231888 -0.41686872 -0.01628393 -0.4117274
debthhr  0.56549149  0.7316371 -0.27017505  0.37014021  0.2923117

               cpi         dep      crisis      debthhr
year    -0.80574136 -0.56464146 -0.09063253   0.5654915
rgd     -0.66135966 -0.76558704 -0.22318881   0.7316371
growth  -0.23114267 -0.08383818 -0.41686872  -0.2701751
pop     -0.83492508 -0.36612868 -0.01628393   0.3701402
school  -0.73013182 -0.43053310 -0.41172741   0.2923117
cpi      1.00000000  0.22800547  0.03813661  -0.1943855
dep      0.22800547  1.00000000  0.40316031  -0.8740554
crisis   0.03813661  0.40316031  1.00000000  -0.1820087
debthhr -0.19438550 -0.87405538 -0.18200866.  1.0000000
```

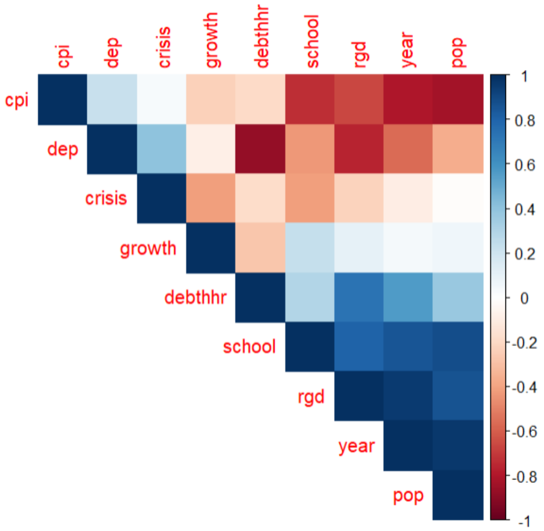


Figure 4- Correlation matrix

5 Results and Discussion

5.1 Inflation vs house hold dept analysis (Question 1)

According to the above result The correlation coefficient between CPI (Consumer Price Index) and Household Debt is approximately -0.19.This negative correlation suggests a weak inverse relationship between CPI and Household Debt. In other words, as CPI decreases (indicating lower inflation rates), Household Debt tends to increase slightly, and vice versa.

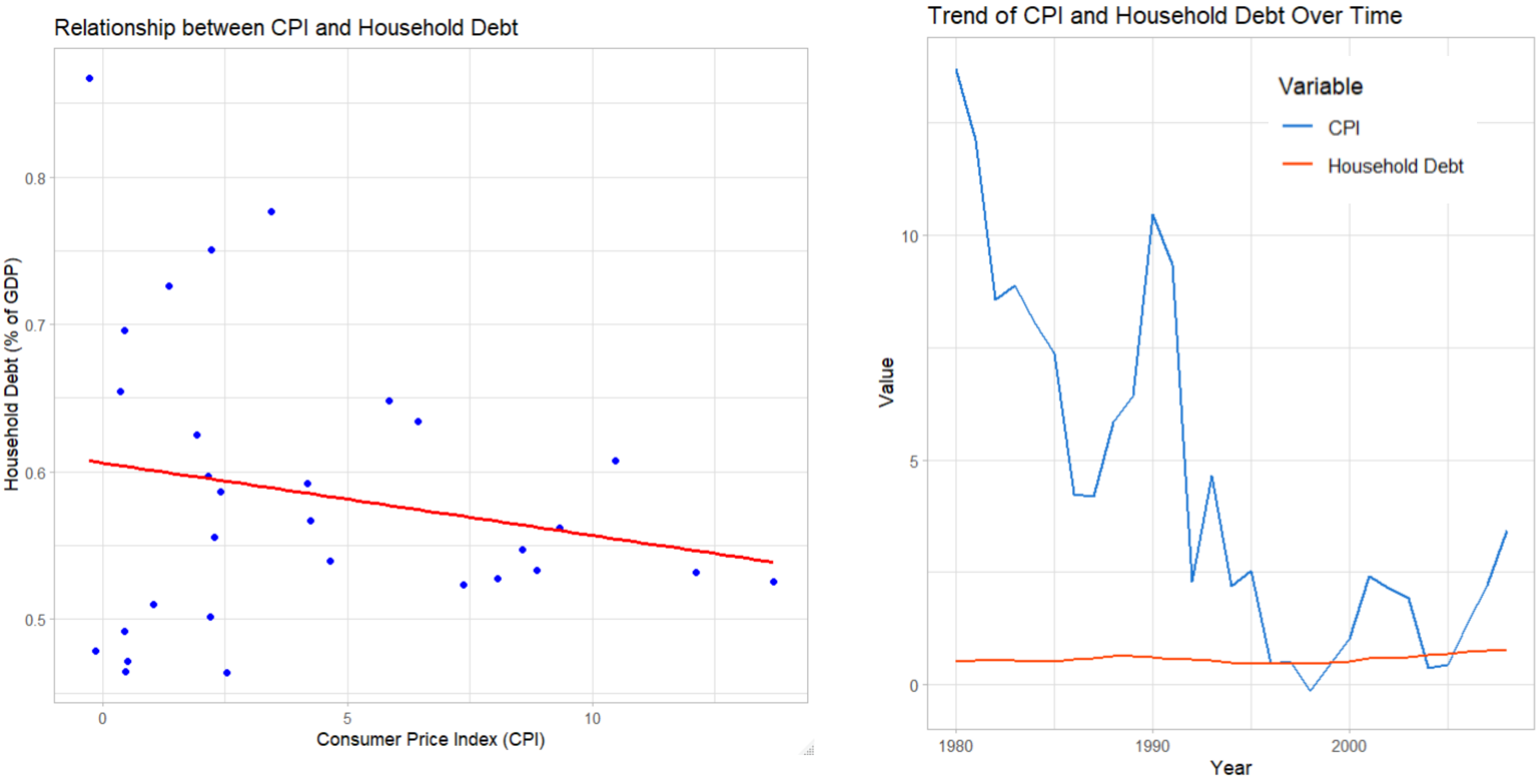


Figure 5- Relationship between Household Debt and CPI / Tend of Household Debt and CPI over the time

Based on the analysis, the correlation between Household Debt and CPI is quite weak (-0.19), indicating a slight negative relationship. However, over time, while inflation (measured by CPI) fluctuates, household debt demonstrates a more consistent pattern with a slight upward trend.

Therefore, while there may be some tendency for Household Debt to move in the opposite direction of CPI changes.

5.2 Predict the Recent Household Debt (Question 2)

To predict the Household Dept Conducted exploratory data analysis to understand the dataset's numerical features, focusing on household debt trends in Sweden over time.

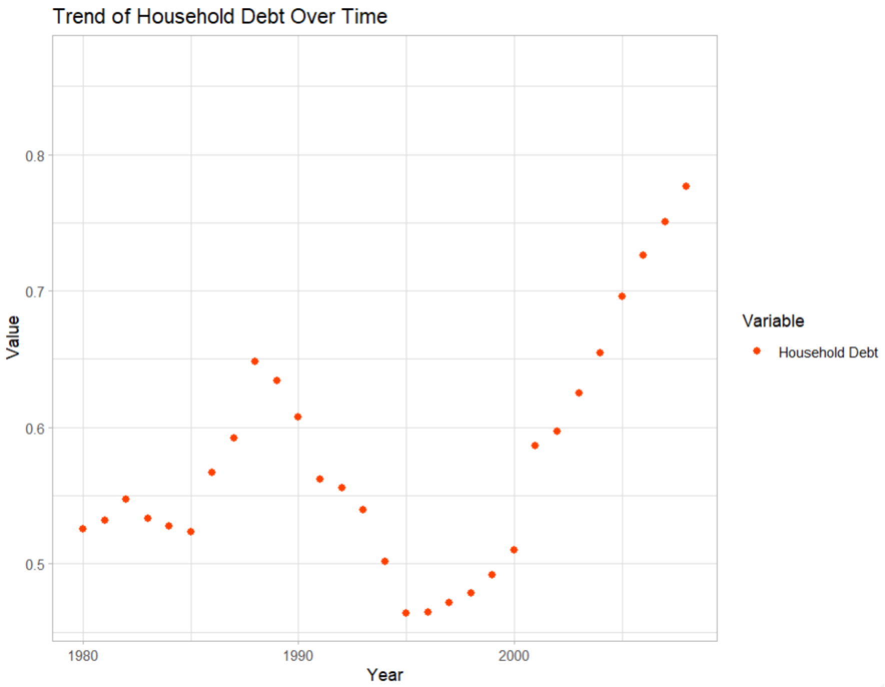


Figure 6- Household Debt over the time

5.2.1 Feature selection:

To select the features for prediction model, correlation analysis has been conducted against to the Household Debt as *Figure 4*

dep (-0.87): There is a strong negative correlation between household debt (debthhr) and dependency ratio (dep). This suggests that a higher dependency ratio (indicating a larger proportion of the population not in the workforce) is associated with lower household debt.

rgd (0.73): There is a moderately strong positive correlation between household debt (debthhr) and real GDP (rgd). This suggests that as real GDP increases, household debt tends to increase as well.

pop (0.37): There is a moderate positive correlation between household debt (debthhr) and population (pop). This indicates that as the population grows, household debt tends to increase.

school (0.29): There is a weak positive correlation between household debt (debthhr) and school enrollment (school). This suggests that higher levels of school enrollment may be associated with slightly higher household debt.

growth (-0.27): There is a weak negative correlation between household debt (debthhr) and economic growth (growth). This suggests that higher economic growth may be associated with slightly lower household debt.

crisis (-0.18): There is a weak negative correlation between household debt (debthhr) and the occurrence of a crisis (crisis). This indicates that during periods of crisis, there may be slightly lower household debt.

Features significancy testing with Random Forest model

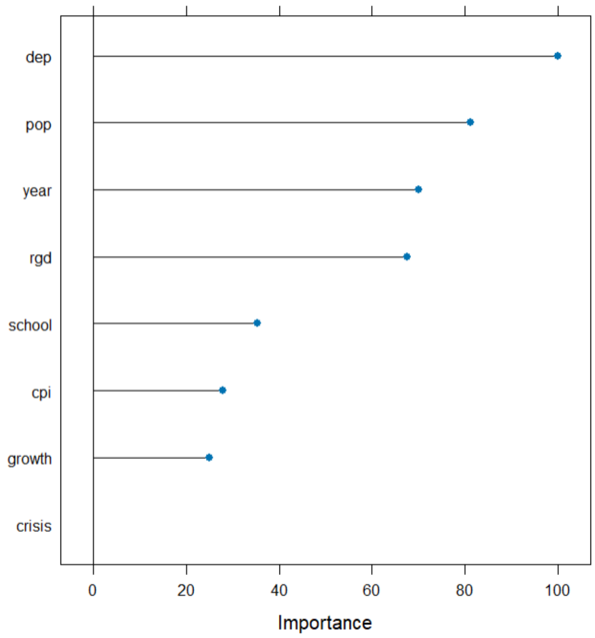


Figure 7- Feature impotency to predict Household Debt

5.2.2 Model designing and evaluation

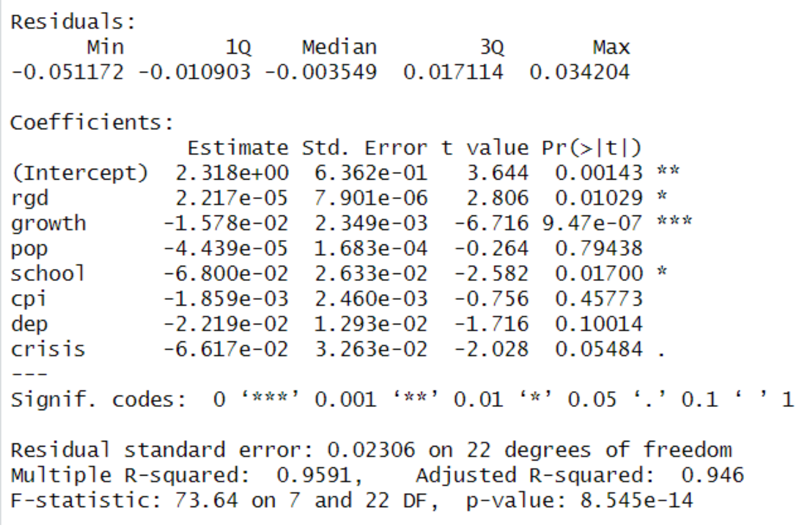
Developed predictive models using Linear Regression, Lasso Regression, and Ridge Regression techniques to predict the Household Debt.

5.2.2.1 Linear Regression

Before running the Linear Regression, the records corresponding to the year 2009 have been removed from the dataset to ensure that they remain unseen during model training

year	rgd	growth	pop	school	cpi	dep	crisis	debthhr
2009	35225	-5.40941	9045	11.41	-0.27843	52.8206	0	0.867368
2021	61418	8.13	1042	11.41	1.6	61	0	0.9748

Liner regression model performance summary based on the K-Fold cross validation



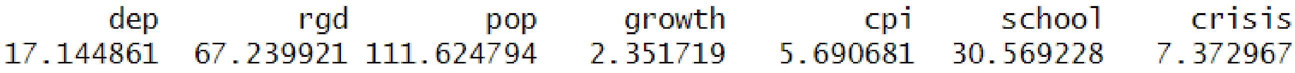
```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.051172 -0.010903 -0.003549  0.017114  0.034204 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.318e+00  6.362e-01   3.644  0.00143 ** 
rgd          2.217e-05  7.901e-06   2.806  0.01029 *  
growth      -1.578e-02  2.349e-03  -6.716 9.47e-07 ***
pop         -4.439e-05  1.683e-04  -0.264  0.79438    
school      -6.800e-02  2.633e-02  -2.582  0.01700 *  
cpi         -1.859e-03  2.460e-03  -0.756  0.45773    
dep         -2.219e-02  1.293e-02  -1.716  0.10014    
crisis      -6.617e-02  3.263e-02  -2.028  0.05484 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02306 on 22 degrees of freedom
Multiple R-squared:  0.9591,	Adjusted R-squared:  0.946 
F-statistic: 73.64 on 7 and 22 DF,  p-value: 8.545e-14
```

The residuals (errors) range from -0.051172 to 0.034204, indicating that the model's predictions vary from the actual values by up to approximately 3.42%.The model's multiple R-squared value of 0.9591 indicates that approximately 95.91% of the variability in the response variable (debthhr) is explained by the predictor variables. With a value of 73.64 and a p-value of 8.545e-14, the model is considered statistically significant. Overall, this summary suggests that the model has a good fit and includes several significant predictors.

Predicted result with unseen data 2009 is - 0.7748666 and 2021 is 1.48783 but the actual value should around 0.867368 for 2009 and 0.9748 for 2021 , that indicate this model may has some overfitting issue or Multicollinearity

5.2.2.1.1 Multicollinearity testing

Multicollinearity has been tested using Variance Inflation Factor (VIF)



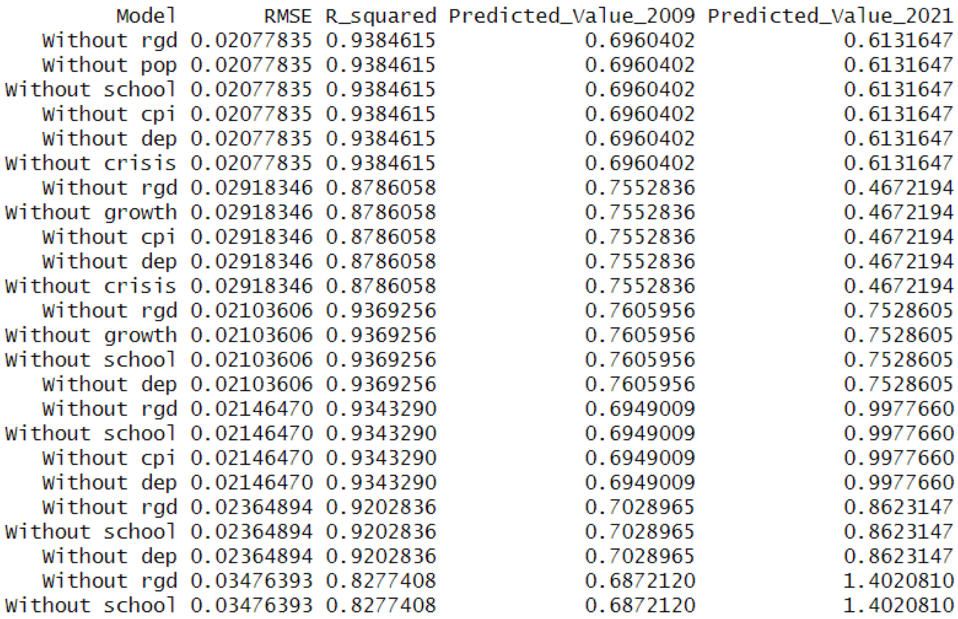
```
      dep       rgd       pop    growth       cpi    school    crisis 
17.144861 67.239921 111.624794  2.351719  5.690681 30.569228  7.372967 
```

In particular, **pop**, **rgd**, and **school, dep** appear to have very high levels of multicollinearity based on the provided VIF values. This should be the reason for this variance

After removing highly correlated variables such as **'pop'** and **'rgd'** and predicting on unseen records, the model performance improved. The predicted value of 0.7943215 for 2009 is closer to the actual value of 0.867368 but the 2021 predicted value deviate form 0.9748 drastically showing 0.4790. However, the R-squared value decreased to 0.8646, and the root mean squared error (RMSE) increased to 0.03461. These changes suggest that while the model's accuracy decreased and its fit to the data weakened (bias decreased), it has effectively mitigated the issue of overfitting.

5.2.2.1.2 Backward feature selection to test the improvement of the model

This testing has been done to verify whether removing multicollinearity automatically without observing multicollinearity values will improve the performance and reduce the variance

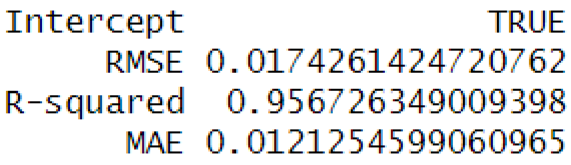


```
          Model       RMSE R_squared Predicted_Value_2009 Predicted_Value_2021
    Without rgd 0.02077835 0.9384615            0.6960402            0.6131647
    Without pop 0.02077835 0.9384615            0.6960402            0.6131647
 Without school 0.02077835 0.9384615            0.6960402            0.6131647
    Without cpi 0.02077835 0.9384615            0.6960402            0.6131647
    Without dep 0.02077835 0.9384615            0.6960402            0.6131647
 Without crisis 0.02077835 0.9384615            0.6960402            0.6131647
    Without rgd 0.02918346 0.8786058            0.7552836            0.4672194
 Without growth 0.02918346 0.8786058            0.7552836            0.4672194
    Without cpi 0.02918346 0.8786058            0.7552836            0.4672194
    Without dep 0.02918346 0.8786058            0.7552836            0.4672194
 Without crisis 0.02918346 0.8786058            0.7552836            0.4672194
    Without rgd 0.02103606 0.9369256            0.7605956            0.7528605
 Without growth 0.02103606 0.9369256            0.7605956            0.7528605
 Without school 0.02103606 0.9369256            0.7605956            0.7528605
    Without dep 0.02103606 0.9369256            0.7605956            0.7528605
    Without rgd 0.02146470 0.9343290            0.6949009            0.9977660
 Without school 0.02146470 0.9343290            0.6949009            0.9977660
    Without cpi 0.02146470 0.9343290            0.6949009            0.9977660
    Without dep 0.02146470 0.9343290            0.6949009            0.9977660
    Without rgd 0.02364894 0.9202836            0.7028965            0.8623147
 Without school 0.02364894 0.9202836            0.7028965            0.8623147
    Without dep 0.02364894 0.9202836            0.7028965            0.8623147
    Without rgd 0.03476393 0.8277408            0.6872120            1.4020810
 Without school 0.03476393 0.8277408            0.6872120            1.4020810
```

But compare to the Full Model results in 3.2.1.1 (95.91% -R-Squired and 8.545e-14 P-value) model performance is not improved based on the R-Squired and RMSE values and also the prediction is not accurate than Full Model and manually reduced model for the 2009 data but some combination has predicted the 2021 data vary accurately as per the above results.

5.2.2.2 Lasso Regression

To test the model performance with Lasso regression



```
Intercept                    TRUE
     RMSE  0.0174261424720762
R-squared   0.956726349009398
      MAE  0.0121254599060965
```

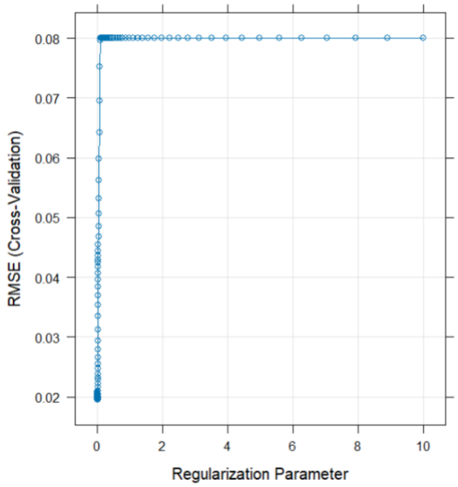
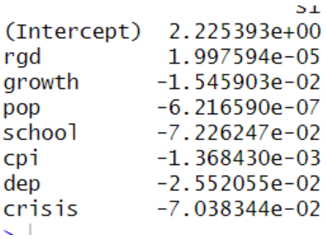


Figure 8- Lasso Regularization graph and model statistics

Testing result with unseen data is 0.7742993 for 2009 but the actual value should around 0.867368 and the predicted value for 2021 is 1.171831 which is much closer to actual value 0.9748, that indicate this model may has some underfitting issue or Multicollinearity.

Based on the results it is clear that there is no significant difference in the R-squired and the RMSE when comparing the Liner Regression Model (R-squared value 0.9591 and RMSE – 0.023)

Following results shows the coefficients of the Lasso regression model at this best lambda value

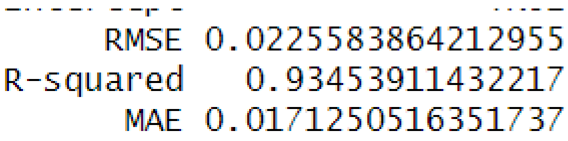


```
                      s1
(Intercept)   2.225393e+00
rgd           1.997594e-05
growth       -1.545903e-02
pop          -6.216590e-07
school       -7.226247e-02
cpi          -1.368430e-03
dep          -2.552055e-02
crisis       -7.038344e-02
```

Variable	Coefficient	Importance
rgd	1.997594e-05	Moderate
growth	-1.545903e-02	High
pop	-6.216590e-07	Low
school	-7.226247e-02	High
cpi	-1.368430e-03	Low
dep	-2.552055e-02	High
crisis	-7.038344e-02	High

5.2.2.3 Ridge Regression

To test the model performance with Ridge regression



```
     RMSE 0.0225583864212955
R-squared   0.93453911432217
      MAE 0.0171250516351737
```

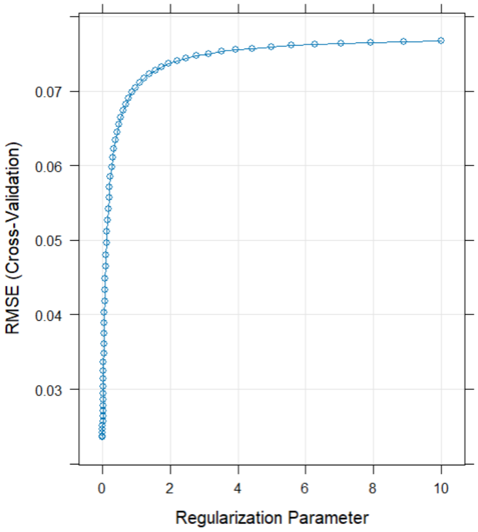


Figure 9- Ridge Regularization graph and model statistics

Based on the results it is clear that there is no significant difference in the R-squired and the RMSE when comparing the Liner Regression Model (R-squared value 0.9591 and RMSE – 0.023)

5.2.3 Model Prediction for unseen data

Based on the evaluation metrics of the model, it's evident that both Linear Regression and Lasso Regression demonstrate favorable outcomes.

Model	R-squired	RMES	P-Value
Liner Regression with all features	0.9591	0.02044	8.545e-14
Liner Regression with selected features by removing the multicollinearity	0.023	0.03461	2.89e-09
Lasso Regression	0.956726	0.0174261	
Ridge Regression	0.93453	0.0225583	

5.2.3.1 How well could you predict the recent household debt in Sweden by using the historical data (now it is about 87%)

The dataset spans from 1985 to 2009, making it challenging to gather the latest reliable data to train the model with updated economic contexts. Consequently, the predictions made using this dataset may exhibit some variations.

year	rgd	growth	pop	school	cpi	dep	crisis	debthhr
2009	35225	-5.40941	9045	11.41	-0.27843	52.8206	0	0.867368
2021	61418	8.13	1042	11.41	1.6	61	0	0.9748

Model	Prediction		
	2009 Actual Valeu- 0.867368406	2021 Actual Value 0.9748	
Liner Regression with all features	0.7748666	1.48783	
Liner Regression with selected features by removing the multicollinearity	0.7943215	0.4790	
Lasso Regression	0.7665429	0.9127475	
Ridge Regression	0.7524397	0.5892599	

In summary, Lasso Regression successfully predicted household debt for both 2009 and 2021 compared to other models. The resulting statistics demonstrate a high level of performance, with an R-squared value of 0.956726, RMSE of 0.0174261