

Hard Statistics in Economics: Investigating Economic Drivers in Individual Confidence in the DOW in the US*

A time series analysis of economic trends and their effect on stock market confidence.

Liam Wall

December 3, 2024

Over the pervious decades, the investor being able to forecast the US stock market has been a way to become very successful very fast, however we rarely focus on being able to predict how the stock market can forecast the investor. This paper investigates how well hard statistics, in regards to economics, can predict a soft statistic, the consumer confidence index in the stock market for the next year. Using the Federal Reserve Economic Data API and a long time survey managed by the Yale International Center for Finance, we employ a vector autoregression we find eveidence that despite a large effort to find a good predictive model, the best predictor of the consumer confidence index is last year's consumer confidence index. This paper also identifies nuances in the time series of some of the variables used in analysis and shows investor's just how difficult consistent predictive modeling is in economics.

1 Introduction

In recent decades, a well-known goal of statisticians in economics has been to develop reliable predictive models for various aspects of the stock market or the U.S. economy. Using consistently available economic indicators such as the unemployment rate or the VIX, these models aim to forecast whether the market will rise, fall, or remain stable. However, less attention has been given to the inverse problem: using hard data from the stock market or the economy to predict how investors will feel.

*Code and data are available at: <https://github.com/Lwall02/US-Economics>

Investors often rely on external data and mathematical models to guide their decisions, substituting raw confidence with quantifiable evidence to validate their instincts. The investor hopes that their own personal biases are avoided in relying on such methods however I argue that this is unavoidable. In this paper, I take the opposite approach: how can the stock market forecast the investor? I aim to use economic and stock market data not to predict market movements directly, but to forecast investor sentiment. By understanding how investors might feel, one can anticipate their likely reactions to market conditions. This information provides a strategic advantage, enabling investors to predict the behavior of others and act accordingly to capitalize on those expectations. Instead of deriving predictions purely from economic data, this paper focuses on deriving predictions based on the projected confidence of investors.

The estimand of this paper is the consumers' confidence in the US stock market one year from now. In other words, and which will be described in detail later, we want to be able to find the best predictor of the percentage of the population that is confident that the US stock market will increase one year from today.

The remainder of this paper is structured as follows. Section 2 discusses the raw data, cleaning process, variables of interest, and offers visual representations of the data through tables and graphs. Section 3 introduces and justifies the vector autoregression model used in the analysis of the data in predicting the relationship between variables as well as the confidence index. Section 4 deals with analyzing the trends and correlations showcased by the model in more detail. Section 5 discusses the real-world implications of the results uncovered in the prior sections, and finally Section 5.4 discusses the limitations and weaknesses of the analysis conducted, as well as the next steps that could be taken to improve the overall reliability of the paper. In the Appendix you can find additional information regarding the model details as well as a discussion of the survey used as our training and test data.

2 Data and Measurement

This paper uses six datasets to begin its analysis. Five datasets are provided through the Federal Reserve Bank of St. Louis, specifically through the Federal Reserve Economic Data (FRED) API, and one dataset is provided by the Yale International Center for Finance. The FRED API provides easy access to all of the Federal Reserve Bank of St. Louis's datasets which include current and historical financial data for the US all around the world. In particular in this paper we obtain datasets regarding core consumer price index, real GDP growth rate, the VIX valuation, the federal funds effective rate, and the unemployment rate. From Yale's International Center for Finance we obtain a dataset regarding the estimated individual and institutional confidence index in the US stock market.

All data analysis was done through R (R Core Team 2023) with the aid of the following packages: `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `ggplot2` (Wickham 2016), `knitr` (Xie 2014), `vars` (Pfaff 2008), `arrow` (Richardson et al. 2023), `kableExtra` (Zhu 2021), and `janitor` (Firke 2023). Moreover, the FRED API was accessed using the package

`fredr` (Boysel and Vaughan 2021) and the shiny web app associated with this paper was made with the aid of the `shiny` package (Chang et al. 2024).

2.1 FRED API

In order to first obtain data from the Federal Reserve Bank of St. Louis, we were able to use the `fredr` package to access the API and easily obtain any FRED datasets. In order to use their API, we first needed to create an API key through the FRED’s website [FREDwebsite]. After saving and storing the key, we can then use the `fredr` package to obtain the data. We use a code associated with each and every dataset on the FRED website to quickly download the dataset through R. We can further specify the dates we are looking for as well as the frequency of the data. To have consistent data across all variables, we use quarterly data in the paper. This gives us data for each year on January 1, April 1, July 1, and October 1.

The process of obtaining the data from the FRED API is the exact same for each of the five FRED datasets. Below is a discussion of the value and measurement of the dataset.

2.1.1 Real US GDP Growth Rate and Measurement

The Real Growth Domestic Product (GDP) growth rate measures how the inflation-adjusted value of goods and services produced in an economy changes over a specific period, typically from one quarter to the next or year-over-year. It’s one of the most widely used indicators to gauge economic performance. In this paper, we downloaded the US’s Real GDP dataset from FRED and obtained the percent change of each quarter’s real GDP to that of one year ago. Specifically we calculated $(\text{Real GDP in current quarter} - \text{Real GDP in quarter 1 year ago}) / (\text{Real GDP in quarter 1 year ago})$.

The US’s Real GDP data we obtain from FRED is in units of billions of chained 2017 dollars, seasonally adjusted annual rate. What this means is that all data in dollars from this dataset is expressed as their purchasing power in 2017, hence adjusted for inflation up to 2017. The annual rate simply means that the GDP listed for each quarter is the estimated or extended annual rate. Since we are looking at the growth rate, the data in this paper is in percents and we do not need to focus on the units of Real GDP.

We can see the Real GDP Growth Rate for all years considered in this dataset below Figure 1

2.1.2 Core CPI Data and Measurement

The Core Consumer Price Index (Core CPI) is a measure of inflation that tracks the change in prices for a basket of goods and services, excluding food and energy items. Food and energy prices tend to be highly volatile and can fluctuate significantly due to factors like weather conditions, geopolitical events, or supply disruptions (e.g., oil price shocks or crop failures).

Time Series of Quarterly GDP Growth Rate From 1990 to 2024

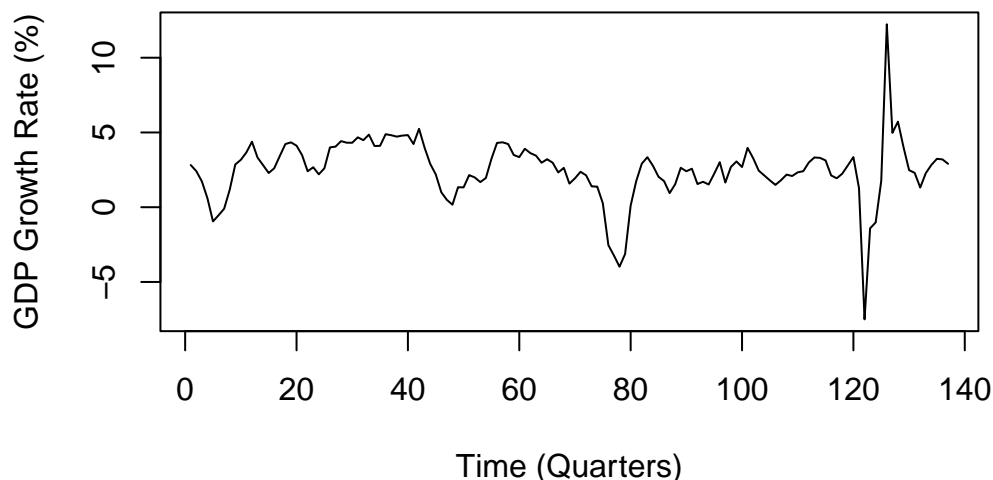


Figure 1: The quarterly US Real GDP Growth Rate for 1990 to 2024 measured in percent change compared to that of one year ago.

By excluding these items, Core CPI aims to provide a clearer picture of underlying, long-term inflation trends that are not distorted by these short-term price movements. Core CPI is one of the most widely used measures of inflation because it is considered a more reliable indicator of the general inflation trend. Economists and policymakers often use Core CPI to gauge inflation pressures that are likely to affect the economy in the long term.

Core CPI is calculated using the same methodology as the overall CPI but with food and energy prices removed. The Bureau of Labor Statistics (BLS) collects data on prices from a wide range of goods and services purchased by urban consumers and lists the core cpi as the percent increase of the typical basket of these goods and services from a listed baseline year. In this case 1984 is the baseline year so a core cpi of 264 implies that the typical basket of goods and services is 164% more expensive than it was in 1984.

We can see the Core CPI for all years considered in this dataset below [Figure 2](#)

2.1.3 Unemployment Rate and Measurement

The unemployment rate measures the number of unemployed person in the US as a percentage of the labor force. Labor force data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in

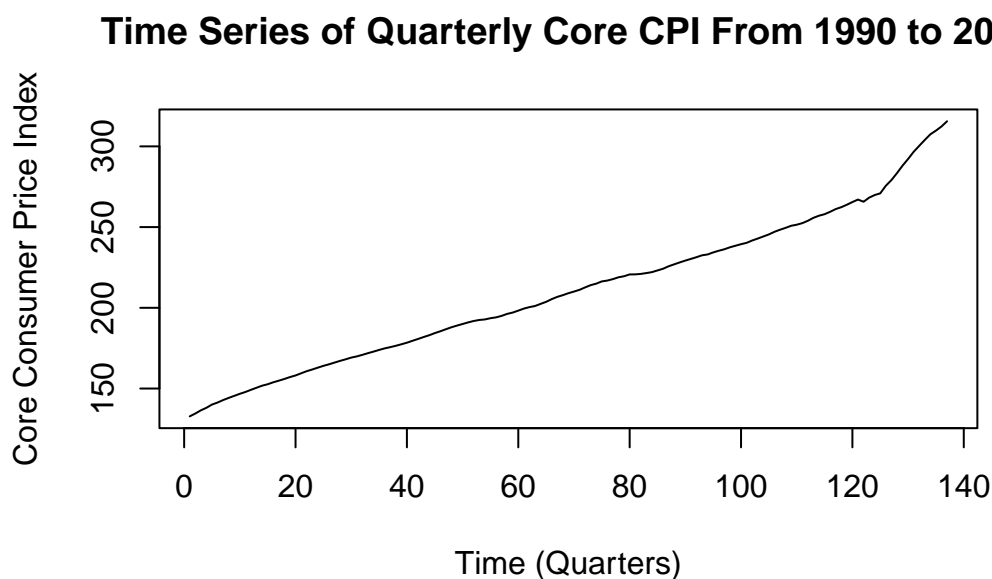


Figure 2: The quarterly Core CPI for 1990 to 2024 measured as an index of change from the baseline year of 1984.

institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces.

We can see the Unemployment Rate for all years considered in this dataset below [Figure 3](#)

2.1.4 Federal Funds Effective Rate

The Federal Funds Rate is the interest rate at which depository institutions, such as commercial banks, lend reserves to each other overnight. These reserves are balances held at Federal Reserve Banks. When a bank has surplus reserves, it lends to other banks that need additional liquidity to meet reserve requirements. The rate at which these transactions occur is determined through negotiations between the lending and borrowing banks. The Federal Funds Effective Rate is the weighted average of all such rates across the market, and while it is largely influenced by market forces, the Federal Reserve may manipulate it to adjust the rate to its target. This target is determined by the Federal Open Market Committee (FOMC), which meets eight times a year to assess economic conditions and set the federal funds rate target. The rate is a central tool for influencing the broader economy, impacting the cost of borrowing and the level of liquidity in the financial system.

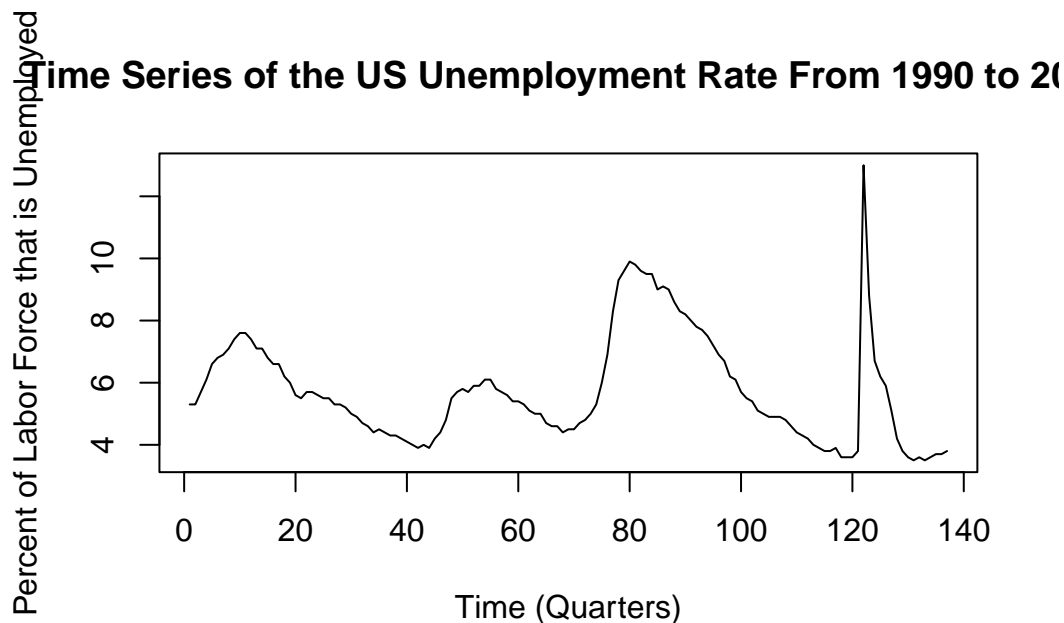


Figure 3: The quarterly Unemployment Rate for 1990 to 2024 measured as a percent of the labor force.

The Federal Reserve manipulates the federal funds rate through buying or selling government bonds to influence the liquidity available in the market. If the economy is overheating, the FOMC may sell government bonds to reduce liquidity and raise the federal funds rate. On the other hand, if the economy is sluggish, the FOMC may buy bonds to increase liquidity and lower the federal funds rate, encouraging borrowing and investment. The federal funds rate is the most important interest rate in the U.S. financial system, as it influences other rates like, mortgages, loans, and savings rates. Changes in the federal funds rate can affect consumer wealth, confidence, and overall economic activity, making it a critical tool for the Federal Reserve in managing the economy's growth while maintaining price stability and maximum employment, as mandated by Congress.

We can see the Unemployment Rate for all years considered in this dataset below [Figure 4](#)

2.1.5 The VIX Dataset and Measurement

The VIX (Volatility Index) measures the market's expectation of near-term volatility based on the prices of stock index options. Often referred to as the "fear gauge," the VIX reflects the market's consensus on the level of volatility expected in the coming 30 days. It is calculated by the Chicago Board Options Exchange (CBOE) using the prices of options on the S&P 500 index. The VIX is expressed as an annualized percentage, which indicates the expected

Series of the quarterly Federal Funds Effective Rate From 19

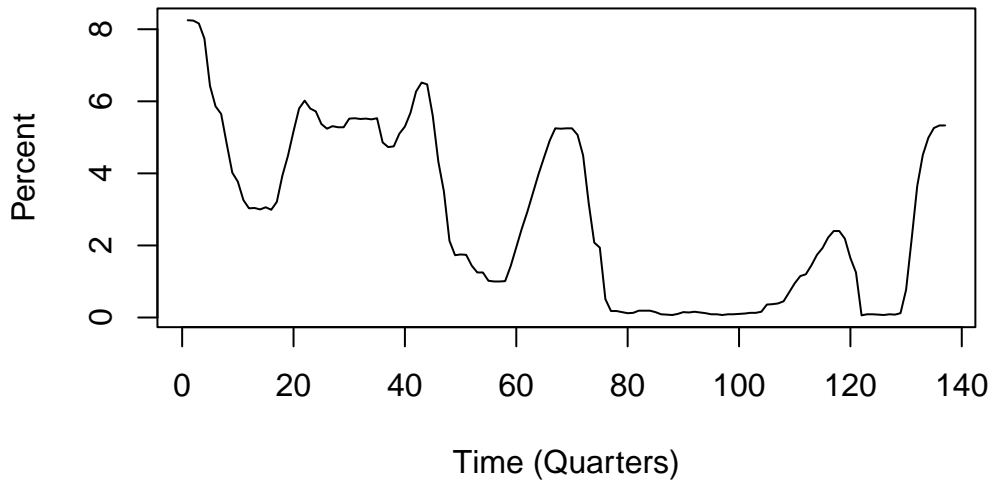


Figure 4: The quarterly Federal Funds Effective Rate from 1990 to 2024.

volatility (in percentage points) over the next 30 days. The formula to calculate the VIX incorporates the prices of both calls and puts, weighing options with different strike prices and maturities. Higher VIX values indicate a higher expected volatility, signaling increased uncertainty or fear in the market, while lower values suggest stability or confidence. As such, the VIX is widely used by investors, analysts, and policymakers as a gauge of market sentiment, helping to assess the potential for market swings and guide investment decisions.

We can see the VIX for all years considered in this dataset below [Figure 5](#)

2.2 Yale International Center for Finance

The sixth and last dataset used in this paper comes from the Yale International Center for Finance (ICF). Specifically it comes from survey data that the ICF publishes coming from a longstanding survey of wealthy investors and institutions. From 1989 to 1990 the survey was taken from a sample of investors purchased from a survey company called W.S. Ponton, Inc. and in 1999 the sample changed to a survey of wealth American Investors purchased from Survey Sampling, Inc. Both individuals and institutions were included in the sample and their responses are recorded separately.

Starting in July 2001, the beginning date of analysis for this paper, the survey was given monthly and the results were the average of the previous six months of surveys. For example

Time Series of the quarterly VIX From 1990 to 2024

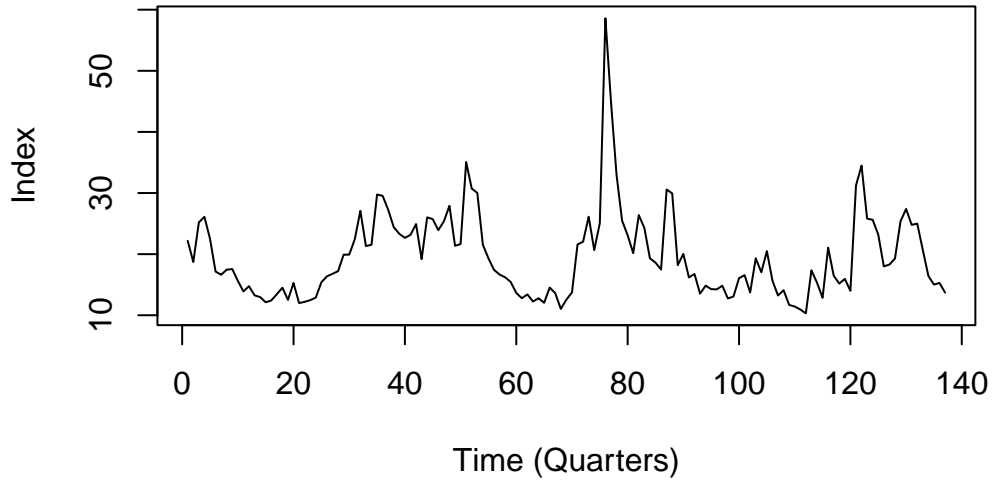


Figure 5: The quarterly VIX from 1990 to 2024.

the number for January 2018 is an average of results from surveys between August 2017 and January 2018. We go into more detail about the questions and calculation of the answers in the next section as well as in the appendix.

2.2.1 US Confidence

The data for the estimate of the US confidence index in the stock market for the next year comes from the survey data released by the ICF. Specifically, the index is the percent of people who said that the DOW would increase one year from taking the survey. This question was always asked on the survey from 1989 to present. The question is offered to both institutions and wealthy investors, however they consistently display similar findings.

Below we can see the one year confidence index for institutions [Figure 6](#)

3 Model

As the purpose of this paper is to investigate the relationship between various economic factors and the U.S. consumer confidence index in the stock market, we will employ a model that allows us to examine how these factors contribute to predicting investor sentiment over time.

Time Series of the 1 Year Confidence Index From 2001 to 2020

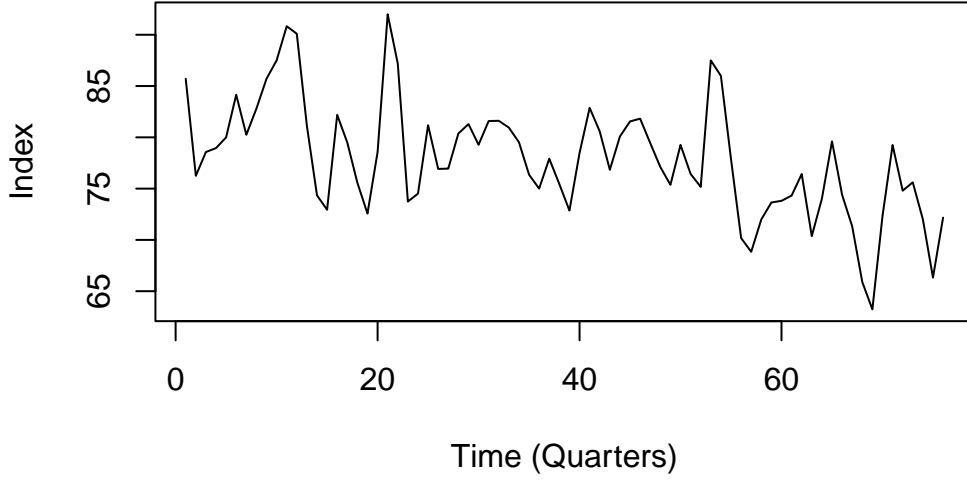


Figure 6: The quarterly 1 year confidence index from 2001 to 2020.

Specifically, we aim to understand how variables such as GDP growth, unemployment rates, the VIX, and the core CPI interact with the consumer confidence index. By using this information, we will develop a predictive model that estimates the future confidence index based on current and past economic data. The goal is to determine whether these economic indicators are effective predictors of consumer confidence, or if other factors beyond the economic data—such as investor psychology or external market events—play a larger role in shaping market sentiment.

Here we briefly describe the vector auto regression (VAR) model used to investigate predictors of market confidence. Background details and diagnostics are included in the Appendix.

3.1 Model set-up

The particular model that this paper will utilize is a vector auto regression model of consumer one-year confidence index as a function of the economic variables present in our analysis. In particular, the model is as follows:

$$confidence_t = \beta_0 + \beta_1 gdp_{t-1} + \beta_2 cpi_{t-1} + \beta_3 unemployment_{t-1} \quad (1)$$

$$+ \beta_4 federal - rate_{t-1} + \beta_5 vix_{t-1} + \beta_6 confidence_{t-1} \quad (2)$$

In the above model:

- β_0 is the coefficient for the intercept.
- β_1 is the coefficient for the predicted change in confidence index given a one unit increase in the GDP Growth Rate at time $t - 1$
- β_2 is the coefficient for the predicted change in confidence index given a one unit increase in the Core CPI at time $t - 1$
- β_3 is the coefficient for the predicted change in confidence index given a one unit increase in the Unemployment Rate at time $t - 1$
- β_4 is the coefficient for the predicted change in confidence index given a one unit increase in the Federal Funds Effective Rate at time $t - 1$
- β_5 is the coefficient for the predicted change in confidence index given a one unit increase in the VIX at time $t - 1$
- β_6 is the coefficient for the predicted change in confidence index given a one unit increase in the One-Year Confidence Index at time $t - 1$

Note that we use `vars` to run the regression. Since all of the variables are a function of time, we are running a model on a multivariate time series. Discussion of choosing the lag as well as making sure the input data is appropriate for such a model is all in the appendix.

3.1.1 Model justification

The variables chosen for this model—GDP growth rate, core CPI, unemployment rate, VIX, federal funds effective rate, and previous confidence index values—are all grounded in economic theory and literature regarding their influence on consumer confidence. The confidence index reflects individuals' expectations about the stock market, specifically their belief that the Dow Jones will increase over the next year, making it closely linked to these macroeconomic variables.

We expect a linear relationship between these variables and the confidence index. Specifically, we anticipate that higher GDP growth and a lower unemployment rate will be associated with higher confidence levels, as they signal economic stability and opportunities. Conversely, increases in the VIX and federal funds rate are expected to reduce confidence, as they indicate higher uncertainty or a tightening economic environment.

We chose the VAR model because it is able to model the time dependent interplay of multiple interdependent variables. The variables in this analysis are closely related with one another, as well as being dependent on time, and we wanted a model that could reflect this. Not only do the variables effect each other, but more importantly they effect themselves. As the percentage of people who lose confidence in the market increases it is more than likely that percentage will continue to increase as time continues. A similar argument can be made for each of the other variables. That is why we chose to incorporate a lag into the model, seeing that the the values at previous times do have an effect on the value currently.

4 Results

4.1 Model Coefficients

After running the regression based on the above model, we receive the following coefficient values as showcased in `?@tbl-summary`.

We can see that in the last estimate of coefficients, the VAR model estimates the lagged coefficients for each of the variables including the lagged one-year confidence index as a predictor. What we find it rather surprising that increases in unemployment positively correlate with increases in the stock market confidence. All other variables have less of an effect except for the lagged one-year confidence interval. In fact, it is the only statistically significant finding from this model with a p-value much less than 0.05. Thus, this model tells us that the previous quarter's confidence index is the biggest predictor of the current confidence index. Followed by the unemployment rate as the second most likely predictor.

4.2 Actual vs. Predicted Confidence Indexes

In order to determine predictive performance of this model is, we can look at the predicted confidence indexes as given by the VAR model compared to the observed confidence indexes we purposely omitted from the dataset. This can be seen in `?@fig-predicted`.

	Time	Predicted	Actual
1	1	72.99945	69.70
2	2	73.29827	76.92
3	3	73.91431	71.28
4	4	74.42457	63.89
5	5	75.00227	66.30
6	6	75.57795	70.10
7	7	76.17936	75.38
8	8	76.79510	75.93
9	RMSE	5.57948	NA

Although this does not seem to be the most accurate model predictions, considering how noisy and volatile

Table 1: ?(caption)

VAR Estimation Results:

=====

Endogenous variables: gdp_diff, unemp_diff_2, vix_diff, eff_rate, inst_conf_index

Deterministic variables: trend

Sample size: 73

Log Likelihood: -539.103

Roots of the characteristic polynomial:

0.9782 0.9782 0.4008 0.3347 0.1188

Call:

VAR(y = data_matrix_2, type = "trend")

Estimation results for equation gdp_diff:

=====

gdp_diff = gdp_diff.l1 + unemp_diff_2.l1 + vix_diff.l1 + eff_rate.l1 + inst_conf_index.l1 + 1

	Estimate	Std. Error	t value	Pr(> t)
gdp_diff.l1	0.309947	0.119924	2.585	0.0119 *
unemp_diff_2.l1	-1.031733	0.412169	-2.503	0.0148 *
vix_diff.l1	0.013640	0.017118	0.797	0.4284
eff_rate.l1	-0.049705	0.062793	-0.792	0.4314
inst_conf_index.l1	0.001808	0.003008	0.601	0.5499
trend	-0.001083	0.004359	-0.249	0.8045

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7683 on 67 degrees of freedom

Multiple R-Squared: 0.1916, Adjusted R-squared: 0.1192

F-statistic: 2.647 on 6 and 67 DF, p-value: 0.02303

Estimation results for equation unemp_diff_2:

=====

unemp_diff_2 = gdp_diff.l1 + unemp_diff_2.l1 + vix_diff.l1 + eff_rate.l1 + inst_conf_index.l1

	Estimate	Std. Error	t value	Pr(> t)
gdp_diff.l1	-0.0096834	0.0354727	-0.273	0.78571
unemp_diff_2.l1	-0.3517849	0.1219169	-2.885	0.00525 **
vix_diff.l1	0.0025648	0.0050635	0.507	0.61416
eff_rate.l1	0.0324634	0.0185737	1.748	0.08508 .
inst_conf_index.l1	-0.0016989	0.0008898	-1.909	0.06049 .
trend	0.0020240	0.0012894	1.570	0.12118

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2272 on 67 degrees of freedom

Multiple R-Squared: 0.1457, Adjusted R-squared: 0.06916

F-statistic: 1.904 on 6 and 67 DF, p-value: 0.09293

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

B.2 Diagnostics

References

- Boysel, Sam, and Davis Vaughan. 2021. *Fredr: An r Client for the 'FRED' API*. <https://CRAN.R-project.org/package=fredr>.
- Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2024. *Shiny: Web Application Framework for r*. <https://CRAN.R-project.org/package=shiny>.
- Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://CRAN.R-project.org/package=janitor>.
- Pfaff, Bernhard. 2008. *VAR, SVAR and SVEC Models: Implementation Within r Package Vars*. *Journal of Statistical Software* 27(4). <https://www.jstatsoft.org/v27/i04/>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoş Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. <https://CRAN.R-project.org/package=arrow>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.