Statistical Analysis of Mortgage Rates With Time Series and Correlation

by
Raelin Domingue
Advisor: Ranadeep Daw, PhD

Table of Contents

Abstract

This project looks into two different methods of looking into the changes in mortgage rates. The first method being a regression analysis going over the different predictor variables to produce a best fit model. The second method is using a time series to look at the changes in the difference between mortgage rate and treasury yield. The time series then produces a forecast to look into the possibility of predicting the next few months using the data given.

Chapter 1

Introduction

1.1 Statement of Problem

One thing that almost everyone faces in their life is eventually dealing with the cost of housing. For many, this takes the form of a 30-year mortgage, which is the most common standard in the housing market. Naturally, the worry about mortgage costs leads to the problem of understanding the drivers that cause changes in mortgage rates. It is well known that one significant predictor of mortgage rates is the 10-year Treasury yield (Boyarchenko et al. [2019]; Gordon [2023]). However, it is also well-known [e.g., see (Gordon [2023]; Larson and Martinez [2025]; McCarthy and Peach [2002]) that the relationship between the two is not straightforward and there may be other stressors that affect mortgage rates depending on the economic scenario. In this project, we have attempted to model the 30-year mortgage rate using two approaches. We have first studied the relationship between a defined set of variables and the mortgage rate. Additionally, we have also focused on the problem from the perspective of a time-series. The time series gave insight on the behavior of the difference between the 30-year mortgage rate and the 10-year treasury yield over time.

1.2 Relevance of Problem

The mortgage rate is a relevant problem for almost everyone's life as it is closely tied with the American dream of owning an affordable home one day. (Goodman and Mayer [2018]). Therefore, it important to understand when it comes to knowing when the rates are going to rise and fall. Broadly, the mortgage rate is primarily based on the 10-year treasury rate, which is the rate an investor would expect from a federally funded security that matures after 10 years. The overall mortgage rate adds an additional spread to the base rate of the treasury (Boyarchenko et al. [2019]). Rates for individuals are based on the individual's financial conditions, lender costs, risk premiums, and the overall conditions of the market. However, as stated previously, understanding only the 10-year treasury yield may not be enough, and therefore we have also included additional predictors in our analysis.

In a broad scale, there are also effects of the mortgage rate on the housing and economic market. Especially since these rates are crucial to almost every family, new home buyer, or lender companies. (Moench and Soofi-Siavash [2022]) Further, their relevance extends to broader stock market conditions and overall economic stability. Therefore, here we study the behavior of the 30-year mortgage rate against a few possible stressors, as well as another model to understand the temporal patterns of the difference between the

mortgage rate and treasury yeild.

1.3 Literature Review

Apart from the sources mentioned above, readers are encouraged to consult a few additional references as well. When looking at the long run, mortgage rates affect the cost of housing, which is the problem that arises in this study. (McGibany and Nourzad [2004]) However, when it comes to mortgage interest rates shocks, the housing prices in areas where a majority of residents have higher debt amounts (Larson and Martinez [2025]). That study also points out how different factors and predictors come into the picture with mortgage rates. Included in things that affect the mortgage rate is the Federal Reserve and its portfolio of assets. The study shows how much it affects the upward or downward trend of the mortgage rate (Hancock and Passmore [2012]). With the Federal Reserve having control over the federal funds rate, other interest rates can only be influenced and not controlled by them. To determine the future rates, the 10-year treasury rate is important as it is what is normally referred to when predicting interest rates (Zulauf [2018]). Further technical details, along with supporting citations, are provided within the respective sections as they are discussed.

1.4 Limitations

Some limitations of our project come from the fact that it is very hard to statistically model an overall trend in such an important, everyday-use data. Such a vast globally demanding variable can often have random, short-term patterns, and can shock the financial market in vast ways. For example, COVID-19 had huge medium-term effects and increased financial uncertainty (Guirguis and Trieste [2020]). In the time series model, there is a local concave trend approximately every four to five years, where the difference between the mortgage and treasury rate initially spikes and then follows the treasury rate downward. These patterns are challenging to model and also highlight that factors beyond the treasury yield can stress mortgage rates.

Another thing that may effect the rate is monetary policy passed by the Federal Reserve Bank alongside small recessions. For example, around 2018, a spike in unemployment had an effect on housing price causing a dip down in rates (Guirguis and Trieste [2020]). There are also other global events not considered here that can play a significant role in interest rate fluctuations. However, despite these challenges, our models are still able to capture some general trends in mortgage rates through both the regression and the time series models. We next present our data and the methodologies used in Section 2, and Section 3 provides the overall summary and key insights gained from this project.

Chapter 2

Data analysis and modeling

In this section, we present our data sources and the models used in this study. We begin with a discussion of the data in 2.1 and two analysis strategies in 2.2. This is followed by an overview of our regression approaches in 2.3 and better fit model in 2.4. Finally, we describe the time series methodology in 2.5.

2.1 Data and resources

All our relevant data has been accessed from the Federal Reserve Economic Data (FRED)[1], a comprehensive repository of U.S. economic data, maintained by the Federal Reserve Bank of St. Louis. It provides historical time series on a wide range of economic indicators, including interest rates, unemployment, inflation, and housing statistics. Apart from the 30-year mortgage rate and the 10-year treasury yield data, we have also used the consumer price index (CPI), producer price index (PPI), unemployment rate, number of housing starts (abbreviated `HousingStarts` in the models), new home sales `NewHomeSales`, and consumer confidence rates. We have used a time period from January 1st 2010 to January 1st 2025 for our models.

Implementation of the project has been done using `R` (R Core Team [2025]). We have used the `fredr` (Boysel and Vaughan [2021]) package in `R`, which made it straightforward to retrieve and manipulate FRED data in an interactive manner. Some additional resources we have used here include the `tseries` (Trapletti and Hornik [2024]) and `forecast` (Hyndman et al. [2020]) packages for time series analysis and the `ggplot2` (Wickham [2016]) package for visualization.

2.2 Preliminary analysis

Our preliminary analysis with this data involved two strategies. First, an initial temporal aggregation and then a subsequent correlation analysis. An early challenge was that different variables were recorded at different time intervals. For example, mortgage rates were recorded weekly, whereas other variables such as the treasury yield, CPI, PPI, etc. were reported monthly by FRED. To align the datasets for analysis, each variable was aggregated to a monthly frequency, i.e., taking the average over each month's data. It should be noted that for implementation, this can be accomplished using `fredR`'s built-in temporal aggregation procedure (in this case, `frequency = "m"`).

[1] https://fred.stlouisfed.org

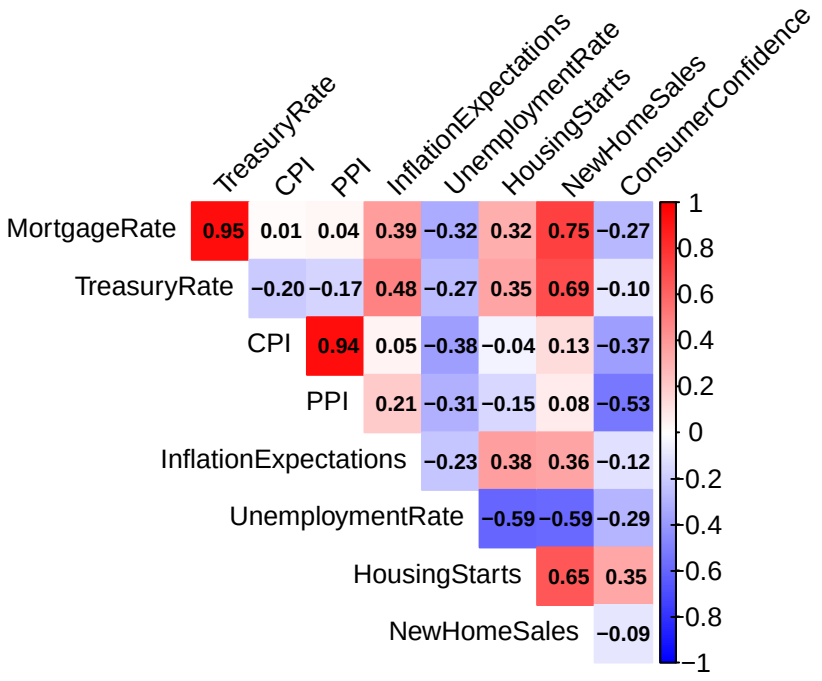

Figure 2.1: Correlation matrix of economic indicators

Our second exploratory analysis involved going through the correlations among all variables. The correlation plot (see 2.1) shows extremely strong (> 0.9) positive correlations between the mortgage rate and the treasury yield which is expected. Another strong correlation is between CPI and PPI. This is also expected as the consumer price index and producer price index tend to move together given their interconnected roles in the cost structure of goods and services. A negative correlation is observed between unemployment and new home sales. This can be explained by the fact that lower household income reduces purchasing power, thereby dampening demand for new homes.

Other notable patterns include a strong positive correlation between new home sales and mortgage rates. This may indicate that when new home sales increase, lenders raise mortgage rates in response to higher market demand. Also, high CPI tends to coincide with lower consumer confidence, reflecting the reduced purchasing power and cautious sentiment of households in times of rising prices.

2.3 Regression models

We began by creating the linear regression model that best fits this data for this project. We started with a linear regression model to predict our mortgage rate using all other variables as predictors in the

following:

$$\begin{aligned}\text{MortgageRate}_i = \beta_0 &+ \beta_1 \,\text{TreasuryRate}_i + \beta_2 \,\text{CPI}_i + \beta_3 \,\text{PPI}_i \\ &+ \beta_4 \,\text{InflationExpectations}_i + \beta_5 \,\text{UnemploymentRate}_i + \beta_6 \,\text{HousingStarts}_i \\ &+ \beta_7 \,\text{NewHomeSales}_i + \beta_8 \,\text{ConsumerConfidence}_i + \varepsilon_i \end{aligned} \tag{2.1}$$

Here β_1, β_2, etc., are all coefficients in the model. These coefficients are what determines how much the predictors affect the response variable which is the mortgage rate. For example, if coefficient β_2 is 3, then the CPI predictor will positively affect the mortgage rate by 3 units of CPI. ε_i is the error term and is the difference between the actual and predicted values. It accounts for the unobserved that affects the mortgage rate.

For model training, we used data up to 2023 and evaluated performance on the remaining data. This standard "train-test" procedure assesses the model's ability to generalize to unseen data. On the training set, the model achieves an R^2 of 0.9502 and an adjusted R^2 of 0.9485, indicating a strong fit. However, not all predictor variables were significant [see results in 3.1. The significant predictors are the 10-year Treasury yield, inflation expectations, `HousingStarts`, and consumer confidence, whereas intuitively related variables such as CPI, PPI, and the unemployment rate were deemed unnecessary by the model (each with $p > 0.05$). The performance on the test data was generally satisfactory and will be discussed and compared in the next subsection.

2.4 A second regression model

The rationale between our second model comes from the fact that one predictors, while intuitively relevant, may not improve the model due to correlations with other variables. This is known as the problem of "multicollinearity", where predictors will high correlation share the explanatory jobs, making individual inference problematic. For example, although CPI and PPI both have high p-values, their strong correlation suggest multicollinearity, so only CPI is removed from the model. After careful experimentation, we have also removed 'NewHomeSales' and unemployment rates as well as their effects were likely captured by remaining variables such as PPI and 'HousingStarts.' This process led to the following revised regression model:

$$\begin{aligned}\text{MortgageRate}_i = \gamma_0 &+ \gamma_1 \,\text{TreasuryRate}_i \\ &+ \gamma_2 \,\text{PPI}_i + \gamma_3 \,\text{InflationExpectations}_i \\ &+ \gamma_4 \,\text{HousingStarts}_i + \gamma_5 \,\text{ConsumerConfidence}_i + \epsilon_i \end{aligned} \tag{2.2}$$

Each predictor variable in Model 2.2 has a p-value less than 0.05, indicating that all variables are statistically significant (see results from 3.2). We also evaluated the models using additional criteria, including AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and prediction error on the test data measured by MSE (mean squared error). Definitions of these metrics are provided in Appendix B. 2.1 presents the comparison between the two models. Since lower values indicate better model performance, almost all the evaluations (except the adjusted R^2) confirm that the second regression model is superior.

Table 2.1: Comparison between two regression models

Metric	Full Model	Reduced Model
AIC	44.98	43.06
BIC	80.07	67.62
MSE	0.0587	0.0522
R^2	0.9502	0.9494
Adjusted R^2	0.9485	0.9483

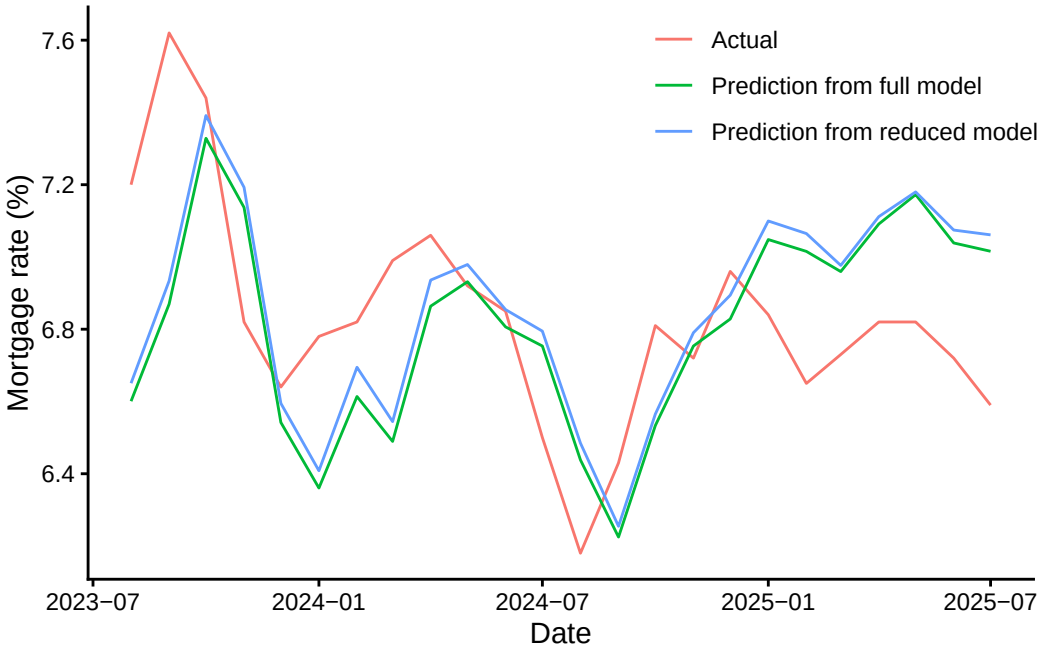

Figure 2.2: Regression models for mortgage rates – original vs predictions.

2.2 shows the original mortgage rate alongside our predictions from the two models. Overall, the predictions closely follow the observed values; however, the models were unable to capture all of the sharp spikes

and dips after the volatile COVID period. Notably, towards 2025, the predicted values are higher than the observed mortgage rates, which may reflect the recent dip in the mortgage rates. Predictions from both models are similar, though, as indicated earlier, the second model performs slightly better according to the evaluation metrics.

2.5 Time series

The final idea we have explored with our data is the use of a time series model. As noted earlier, predicting any temporal trend in mortgage rates alone is quite challenging. Instead of modeling the mortgage rate directly, we considered the spread, which is the difference between the mortgage rate and the treasury rate. Since these two series are highly correlated, the spread shows a more stable pattern and is easier to analyze. For organizational purposes, we present only the model and results below. The detailed theory and other details are provided in Appendix C.

2.5.1 Implementation

For this part, we are going to define the term *spread*, denoted as Y_T, as our response variable for the time series, i.e., the difference between the mortgage and the treasury rates. We first used a seasonal decomposition of the form: $Y_t = T_t + S_t + R_t$, which separates the time series into trend, seasonality, and a residual component. Then, we have used a one-time differencing of the form $(T_t - T_{t-1})$, which helped with our modeling. Following that, we have tried to fit an ARIMA model has been implemented, which is of the form:

$$T_t^{(D)} = \phi_1 T_{t-1}^{(D)} + \phi_2 T_{t-2}^{(D)} + \cdots + \phi_p T_{t-p}^{(D)} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}.$$

Note that, we usually need to difference until the series is stationary, however, for our case, we only differenced it once and allowed the seasonality to be modeled using a more flexible seasonal ARIMA. Due to the complex nature of the seasonal ARIMA model (which involves backshift operators and seasonal polynomials), its full equation is not shown in this paper.

The model has been implemented using the `R` package `forecast` (Hyndman et al. [2020]) and the routine `auto.arima`. The package by default allows seasonal ARIMA model implementation and also helps with selection of the optimal model by evaluating various (seasonal) ARIMA models and choosing the one with minimum AIC. In future work, a more manual route with full autocorrelation and partial autocorrelation plots may be adapted to help choose the best model.

2.5.2 Results

Table 2.2: ARIMA model results

Term	Estimate	Std. Error
MA1	0.783	0.046
SAR1	-0.562	0.126
SAR2	-0.587	0.093
SMA1	-0.162	0.133
INTERCEPT	0.006	0.001

Table 2.2 shows our results. The best ARIMA model has resulted in four temporal term which includes one moving average term (ϵ_{t-1}), one seasonal moving average (SAR_1) term, and two seasonal autoregressive (SAR_1 and SAR_2) terms. The resulting AIC of the model is -811.65 and BIC is -793.1. It should be noted that this model has automatically identified a seasonality of 12 months for the seasonal AR and MA terms, which is sensible since this is monthly data and we have already forced some seasonality by using a decomposition already.

2.5.3 Forecasting

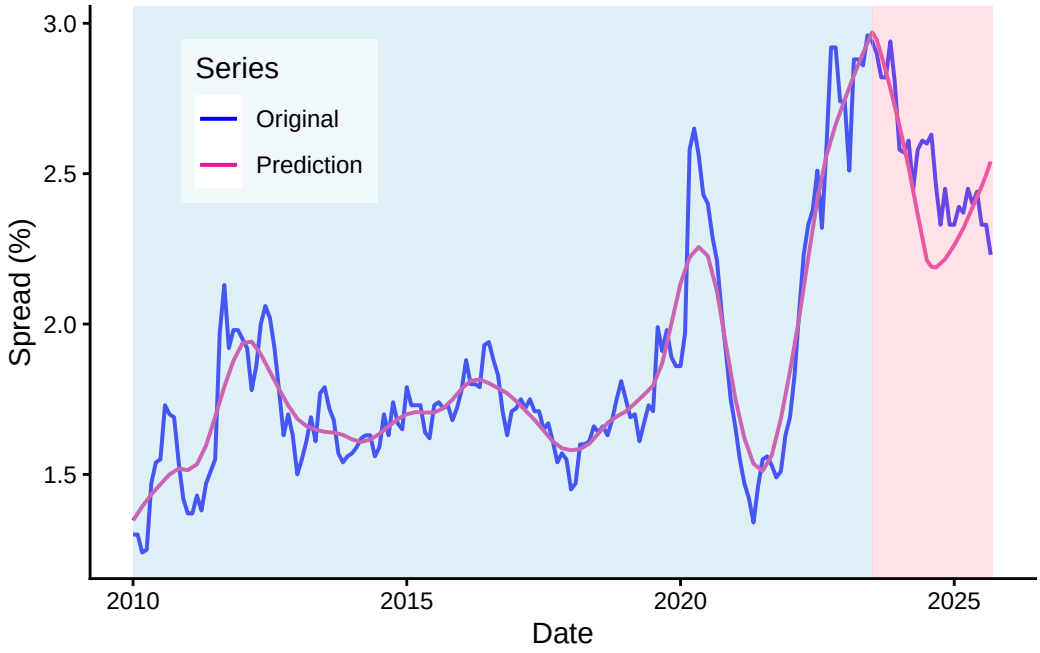

Figure 2.3: ARIMA modeling of the difference of mortgage and treasury rates. Blue region is our training data, whereas the pink region shows the test performance.

As described in Appendix C, forecasting refers to predicting future observations and is a central objective of time series analysis. We forecasted the spread series beyond July 2023, using all data up to that point for model training. 2.3 displays the resulting forecasts. Initially, the model performed reasonably well by capturing the general downward trend. However, more recently, it has failed to capture the market's behavior. While the actual spread has continued to decline over the past year, the model instead predicts a "U"-shaped upturn. This indicates that our model has not been sufficient in capturing the trend and further tuning will be necessary in future work.

Chapter 3

Conclusions

In this section, we conclude our study. In 3.1, we summarize our analyses in this project with all the key findings. 3.2 then outlines potential future directions for extending this work in future research.

3.1 Summary

One of the limitations of our project is that models with only a few predictors are unlikely to accurately predict real-life movements in the mortgage market or how rates will behave in the upcoming months. Much of the variation in mortgage rates is policy-driven, and changes are often influenced by multiple macroeconomic and microeconomic stressors as well as global events. However, the purpose of this project was to explore whether basic statistical methods could provide any insight into how the market might move when considered alongside treasury yields and a small set of economic indicators. The outputs from our models should therefore be viewed as exploratory rather than definitive. They help show the general direction of relationships and whether there is any predictive signal at all, but they should not be interpreted as precise forecasts of future mortgage rate behavior.

3.2 Future directions

In the future, this study could be broadened to further understand what moves the mortgage rate. One such way is to include more predictors of the general economy or from the housing market specifically. Research could be done on the predictors to see if they affect the mortgage rate in the short or long term. Another way the study could be furthered is with different nonlinear approaches. Since the current graphs show how unpredictable rates are just based on the indicators in the current study after COVID-19, regime-switching models or machine learning may show what the linear model may not have been able to. Another consideration is to look into shorter spans of time. Focusing on 3 to 5 year time periods to get a more specific interaction between mortgage rates and the predictors in the short term windows.

For brainstorming purposes, we also asked OpenAI's ChatGPT (OpenAI [2025]) for ideas on what else could be done with our project. Some of the suggestions were similar to what we had already considered. These included adding more predictors to the regression model, trying a nonlinear model (such as a random forest), and improved model validation techniques, e.g., by looking at smaller time windows. Overall, combining time-series ideas with more predictors and using stronger validation methods seems like a natural

direction for future work.

Bibliography

Boyarchenko, N., Fuster, A., and Lucca, D. O. (2019). Understanding mortgage spreads. *The Review of Financial Studies*, 32(10):3799–3850.

Boysel, S. and Vaughan, D. (2021). Package 'fredr'.

Goodman, L. S. and Mayer, C. (2018). Homeownership and the american dream. *Journal of Economic Perspectives*, 32(1):31–58.

Gordon, G. (2023). Mortgage spreads and the yield curve. *Richmond Fed Economic Brief*, 23(27).

Guirguis, H. S. and Trieste, J. (2020). Measuring the impact of monetary policy on mortgage rates. *Journal of Real Estate Research*, 42(2):285–313.

Hancock, D. and Passmore, W. (2012). The federal reserve's portfolio and its effects on mortgage markets.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., and Wang, E. (2020). Package 'forecast'. *Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf*.

Larson, W. D. and Martinez, A. B. (2025). House prices, debt burdens, and the heterogeneous effects of mortgage rate shocks. *Office of Financial Research Working Paper*, (25-02).

McCarthy, J. and Peach, R. W. (2002). Monetary policy transmission to residential investment. *Economic policy review*, 8(1).

McGibany, J. M. and Nourzad, F. (2004). Do lower mortgage rates mean higher housing prices? *Applied Economics*, 36(4):305–313.

Moench, E. and Soofi-Siavash, S. (2022). What moves treasury yields? *Journal of Financial Economics*, 146(3):1016–1043.

OpenAI (2025). Chatgpt (version 5.1). Large language model.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Trapletti, A. and Hornik, K. (2024). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-58.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zulauf, C. (2018). Interest rates in 30-year perspective: The case of us 10-year treasury rates. *farmdoc daily*, 8(115).

Appendix

A Regression model results

The coefficient significance results from our regression model 2.1 is as follows:

Table 3.1: Regression Coefficients from initial model

Term	Estimate	Std. Error	Statistic	P-Value
(Intercept)	2.501	0.532	4.699	0.000
TreasuryRate	1.042	0.033	31.549	0.000
CPI	0.001	0.003	0.218	0.828
PPI	0.005	0.004	1.434	0.153
InflationExpectations	-0.466	0.085	-5.479	0.000
UnemploymentRate	-0.022	0.018	-1.250	0.212
HousingStarts	0.000	0.000	3.836	0.000
NewHomeSales	0.000	0.000	1.032	0.303
ConsumerConfidence	-0.016	0.003	-5.662	0.000

Note the variables `HousingStarts` and `NewHomeSales` appear to have zeroed out values for estimate and standard error. These do have numeric values that are not written due to the three decimal place rounding in the chart and are not actually zero values. The same goes for `HousingStarts` in the new model. Also note the high p values for the intuitively important predictors such as PPI, CPI, etc. After removing the three predictors as explained in Section 2.4, following is the significance results for our new regression model:

Table 3.2: Regression Coefficients from second regression model

Term	Estimate	Std. Error	Statistic	P-Value
(Intercept)	1.880	0.258	7.284	0
TreasuryRate	1.088	0.022	49.596	0
PPI	0.008	0.001	9.282	0

Table 3.2: Regression Coefficients from second regression model

Term	Estimate	Std. Error	Statistic	P-Value
InflationExpectations	-0.544	0.055	-9.970	0
HousingStarts	0.000	0.000	9.157	0
ConsumerConfidence	-0.014	0.002	-8.696	0

B Metrics used

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = -2\ln(\hat{L}) + k\ln(n)$$

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

$$R^2_{\text{adj}} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

AIC : The akaike information criterion. It is a measure of goodness of fit and complexity. It is used to compare two models. The model with the lower number is better fit.

BIC : The bayesian information criterion is similar to the AIC, however it is more strict with complexity. The model with the lower number is better fit

R^2 : The amount of variance explained by the model.

R^2_{adj} : The adjusted R^2 adjusts R^2 for the number of predictors.

MSE : The mean squared error measures model accuracy using the differences between the observed and predicted values, averaged, and squared. A low number means the model is more fit to the data in the model.

C Time series details

Here we present the detailed discussion on how a time series is different from a regular data and a traditional regression modeling problem.

Definition of a time series

A time series is a sequence of data points that has been collected and sequenced in order by time [@bee2016seasonal]. Such data then is analyzed to identify historical patterns and trends, which can then be used to forecast future events. In this study, time is measured at regular intervals, i.e., at a gap of one month. As a result, we can denote the time index as $t = 1, 2, 3, \cdots$ (in units of months) and the observed data as $\{y_1, y_2, y_3, \ldots.\}$ See that compared to regular regression model, here we have an additional constraint that data occurred in a temporal ordered fashion. This makes the analysis fundamentally different from standard regression approaches.

Components of a time series

Some important components of a time series include *trends* and *seasonality* [@diggle2025time]. A trend indicates a long term pattern; for example, Google's stock prices has shown an overall increasing trend or pattern since its inception. Seasonality refers to patterns that repeat at fixed intervals, such as daily, monthly, or yearly cycles; for example, temperatures rise each summer and fall back again each winter. Since seasonality is a repeating behavior, we often tend to analyze the trend component separately by decomposing a time series as:

$$y_t = T_t + S_t + R_t,$$

where T_t and S_t are trend and seasonality component and R_t is the residual term. We proceed with analyzing the trend component in the subsequent part.

Stationary time series

Many time series analysis models [e.g., autoregressive models, moving average models [@box2015time] require assumptions of stationary. With time series, there are two primary forms of stationary, *strong stationary* and *weak stationary.* For a strongly stationary series, the probability distribution remains unchanged. For weakly stationary series, the requirements are that the mean and variance of the time series requires the same, but there is a covariance structure on the lag between the time series. This is what have used here in our models, i.e., a weakly stationary series after calculating the self-lags or differences of the series, i.e., $y_t - y_{t-1}$.

The augmented Dicky-Fuller test is often used to test for stationarity. The null hypothesis is H_0: the series is not stationary, against the alternative H_1: the series is stationary. A small p-value therefore indicates that the series is stationary.

ARIMA models for time series

ARIMA stands for auto-regressive integrated moving average models. This has two components: autoregression (AR) and moving average (MA). An autoregression model is a regression problem of a time series on itself in the form:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

where $\phi_1, \ldots, \phi_p$ are AR coefficients and ϵ_t is white noise. The moving average component models the series as a function of past error terms:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q},$$

where $\theta_1, \ldots, \theta_q$ are MA coefficients. The integrated part (I) refers to the number of self-differencing the series (i.e., $y_t^{(1)} = y_t - y_{t-1}; y_t^{(2)} = y_t^{(1)} - y_{t-1}^{(1)}; \ldots$) that it requires to achieve stationarity. All three together, the entire modeling statement takes the form:

$$y_t^{(D)} = \phi_1 y_{t-1}^{(D)} + \phi_2 y_{t-2}^{(D)} + \cdots + \phi_p y_{t-p}^{(D)} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}.$$

Seasonal ARIMA model

Sometimes, it also helps if we allow a more flexible ARIMA model instead of only capturing the seasonality through the differencing. A seasonal ARIMA (SARIMA) model is a further improvement from a traditional ARIMA model with additional terms of the time that takes care of any remaining seasonality in the data. This has been especially helpful in our modeling case due to the large volatility towards the end of the series within the COVID period. Especially for implementation purposes, this can be easily handled using the `forecast` package in `R` and the `auto.arima` routine. It also automatically guides the selection of the optimal ARIMA(p, d, q) model based on the minimum AIC (similar definition like in Section B) among various choices of (seasonal) ARIMA models with different parameterization.

An additional advantage is that the seasonal component often reduces the need to manually difference the series to achieve stationarity, as the S-part takes care of seasonal trends automatically. This has been helpful in this project since we directly applied the SARIMA model on once-differenced series for better prediction, where as we needed to use differencing two times to pass the ADF test. Therefore, the seasonal ARIMA idea has been immensely helpful in this case.

Forecasting for time series

Once a time series model (such as ARIMA, SARIMA) is fitted on the data, we can try to generate a prediction over the future. This in time series is called forecasting. Note that the forecasting or prediction is done using the differenced series, $\widehat{y}^{(D)}$, i.e., given data up to time T, the ARIMA model first predicts the future $\widehat{y}^{(D)}_{T+1}, \cdots, \widehat{y}^{(D)}_{T+h}$. It is then transformed into the original series by reversing the differencing operation as:

$$\begin{aligned}
\widehat{y}^{(D-1)}_{T+h} &= \widehat{y}^{(D)}_{T+h} + y^{(D-1)}_{T+h-1} \\
\widehat{y}^{(D-2)}_{T+h} &= \widehat{y}^{(D-1)}_{T+h} + y^{(D-2)}_{T+h-1} \\
&\cdots
\end{aligned}$$

Together with decomposition, differencing, seasonal ARIMA model, and forecasting, this has been our workflow for the time series model in Section 2.5.