Time Series Analysis: Final Project

Statistics 170 with Professor Juana Sanchez

Written by Kitu Komya
UID: 404-491-375

November 30, 2018

Contents

1 Introduction

1.1 Goal

The goal of this paper is to understand when certain forecasts are better and more accurate in predicting the future. Time series analysis is a method of modelling data in the form of time series by using its past data points. Several different models exist in time series forecasting, and this paper will examine three of the most commonly used ones: SARIMA + GARCH, Regression with Autocorrelated Errors, and VAR.

1.2 SARIMA + GARCH Model

The SARIMA model only uses historical information of the time series, and the SARIMA + GARCH is an extension of SARIMA by modelling the heteroskedastic variance of the residuals.

1.3 Regression with Autocorrelated Errors Model

Regression with Autocorrelated Errors utilizes a traditional regression model in forecasting a dependent variable by also using other exogenous variables, dummy variables for seasonality, or a polynomial trend for trend. To account for the autocorrelated errors, the residuals of such a model will be modelled by an ARIMA model, which is the approach used in GLS to approach regression with autocorrelated errors.

1.4 VAR Model

Finally, the VAR model is employed when the forecasted variable is both an independent and dependent variable, in which it depends on past values of itself along with past values of other variables which are also both dependent and independent.

1.5 "Consensus" Forecast

This paper will find the most optimal models for each of the three aforementioned models and average their forecasts to attain a "consensus" forecast. Consensus forecasts combine several, different forecasts, which benefit due to diversification gains and reduce the forecast errors of the individual forecasts.

1.6 Application

Understanding the mathematical and practical intuition behind these three techniques is vital in time series analyses. Such analyses arise across multiple domains, such as in economics and meteorology. Such areas that try predicting the future use multiple forecasts to obtain an average for times t + 1 and onward to ensure the most likely, consensus forecast. Similarly, this paper will conclude with a consensus forecast and compare its forecast with the actual, observed predictions to learn when certain forecasts are more accurate.

2 Describing the Data

2.1 Description

In order to answer our question posed in Section 1, economy data will be used. The variable to forecast is housing starts in the United States. Housing starts is considered to be a leading indicator of what might come next in the economy. Economists believe that if housing construction starts flourishing, it indicates that prosperity will come. Unemployment rate and Women's unemployment rate will also be considered. This data comes from FRED (Federal Reserve Economic Data, https://fred.stlouisfed.org). The data starts on January 1, 1959 and ends on August 1, 2018. Each data point for the three variables are collected monthly.

2.2 Summary of Data

var_name	r_name	description	train	test
HOUSTNSA	hs	Housing Starts is the total # of new, privately owned housing units (in thousands), measured monthly; not seasonally adjusted; this is variable to be forecasted	Jan 1, 1959 to Aug 1, 2017	Sept 1, 2017 to Aug 1, 2019
LNU04000002	uw	Women's Unemployment Rate (in percent) measured monthly; not seasonally adjusted & Jan 1, 1959 to Aug 1, 2017 & UNRATENSA & ur & Civilian Unemployment Rate (in percent) measured monthly; not seasonally adjusted	Jan 1, 1959 to Aug 1, 2017	
UNRATENSA	ur	Civilian Unemployment Rate (in percent) measured monthly; not seasonally adjusted	Jan 1, 1959 to Aug 1, 2017	

Table 1. Summary of the three variables used in the dataset.

3 Descriptive Analysis of the Data

3.1 Decomposition of the Time-Series' Variables

A descriptive analysis of the dataset will provide further insight into which forecasting methods will work best.

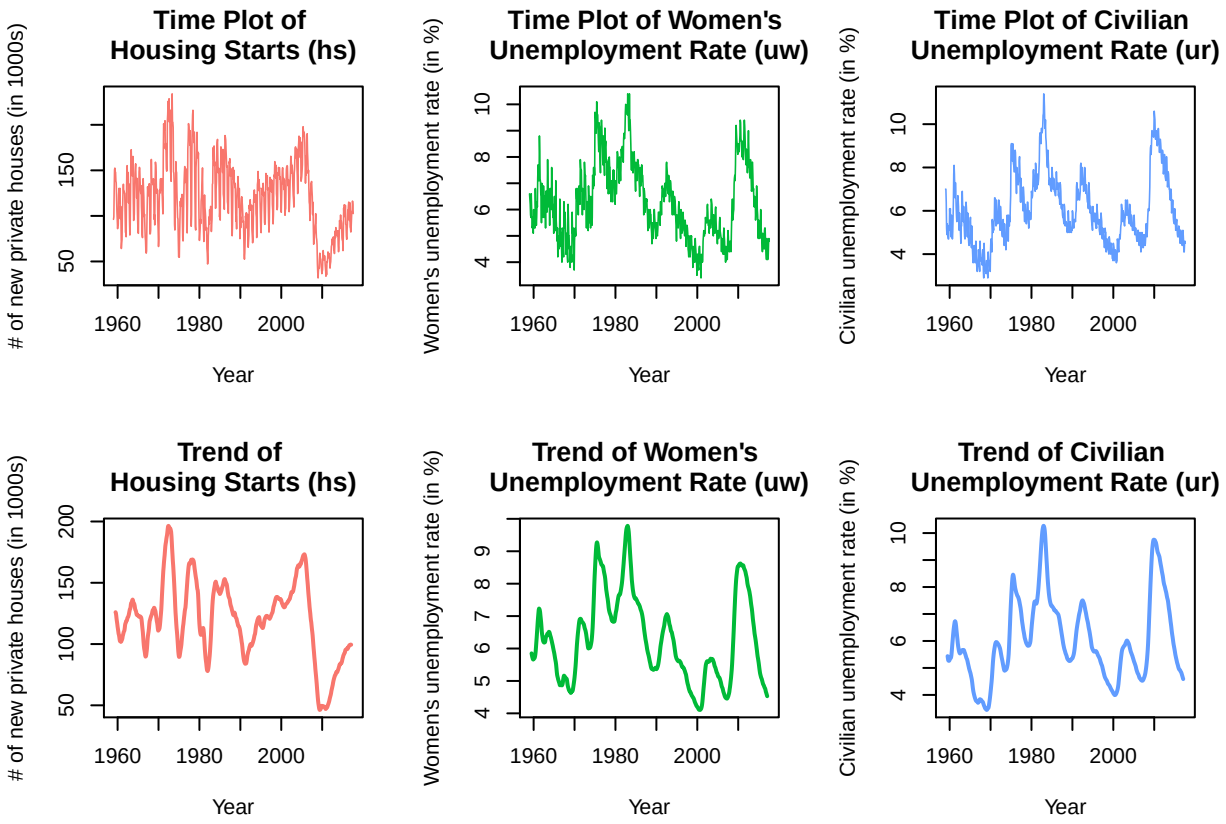


Figure 1. Time plots and trends of the three variables.

plot_type	hs	uw	ur
Time Plot	seasonal and cyclical component evident; large, constant variances; no general trend; not stationary; dip before 2010	seasonal and cyclical component evident; somewhat constant variances; no general trend; not stationary	seasonal and cyclical component evident; somewhat constant variances; no general trend; not stationary
Trend	cyclical component evident; large, somewhat varying variances; no general trend; dip before 2010	cyclical component evident; somewhat constant variances; sinusoidal trend	cyclical component evident; somewhat constant variances; sinusoidal trend

Table 2. Interpreting the timeplots and trends from Figure 1.

Figure 2. Seasonal boxplots and random parts of the three variables.

plot_type	hs	uw	ur
Seasonal Box Plot	seasonal component evident; arches to June then dips down; larger variances in summer months	relatively stationary, but a slight peak in June after which it decreases; larger variance in summer months	relatively stationary, but slightly decreases over the year; constant variance
Random Plot	mean stationary, but variance decreases with time	mean and variance stationary	mean and variance stationary

Table 3. Interpreting the seasonal boxplots and random parts from Figure 2.

3.2 Unit Root and Cointegration Tests

unit_root_hs	unit_root_uw	unit_root_ur	cointegration_test
0.2943	0.119	0.09635	< 0.01

Table 4. p-values of the unit root tests for each variable as well as for the cointegration test among all the variables.

Based off Table 4, since the p-values for all of the variables are larger than 0.05 for the unit tests, we fail to reject the null hypothesis of the Augmented Dickey-Fuller Test. Thus, all three variables are random walks because at least one of their roots is equal to 1, signifying a random walk process. Moreover, because in Phillips-Ouliaris test for cointegration the p-value is less than 0.05, we reject the null hypothesis. Thus, the three random walks are cointegrated. This means that the three variables are related by a stationary linear combination and share an underlying stochastic trend.

3.3 Volatility Check for GARCH Model

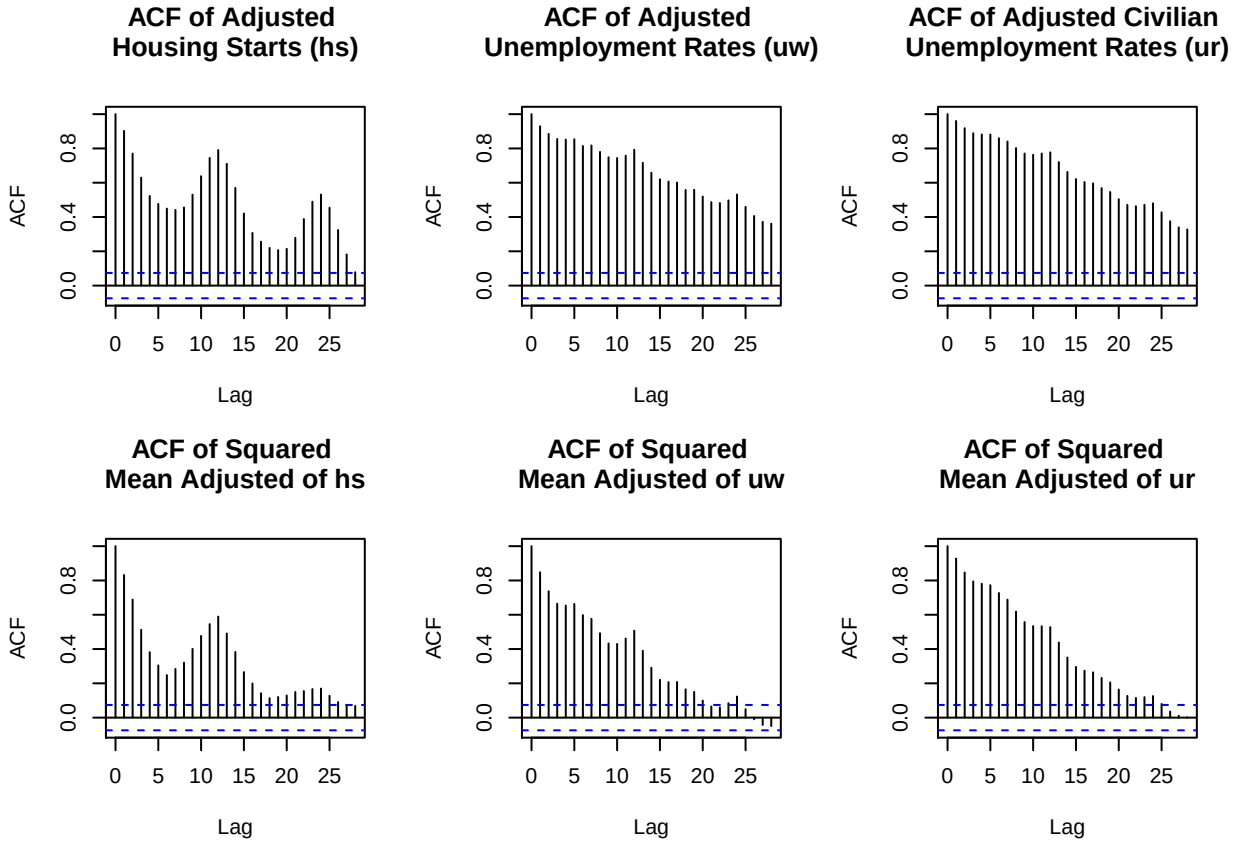


Figure 3. ACFs of adjusted variables and of squared mean adjusted variables.

From Figure 3, there is no suggestion of volatility in our data. Although the ACFs of the squared mean adjusted of the time series have significant lags, the ACFs of our variables in the time series are not white noise. Since both conditions have not been met, no volatility, or random periods of increased variance, exists in our time series. This conclusion is consistent from our interpretations of the time plots and trends from Table 2.

3.4 Dependencies of Variables Endogenously

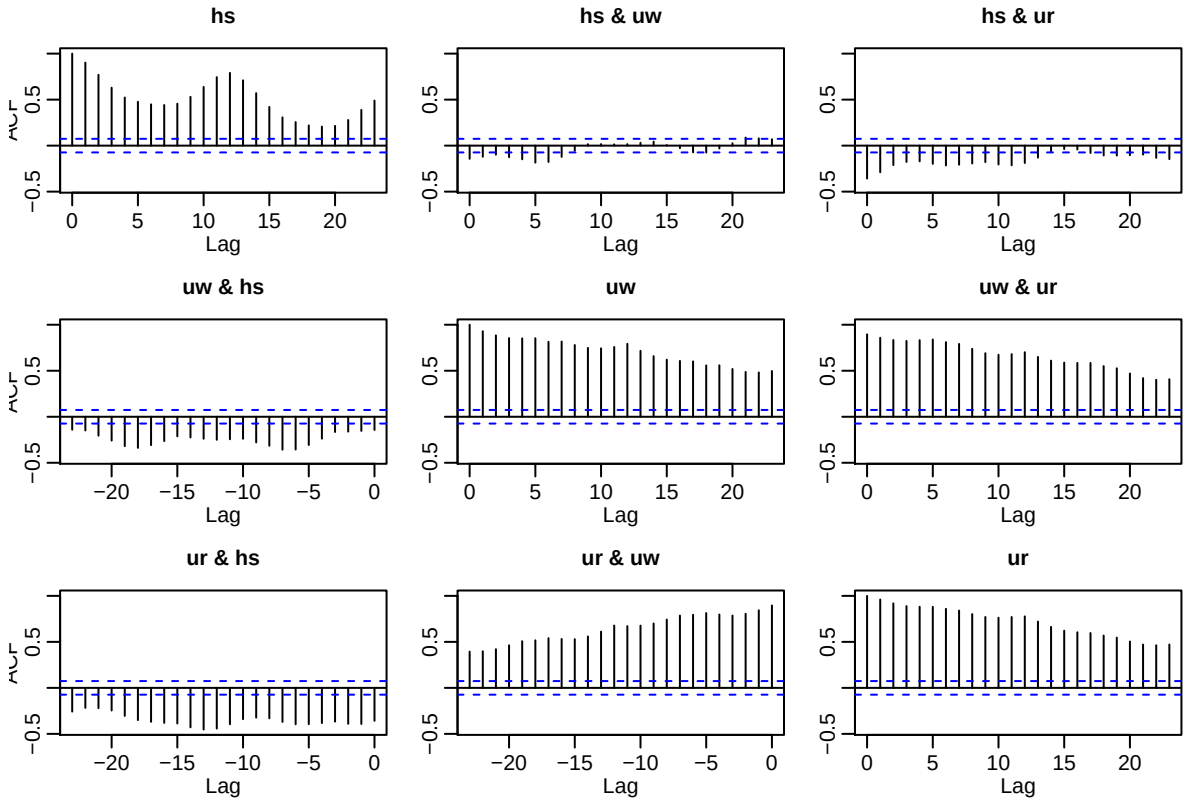


Figure 4. ACFs and CCFs of the three variables.

From Figure 4, we see that because none of the three variables have stationary ACFs, we will need to eventually difference the time series. Moreover, because of the significant autocorrelations in the CCFs, there are dependencies of the variables among each other, making them endogenous and dependent to each other.

4 Modelling SARIMA + GARCH

4.1 Differencing the Time-Series

In order to model SARIMA, we must model using a stationary time-series. As evidenced from Figure 3, the time-series is non-stationary as there are several significant auto-correlations. Thus, we will difference our data to achieve stationarity and to identify a model.

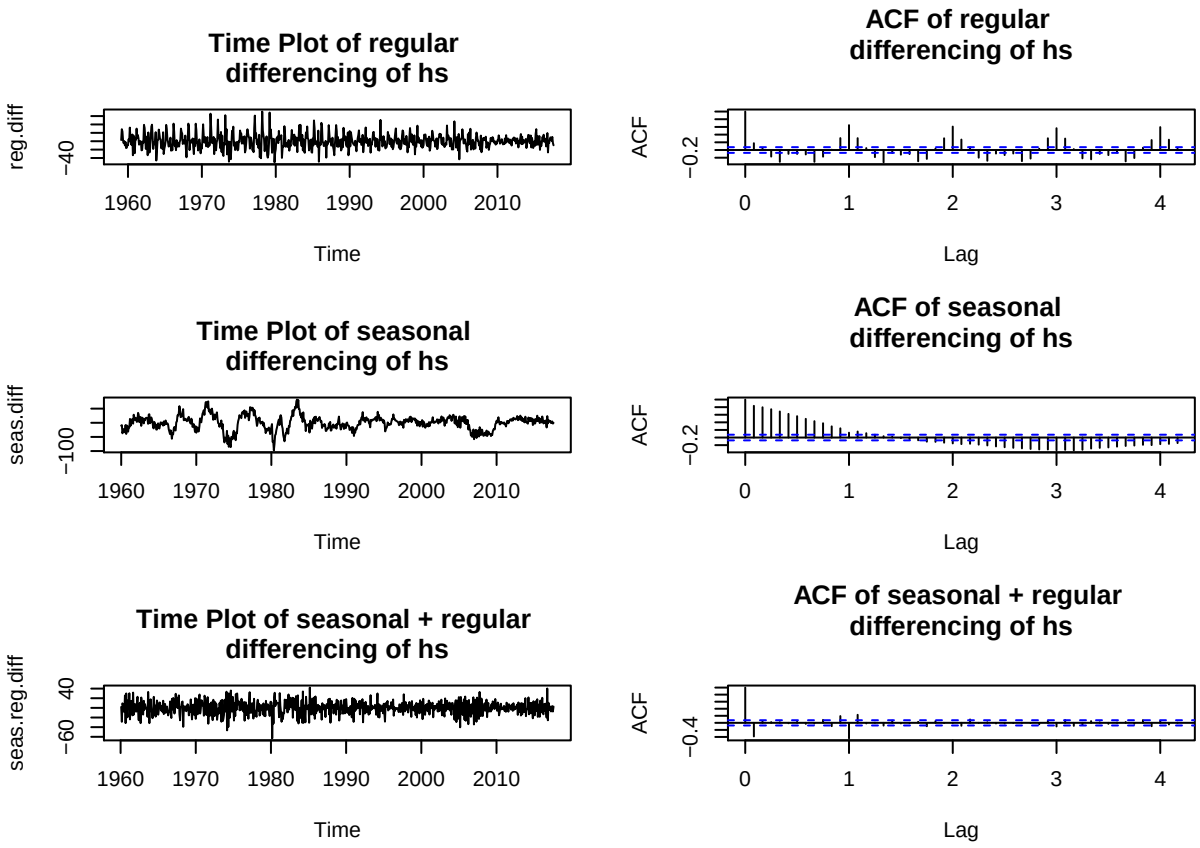


Figure 5. Time plots and ACFs of three differenced time-series.

differencing_type	time_plot	ACF
Regular	relatively mean stationary but not variance stationary	not white noise; seasonal/cyclical trend apparent; many significant autocorrelations
Seasonal	neither mean nor variance stationary	not white noise; many significant autocorrelations
Seasonal + Regular	mean and variance stationary	relatively white noise; few significant autocorrelations due to chance

Table 5. Interpretation of time plots and ACFs of three differenced time-series from Figure 5.

Table 5 suggests that the best difference which leads to stationary data is the seasonal + regular differencing. No pre-transformation is needed because the variance does not increase with time of the original time-series, as seen in Figure 3.

4.2 Identification of SARIMA Model

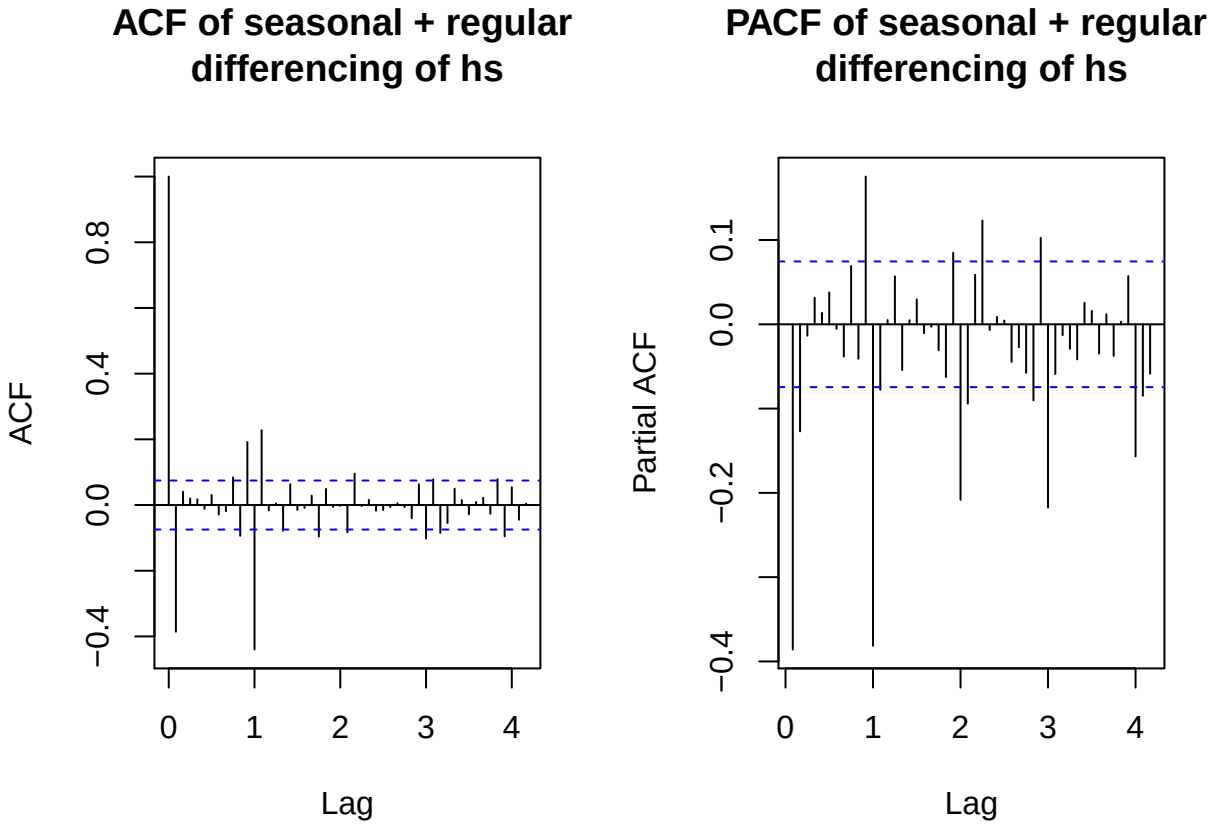


Figure 6. ACF and PACF of seasonal + regular differenced time-series.

Now that we have made our time-series stationary, we may identify a SARIMA model from its ACF and PACF. As evidenced by Figure 6, we will identify a $\mathsf{SARIMA}(0,1,1)(1,1,1)_{12}$ model.

Regular part: We will justify this model selection by first explaining the ARIMA(0, 1, 1) part. When looking at the ACF plot, there is a sharp, significant autocorrelation at lag 1, indicating an MA(1) process. From the PACF plot, we see that the autocorrelations die quickly over many lags, which is typical of MA behaviors for the regular part of SARIMA. No AR structure is identified.

Seasonal part: Moreover, in looking at the seasonal part of the SARIMA model, we notice significant autocorrelated spikes at the seasonal lags in both the ACF and the PACF, indicating both AR(1) and MA(1) in the seasonality.

4.3 Fit SARIMA Model

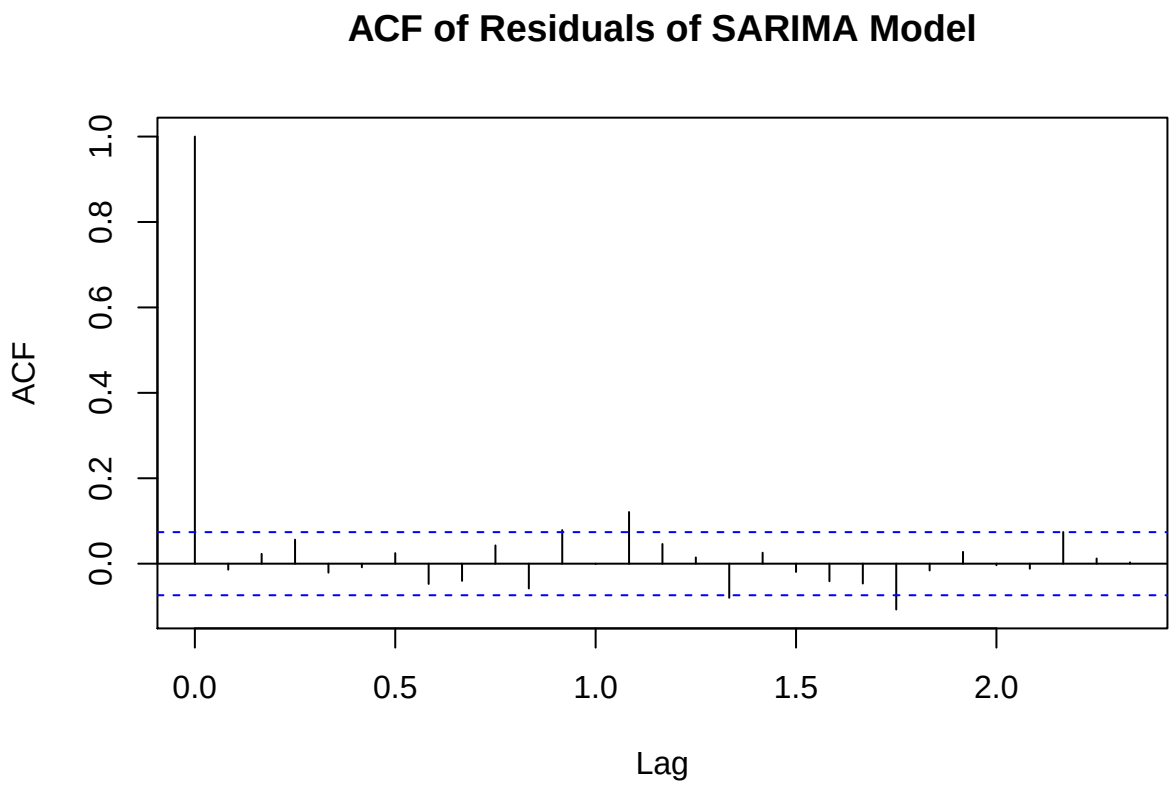


Figure 7. ACF of residuals of SARIMA model.

The plot from Figure 7 shows that the model residuals approximately demonstrate white noise, since the only significant autocorrelations are due to chance. Using the Ljung-Box test in conjunction, we find that although we reject the null hypothesis, the p-value is 0.02123, and the test in itself requires a lot of power and assumptions. For our purposes, visually inspecting the ACF demonstrates that our model is ready to be forecasted.

4.4 Volatility Check of Residuals for GARCH Model

Since the residuals model white noise, it's worth looking at the volatility of the residuals to determine whether fitting a SARIMA + GARCH is required.

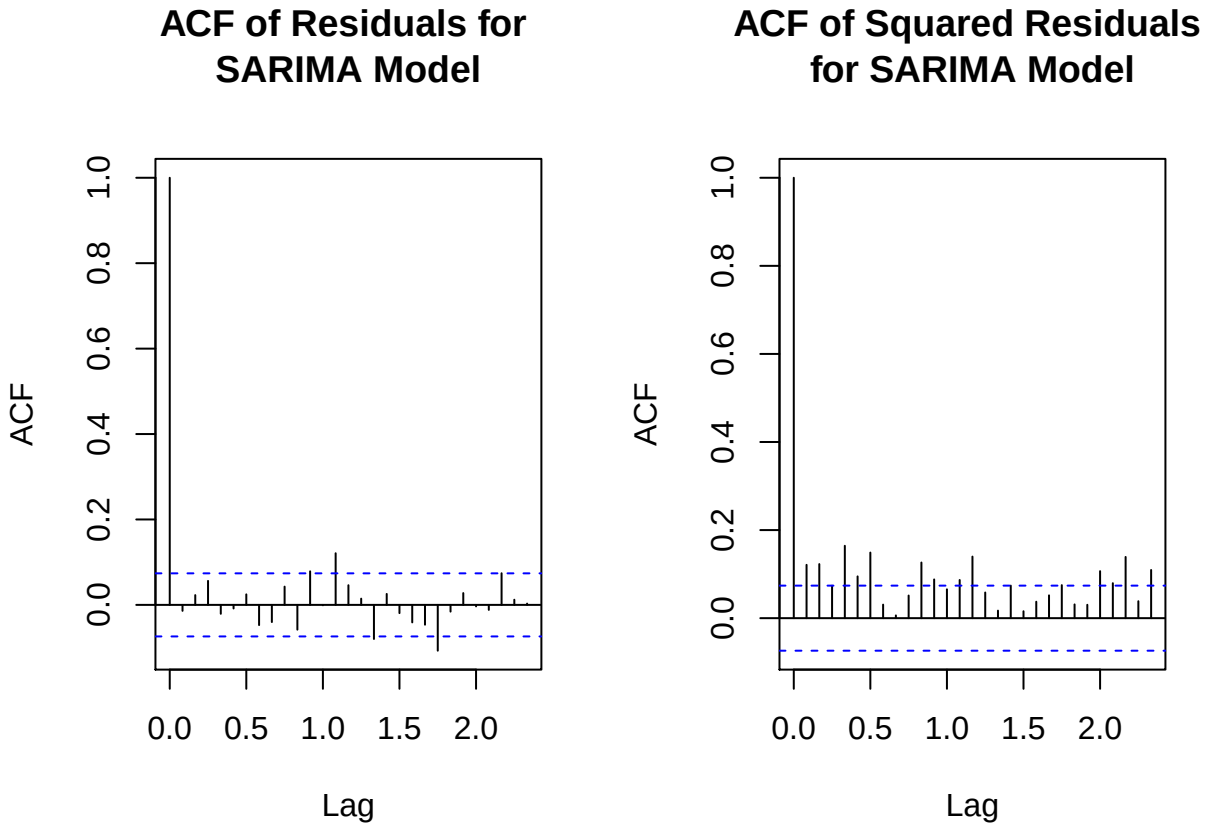


Figure 8. Volatility checking of residuals of SARIMA model.

The residuals of the ACF in the SARIMA model model white noise with significant autocorrelations being due to chance, as mentioned earlier. Moreover, the squared residuals of the ACF of the SARIMA model has many significant autocorrelations, suggesting that the residuals should be modeled by a GARCH to treat the varying variance.

4.5 Fit SARIMA + GARCH Model

GARCH_model	AIC
GARCH(0, 1)	5294.433
GARCH(1, 0)	5302.761
GARCH(1, 1)	5258.687
GARCH(0, 2)	5285.066

Table 6. AIC values for four common GARCH models fit onto SARIMA.

From Table 6 we see that the SARIMA + GARCH model with the lowest AIC and thus most optimal is GARCH(1, 1). Let's visualize what the ACF plots of the residuals and the squared residuals look like now.

4.6 Volatility Re-Check of SARIMA + GARCH Model

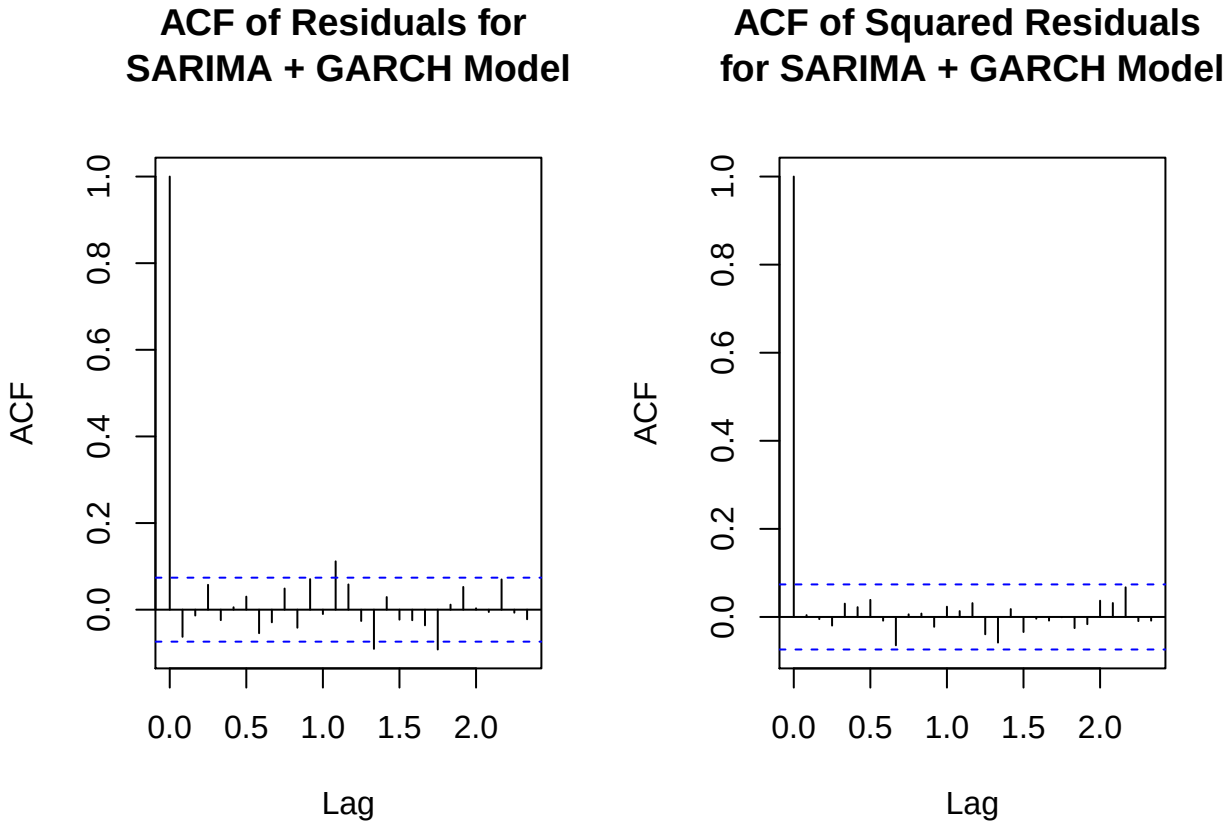


Figure 9. ACFs of residuals and of squared residuals for SARIMA + GARCH model.

Wonderful. Figure 9 demonstrates that after adding a GARCH component onto the residuals after a SARIMA model, the residuals and the squared residuals are white noise. With p-values of 0.03062 and 0.9397 respectively for the residuals and squared residuals for the Ljung-Box test demonstrate reasonable white noise.

4.7 SARIMA Model Equation in Polynomial Form

The equation of our $\mathsf{SARIMA}(0,1,1)(1,1,1)_{12}$ model in polynomial form will be detailled below:

$(1-\alpha_{12}*\mathsf{B}^{12})(1-\mathsf{B}^{12})(1-\mathsf{B})\mathsf{x_t} = (1-\beta_1*\mathsf{B})(1-\beta_{12}\mathsf{B}^{12})\mathsf{w_t}$

$(1-(0.1405)*B^{12})(1-B^{12})(1-B)\mathsf{x_t} = (1-(-0.3363)*\mathsf{B})(1-(-0.8885)\mathsf{B}^{12})\mathsf{w_t}$

$(-1.686*B^{25})\mathsf{x_t} + (1.686*\mathsf{B}^{24})\mathsf{x_t} + (2.686*\mathsf{B}^{13})\mathsf{x_t} - (2.686*\mathsf{B}^{12})\mathsf{x_t} - \mathsf{B}\mathsf{x_t} + \mathsf{x_t} = (3.58563*\mathsf{B}^{13})\mathsf{w_t} + (10.662*\mathsf{B}^{12})\mathsf{w_t} + (0.3363*\mathsf{B})\mathsf{w_t} + \mathsf{w_t}$

4.8 Forecasting SARIMA Model

Because SARIMA + GARCH provides the same coefficient estimates/forecasts as SARIMA, we will simply forecast for SARIMA. SARIMA + GARCH only reduces the variance in the coefficient estimates/forecasts.

time_ahead	sarima_prediction
1	101.69519
2	104.44352
3	88.12797
4	80.42764
5	77.56587
6	81.02832
7	96.02236
8	106.16091
9	108.74752
10	112.49573
11	110.25166
12	103.62926

Table 7. Forecasts of the next 12 data points for SARIMA or SARIMA + GARCH model.

4.9 Forecasted Plot of SARIMA Model

Below is the plot of the forecasts of the SARIMA or SARIMA + GARCH model.

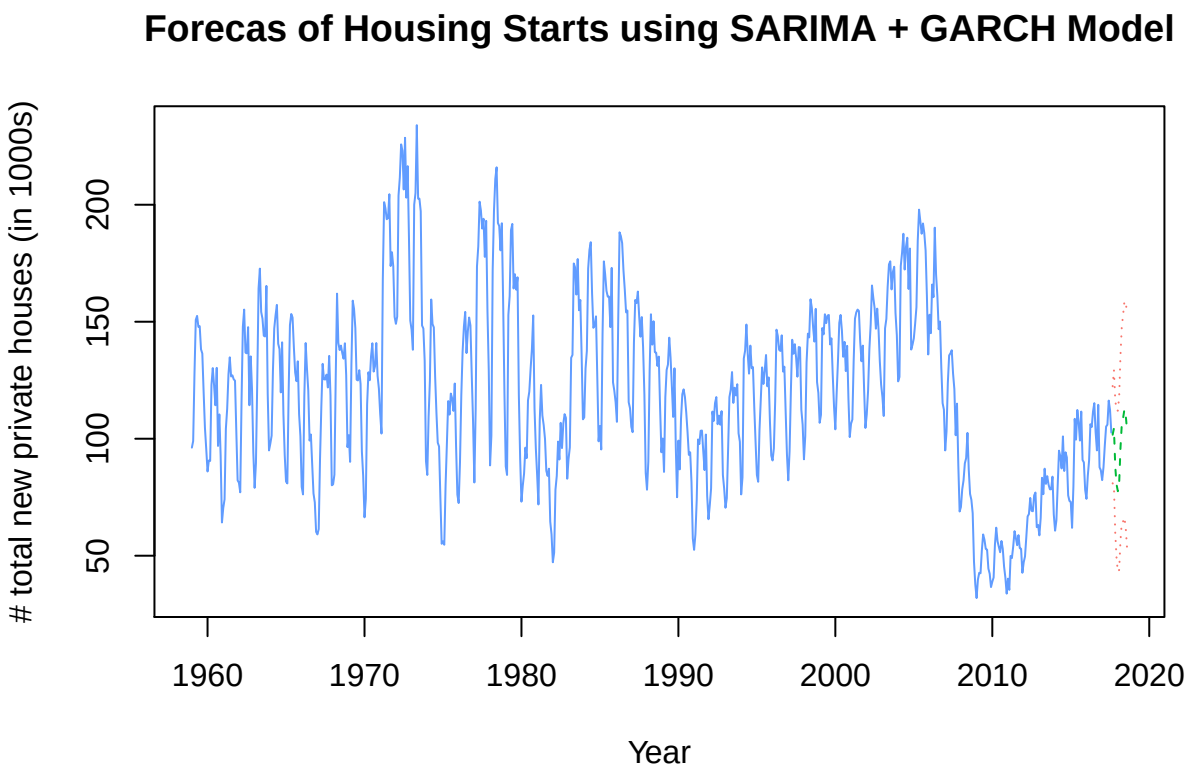


Figure 10. The forecasted plot of the SARIMA model.

Figure 10 shows the forecast of the SARIMA model. Note that although we also used SARIMA + GARCH model, SARIMA + GARCH model does not change the coefficeints of the estimates/forecasts; it only reduces the standard error. However, as of now, R does not have capability to forecast GARCH with seasonality time-series, rendering any outputted standard error absolutely useless since we are modeling with seasonality in our SARIMA model. An interesting feature of the plot is how large the confidence interval bands are; due to the high and inconsistent variance throughout the time plot, such a large range is required.

4.10 RMSE of SARIMA Model

model	RMSE
SARIMA	8.958358

Table 8. RMSE value of SARIMA model's predictions against test data set.

Calculating the RMSE will tell us how accurate our prediction is. We use the test set data as the observed data and our forecasted data as predicted. Table 8 tells us that we obtain an RMSE value of 8.958358. Although this value is not meaningful in itself despite being a small number, we will compare this value to see relatively how accurate the SARIMA model is.

5 Modelling Regression with Autocorrelated Errors

5.1 Exploring Autocorrelated Features in Housing Starts

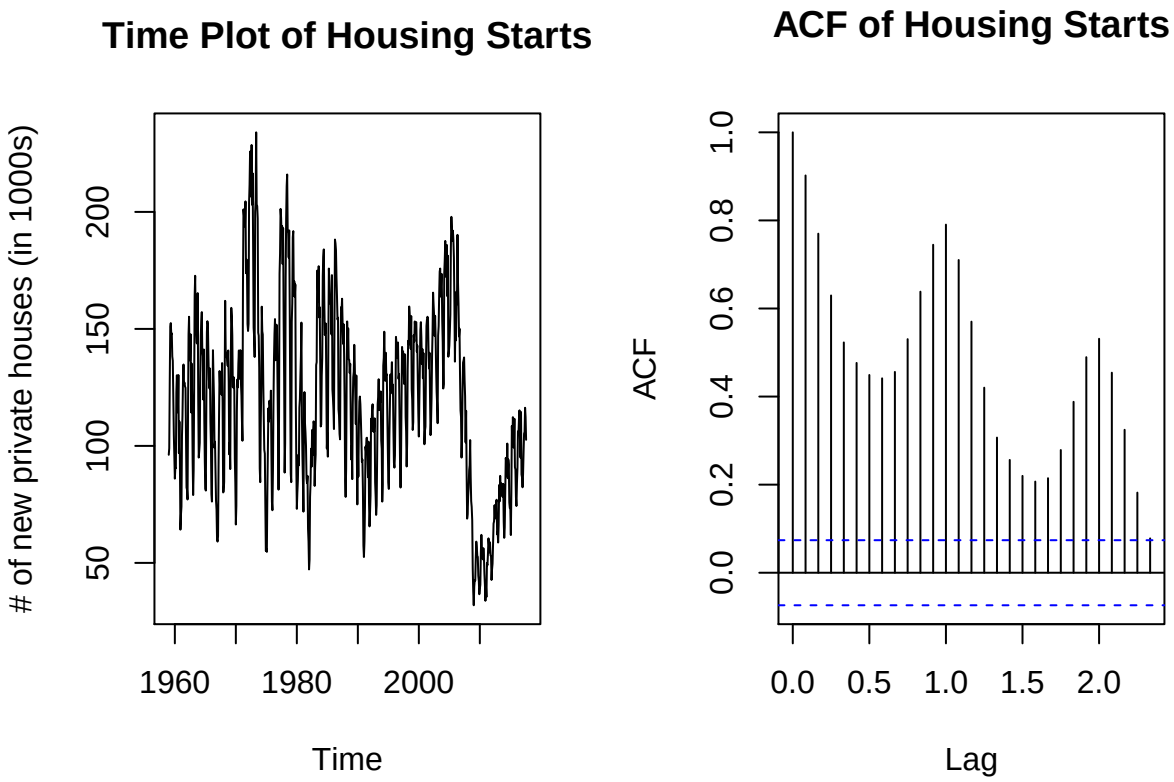


Figure 11. Time plot and ACF of the housing variable.

We revisit the time plot in Figure 11 in order to model based on its autocorrelated errors. We have already explored the main features of these tWo graphs in Sections 3.1 and 3.3, but as a quick refresher, the data is not stationary as there are many significant autocorrelations in the early lags. Seasonality is present and trend is uncertain.

5.2 Fit Classical Regression Model

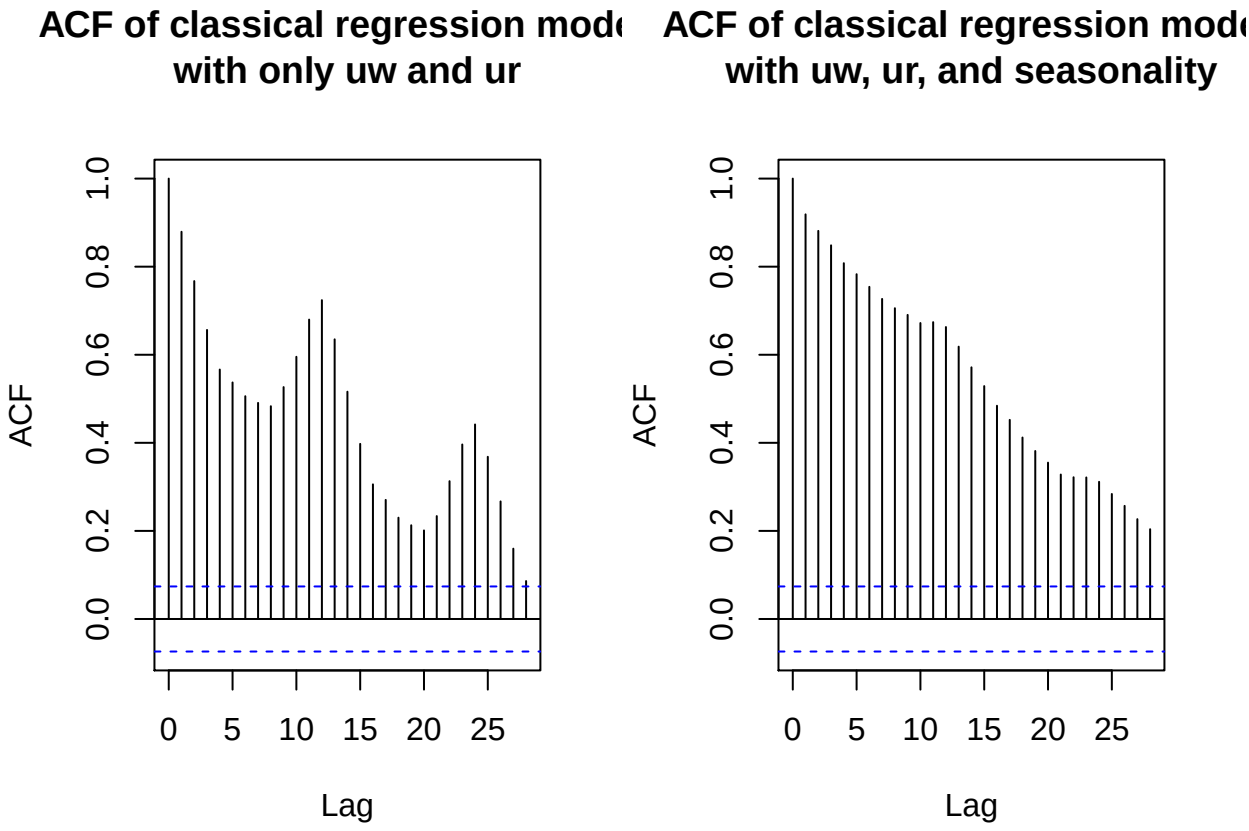


Figure 12. ACF plots of two candidate classical regression models.

Figure 12 shows two classical regression models: the one on the left only uses the other two variables (uw, ur), while the right model also includes seasonality. Comparing these two models to find the optimal one is necessary in order to move forward. The reason why there is a second model that captures seasonality is that from Figure 11, there is clear seasonality in the time plot that needs to be modelled. Trend is not included because there is no clear trend for the timeplot, so adding it does not make intuitive sense. Both of the ACFs are clearly nonstationary due to their many significant autocorrelations. However, for multiple time-series analyses, having stationary variables is not a prerequisite. Thus, from these two models, the right model looks better because it follows an AR structure which is easier to model. Disregarding the nonstationarity of it, the right plot identifies as an AR(1) model. Thus, we will proceed to correct the residuals by modelling an AR(1) model onto the classical regression model that incorporates uw, ur, and seasonality.

5.3 Fit AR(1) Model onto Classical Regression Model

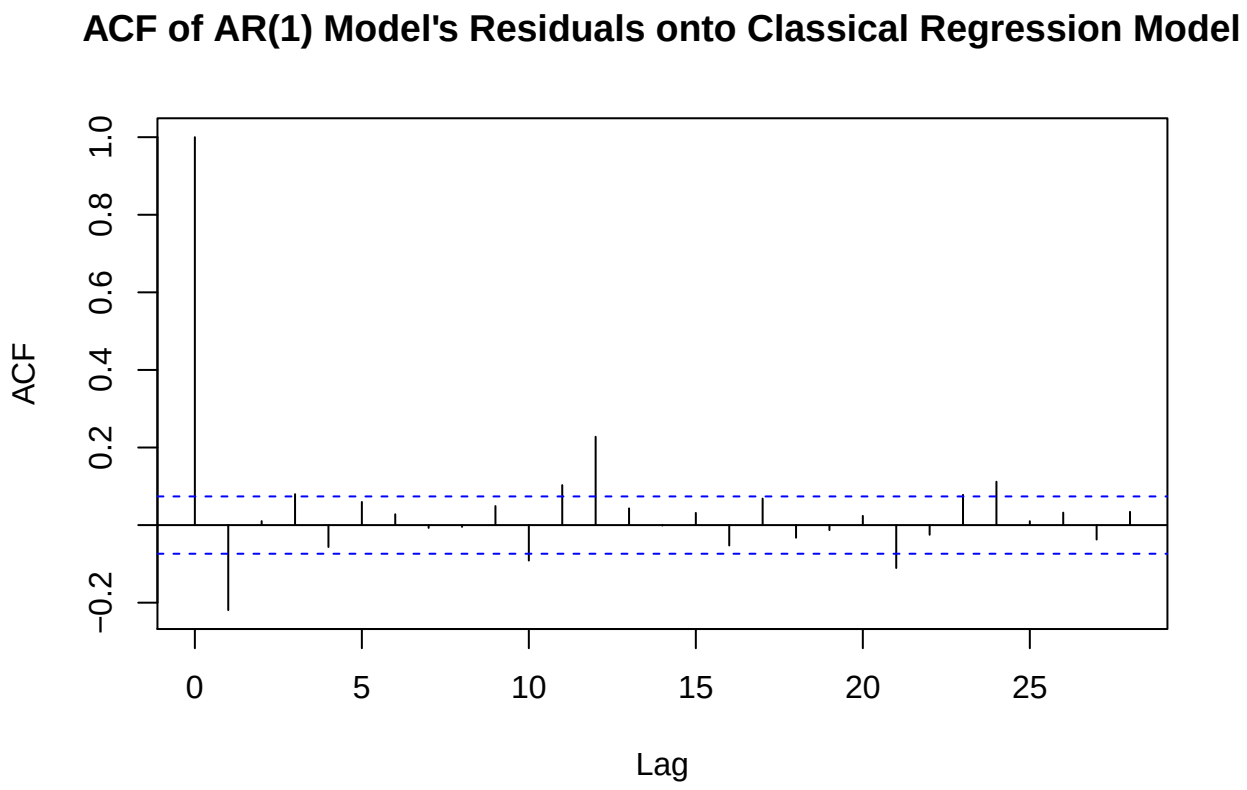


Figure 13. ACF plot of AR(1) model's residuals onto classical regression model.

Figure 13 showcases the residuals obtained after fitting an AR(1) onto the classical regression model. Applying an AR(1) model will clarify the true structure of the data so that we can use the coefficient estimate of the AR(1) into a GLS model to explain other autocorrelations that is not explained by the other variables. Thus, this is not our final model since it hasn't yet taken into account that correlation.

5.4 Fit GLS using AR(1) Model Structure

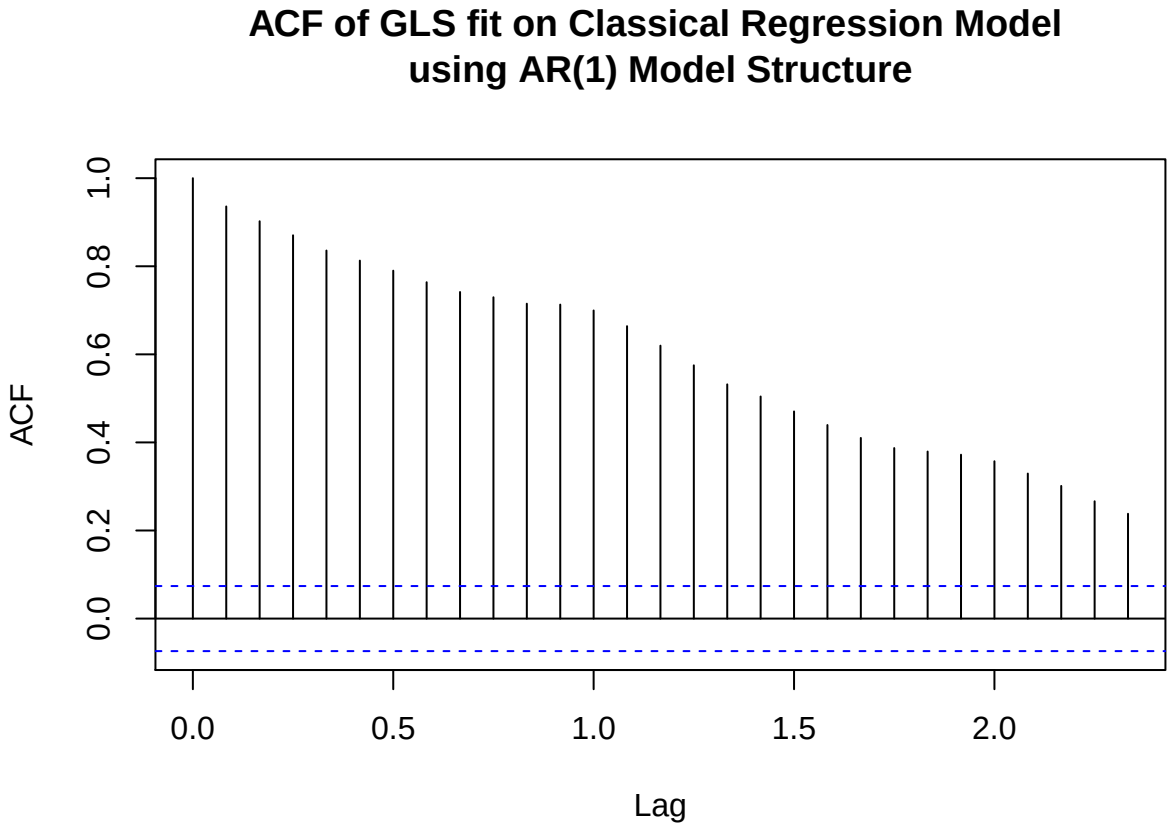


Figure 14. ACF of GLS fit on classical regression model using AR(1) model structure.

The ACF from Figure 14 is not the most beautiful since it is not white noise, but it definitely highlights an AR structure as seen through its significant autocorrelations. Other models tried could not remove the structure of this pattern in the residuals, so this is the best GLS fit for this model. GLS fit a model on the original time-series by using the structure of the AR(1) model. Other AR models do not do a better job of fitting this data, and thus we choose the simplest model, which is this.

5.5 GLS Model Equation

The final GLS model obtained is as follows:

$$hs = 114.72 + 1.66 * uw - 5.89 * ur + 3.08 * Feb + 31.39 * March + 45.05 * April + 49.47 * May + 50.78 * June + 44.24 * July + 40.86 * August + 33.46 * September + 37.37 * October + 16.92 * November + 1.48 * December$$

where the month variables are dummy variables. Interestingly enough, in comparison to the first month, January, each month has a higher coefficient estimate, suggesting that winter months have lower housing starts, while the higher coefficient estimates, which are found in the summer months, have higher power. This conclusion is consistent with the analyses done on the seasonal boxplots in Table 3. Also interestingly, all variables in this model are significant (with a p-value less than 0.05) excpet for uw and December. This leads us to believe that most of these variables are truly useful in the model.

5.6 Forecasting GLS Model

time_ahead	gls_prediction
1	84.52104
2	87.26864
3	118.77163
4	138.82954
5	144.83975
6	144.71710
7	138.51507
8	136.65931
9	129.34525
10	133.42196
11	109.93937
12	94.68026

Table 9. Future forecasts using the GLS model.

5.7 Forecasted Plot of GLS Model

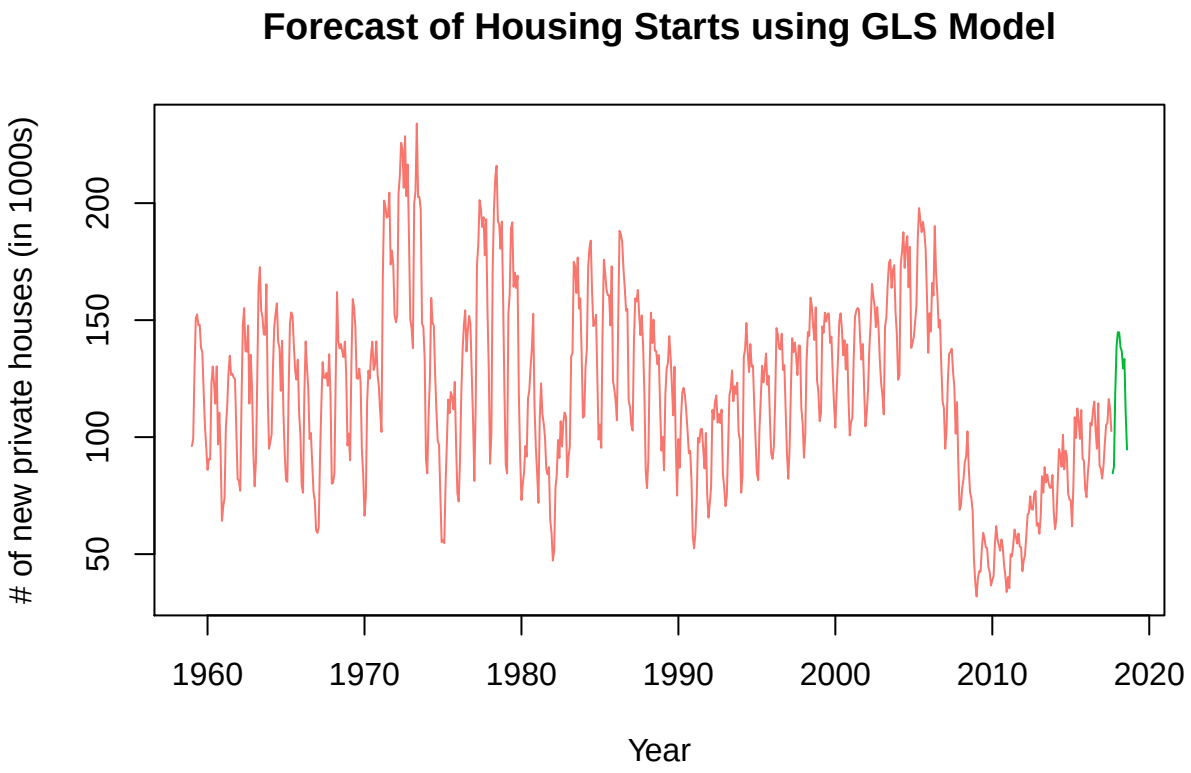


Figure 15. Forecasted plot of GLS Model.

The green illustrates the forecasted data points of the GLS model, while the red are the training data. The forecasted data points seem to have higher variability than the cycles before it. Unfortunatley, finding confidence interval bands for GLS models is not yet doable in R. Regardless, the forecast seems to be a good fit because it follows the general trend of the time plot.

5.8 RMSE of GLS Model

model	RMSE
GLS	32.47106

Table 10. RMSE of the GLS model.

Compared to the RMSE of the SARIMA model, Table 10 indicates that the RMSE of the GLS model is higher, demonstrating that it is a weaker fit than the SARIMA model. This is not too surprising, given that finding the most optimal GLS model was nearly difficult when all of the model's residuals had some sort of AR(1) processes, completely devoid of any white noise pattern.

6 Modelling VAR

6.1 Identication of VAR Model

We have already analyzed the time plots of the three variables in Table 2, but as a refresher, the three follow similar seasonal trends, leading us to believe that they are cointegrated. Table 5 proved to us to use the seasonal + regular differencing because it is the most stationary.

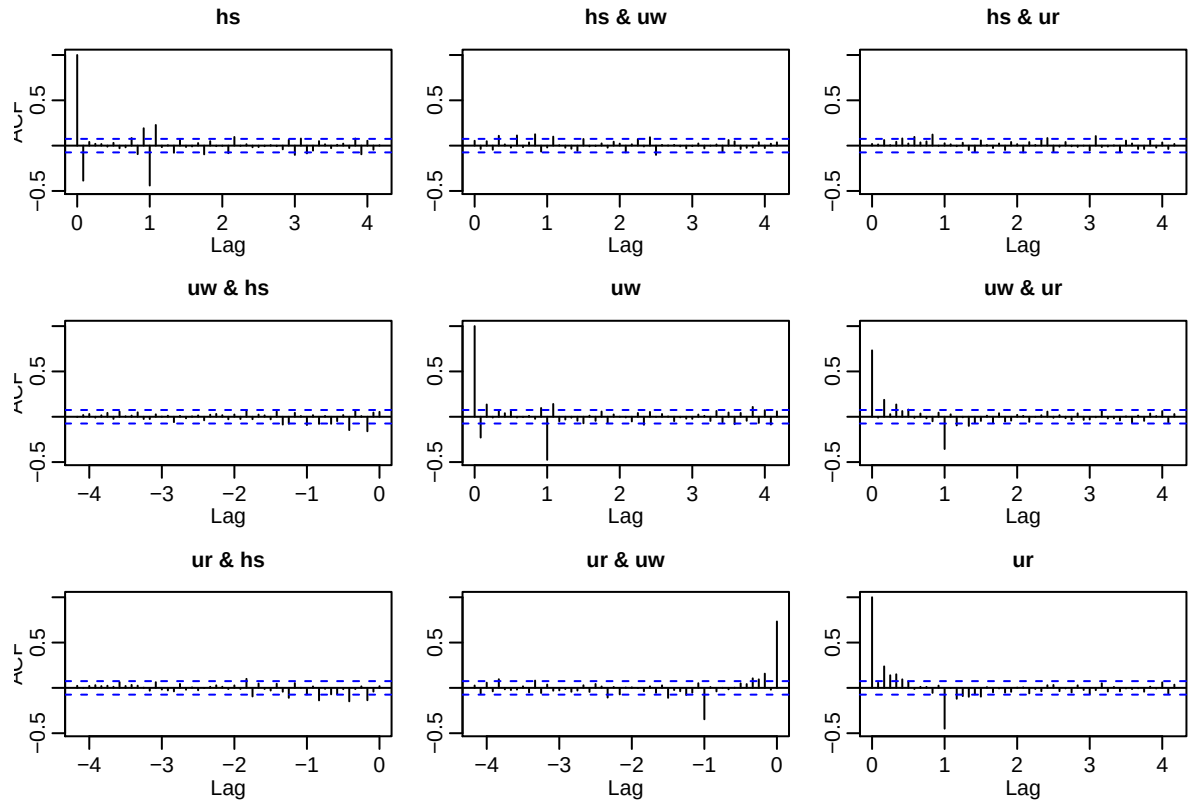


Figure 16. The ACFs and CCFs of the seasonal + regular differenced time-series.

Figure 16 shows that there exist some dependency among the endogenous variables. The variables are also dependent on themselves since they all damp down in a sinusoidal fashion. Identification of a model will be done by analyzing the lags that are significant for each plot.

input	response	lags
hs	hs	1, 2
hs	uw	4
hs	ur	none
uw	hs	2
uw	uw	1, 2
uw	ur	1
ur	hs	2
ur	uw	2, 3
ur	ur	1, 2

Table 11. Shows which variables are involved with each other, as inspected from their ACFs and CCFs.

Based on Table 11, although there is one lag that goes until lag 3 and one that goes until lag 4, I will choose VAR(2) because most of the lags are only until t-1 and t-2.

6.2 Fit VAR(2) Model

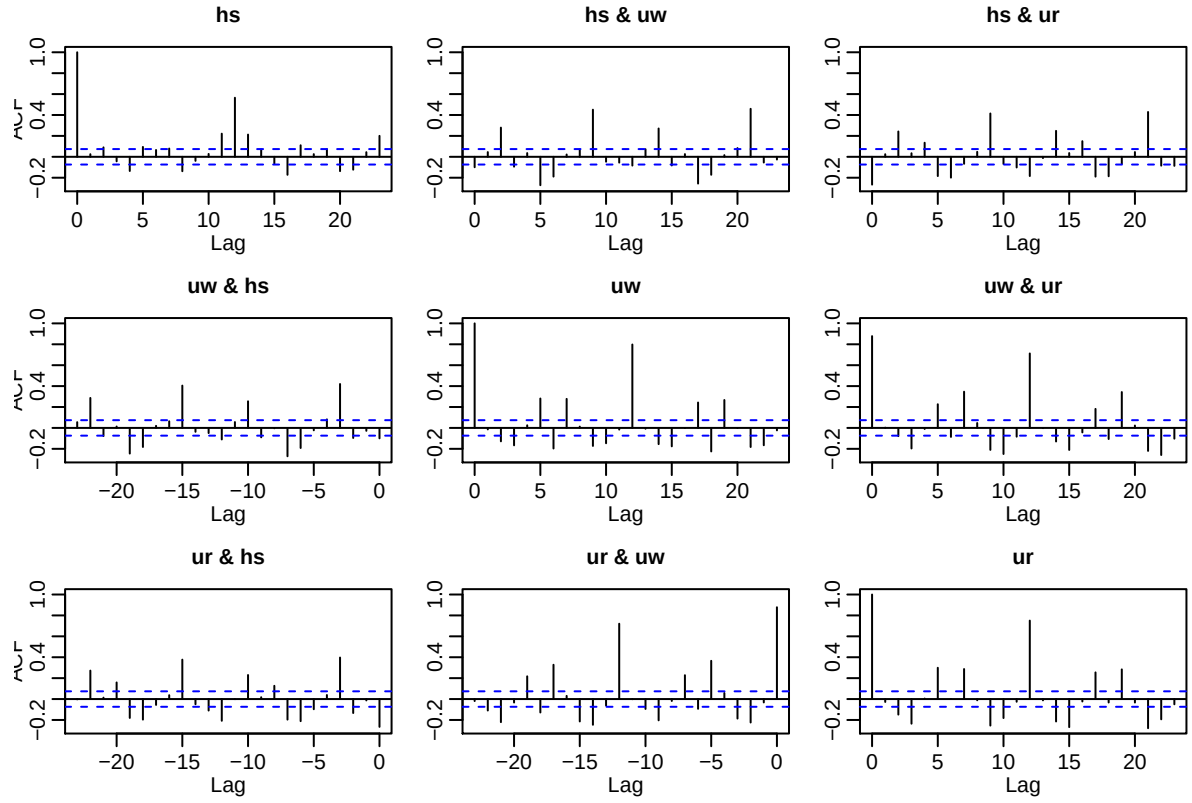


Figure 17. ACFs and CCFs of VAR(2) with constant term.

Figure 17 is the best model in terms of white noise residuals in its ACFs and PACFs. The other three models (none, trend, both) are either not white noise or are more complex than constant term without significant improvement. This makes sense too, given Table 2, because there was no real trend in the timeplots, but the values did have a constant term from which they started.

6.3 Leading and Lagging Variables

input	response	significant_lags
hs	hs	1, 2
hs	uw	1, 2
hs	ur	1, 2
uw	hs	none
uw	uw	1, 2
uw	ur	1
ur	hs	1, 2
ur	uw	1, 2
ur	ur	1, 2

Table 12. Demonstrates the significant coefficients for each of the variables.

From Table 12, we can decipher which variables are leading which variables.

hs is leading uw and ur across both lags.

uw is leading ur across 1 lag

ur is leading hs and uw across both lags.

Thus, both hs and ur lead uw; also, a VAR(2) has fit well since there are many t-1 and t-2 significaint lags.

6.4 VAR(2) Model Equation

The following equation makes up the final model to predict housing starts:

$$hs = 1.02760 * \mathsf{hs}_{t-1} - 0.13663 * \mathsf{hs}_{t-2} + 6.27333 * \mathsf{uw}_{t-1} - 9.62279 * \mathsf{uw}_{t-2} - 10.19133 * \mathsf{ur}_{t-1} + 13.58082 * \mathsf{ur}_{t-2}$$

All variables in this equation are significant. In comparing with the results we obtained in Section 3.2, our conclusions match up: these variables are cointegrated since they are all significant in each other's equations. Thus, the relations are not spurious and are genuine because there are leading and lagging variables.

6.5 Forecasting VAR Model

time_ahead	var_prediction
1	103.2907
2	104.0762
3	105.4022
4	106.8394
5	108.2433
6	109.5391
7	110.6987
8	111.7179
9	112.6038
10	113.3687
11	114.0268
12	114.5920

Table 13. Forecast estimates from the VAR(2) model.

6.6 Forecasted Plot of VAR Model

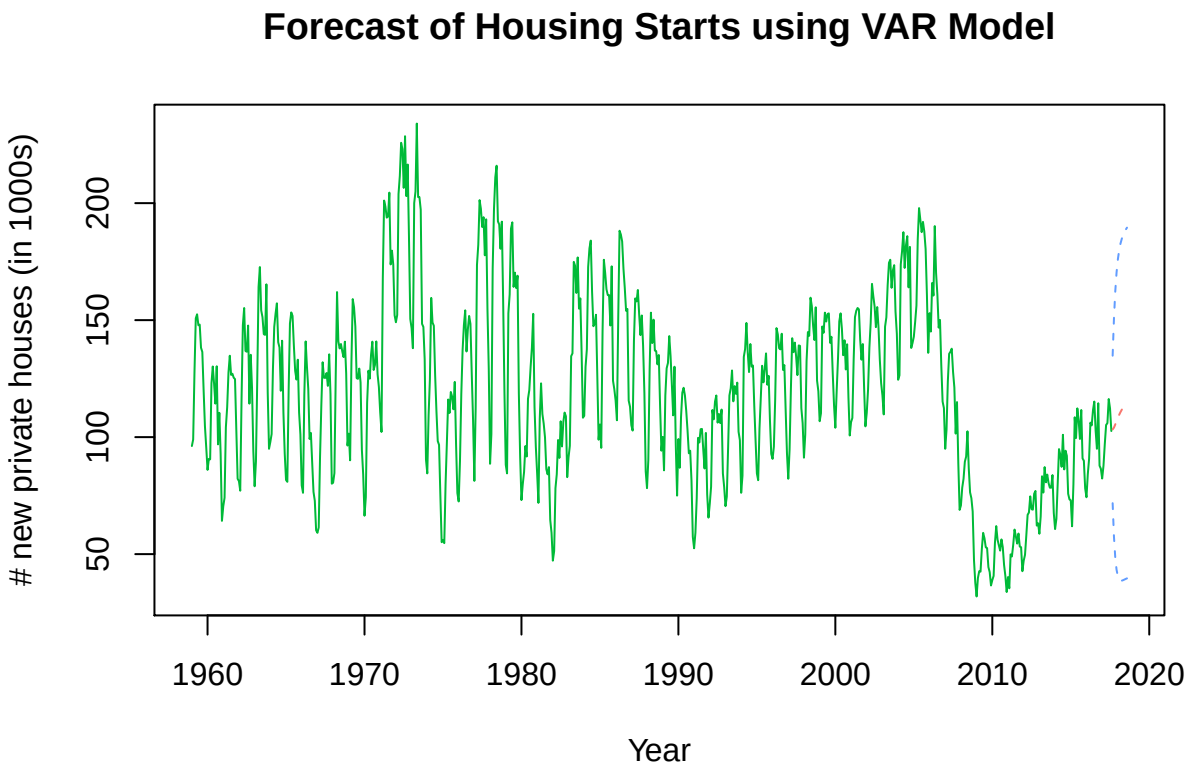


Figure 18. Forecasted values of estimates using VAR model.

Figure 18 shows the plot of VAR(2) forecasts. The confidence interval is also large due to the nature of high variance in the time plot. Overall though, it is still not a bad forecast as it seems to resemble the pattern of the time plot.

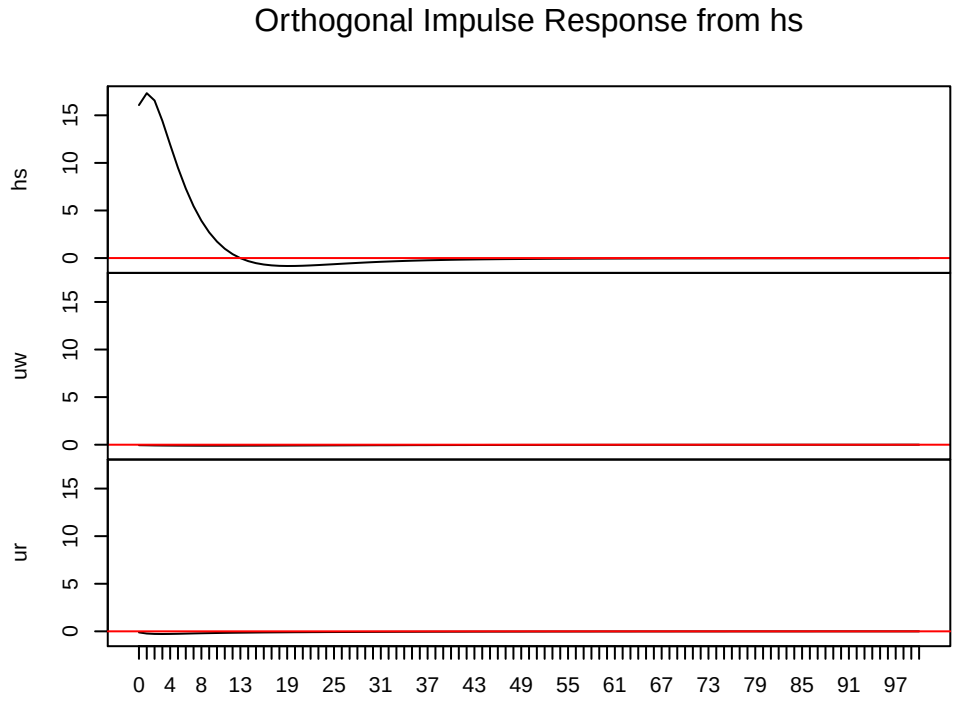
6.7 RMSE of VAR Model

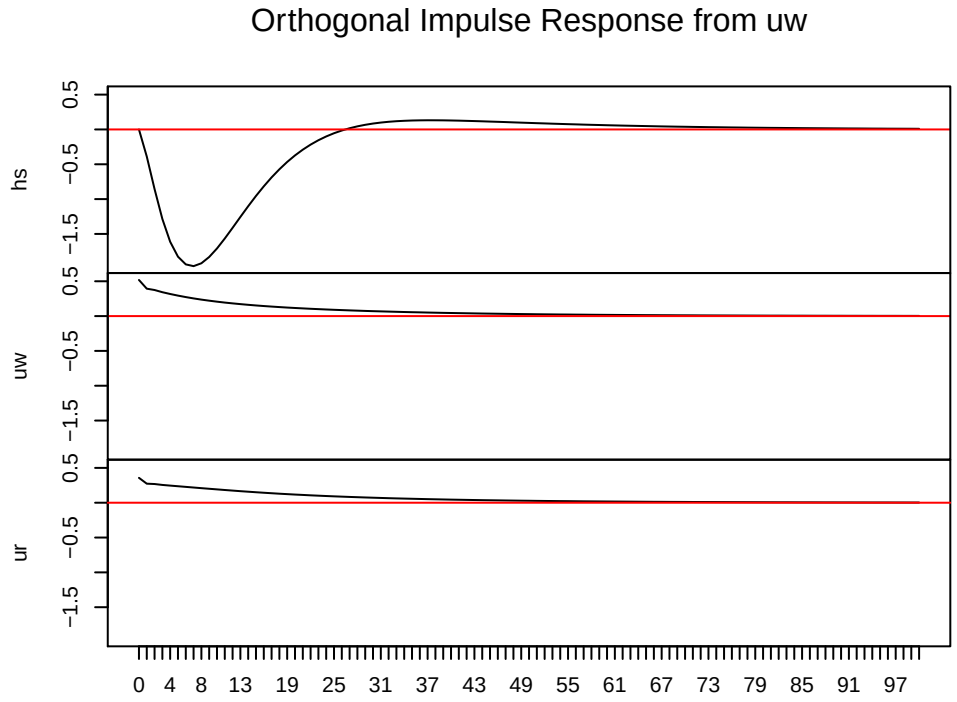
model	RMSE
VAR	11.49255

Table 14. RMSE value of VAR model.

Table 14 shows that the final RMSE for the VAR model is in between the other two models. This values of all three RMSE values are in the same ballpark, which helps ascertify that these forecasts together are more accurate.

6.8 Impulse Response Analysis





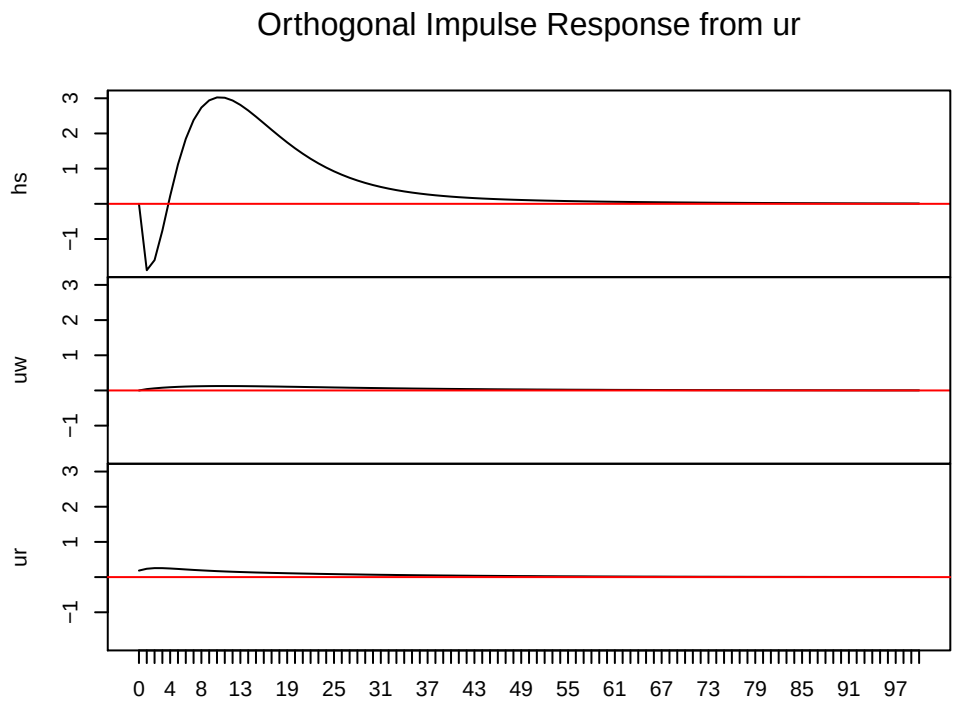


Figure 19. The impulse response function plots for the three variables.

Figure 19 illustrates all the impulse response functions plots. Here are some interpretations to be drawn from them:

A one unit increase in hs at time 0 does not affect uw or ur at all since these variables remain constant and stable and stay at their original states. It does, howveer, affect hs by starting with an initial increase that gradually dips down a bit below the original states until it very slowly climbs back up until it reaches its original state where it stabilizes around time 30.

A one unit increase in uw at time 0 affects all three variables. In hs, there is a sudden drop from its original state until time 5, after which it steadily climbs back up to its original state and remains stable from time 70 onward. In both uw and ur, the impact starts a bit above the original state and gradually dips until it stabilizes to its original state after time 50.

A one unit increase in ur at time 0 affects mostly only hs. uw and ur are slightly disturbed at time 2, but they quickly climb down to their original stable position from time 25 onward. hs, however, starts with a sudden dip below its original state until about itme 2, after which it rises and passes its original state at 4 and keeps increasing until time 13, after which it gradually comes down to its original, stable state from time 40 and onward.

Piecing in all of these three analyses, we see that none of the variables have a permanent effect on the other since they all tend to go back to their original state soon. Moreover, only hs is most strikingly disturbed across all three variables, implying that all the variables have a huge affect on hs. uw also has some more power over all three variables, suggesting it has somewhat more of an influencing behavior than the other variables.

7 Forecasts and Conclusions

7.1 Conclusions

Time Forecasted	SARIMA	Classical Regression + GLS	VAR	Average	Actual Data
Sept 1, 2017	101.7	84.52	103.29	98.48	104.4
Oct 1, 2017	104.44	87.27	104.08	101.35	109.6
Nov 1, 2017	88.13	118.77	105.4	102.55	97.9
Dec 1, 2017	80.43	138.83	106.84	101.87	81.4
Jan 1, 2018	77.57	144.84	108.24	105.56	91.6
Feb 1, 2018	81.03	144.72	109.54	106.25	89.7
March 1, 2018	96.02	138.52	110.7	113.11	107.2
April 1, 2018	106.16	136.66	111.72	118.01	117.5
May 1, 2017	108.75	129.35	112.6	118.6	123.7
June 1, 2017	112.5	133.42	113.37	117.82	112
July 1, 2017	110.25	109.94	114.03	111.53	111.9
Aug 1, 2017	103.63	94.68	114.59	106.38	112.6

Time Forecasted	SARIMA	Classical Regression + GLS	VAR	Average	Actual Data
RMSE	8.96	32.47	11.493	5.83	N/A

Table 15. All forecasts from the three models, along with the average and the actual data, supplemented by the RMSE values.

Based off Table 15, we see that the model with the lowest RMSE is the SARIMA model. If I had to only choose one model and forecast, I would choose SARIMA. SARIMA also makes sense because it is the only that truly accounts for seasonality in both the AR and MA processes, unlike VAR. And from Table 2, it is clear that seasonality must be modelled. However, a more sound forecast is the average, or the "consensus" forecast of the three forecasts. This is the forecast I would use since it combines multiple metholodogies, overriding assumptions made for each model since no data can really be modelled perfectly to begin with.

7.2 Practical and Future Applications

It's difficult to discern the importance of housing starts in the economy without actually using variables more directly involved in the economy. However, through our approach in VAR, we discussed how hs is a leading variable for ur and uw which are two variables somewhat representing the economy. Analyzing further to what extent hs plays an influence in the economy and market by introducing more variables could be a fascinating project.