Soybean Futures Returns Modeling

Sandeep Girada, Krishnachaitanya Pullakandam and Alex Tunnell

November 2019

1 Introduction

Soybean is a globally important crop. It is a great source of both edible oil and vegetable protein. Soy Oil is the second most consumed edible oil in the world and Soy Meal is an essential ingredient in animal feed. USA and Brazil are two of the biggest producers and exporters of Soybeans whereas China and Europe are the two biggest consumers.
The prices of Soybean are highly sensitive to supply and demand, consumption trends and trade restrictions. The forecasting of Soybean Prices or returns can be of great importance to decision making in the soy-complex industry. Accurate forecasting can play a major role in effectiveness of hedging practices to both producers and consumers.
In this study we aim to analyze the series of Monthly Prices or Returns of Soybean Futures (Active Continuation) in the USA from Jan 2000 to Oct 2019 and describe their behaviour and fitting using different time series models.

2 Data

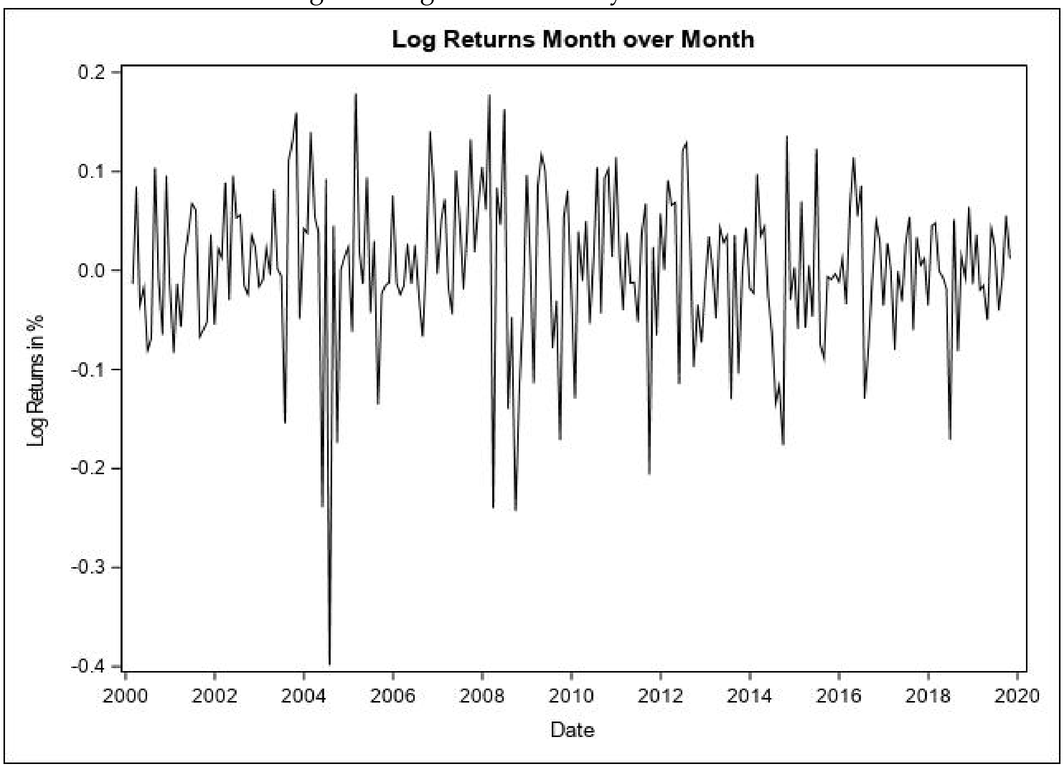
The Monthly closing Prices of the Soybean (Active Continuation) Futures Contract was collected for the time period Jan 2000 to Oct 2019 from Bloomberg. We collected several other Futures Prices such as Lean Hog and Soybean by-products such as Meal and Oil to check for possible leading indicators. Once we analyzed the data we concluded that we would model the returns (change in price), as the prices themselves were not stationary.

We used an approximation (log returns) for the change in price given by:

$$R_t = ln(\frac{S_t}{S_{t-1}}) \tag{1}$$

Where S_t is the price of the soybean futures contract and R_t is the return for month t. The plot for the log returns of the soybean futures are given in the figure below:

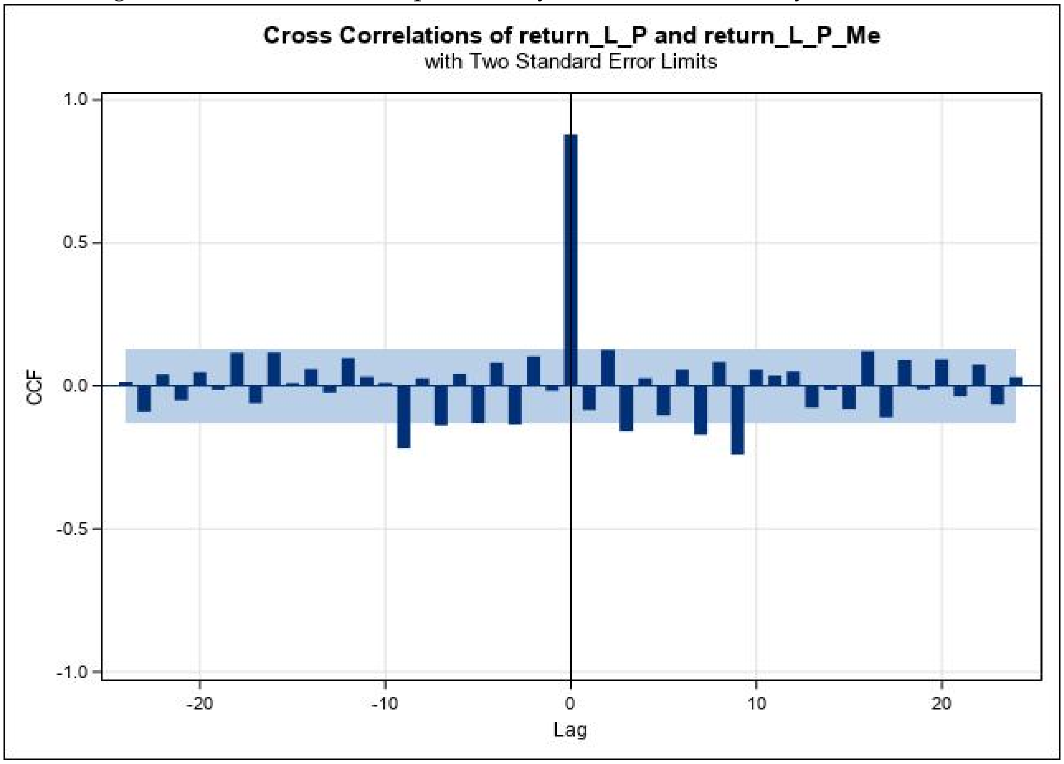
Figure 1: log Returns for Soybean Futures



From the figure above, we observe non-constant variance (heteroscedasticity) in the data. Given that the volatility seemed to change through time, we decided to look into modeling heteroscedastic processes in an attempt to account for this behavior. We also see the presence of some outliers in the data. In this report, we did not use any smoothing techniques or data replacement to reduce the effects of the outliers, as this would manipulate the data and move us further from modeling the true series.

We also analyzed potential leading indicators. One would expect that a change in Soy Meal, Oil or Lean Hogs prices would lead to a change in the soybean futures price. We analyzed the cross correlations between these and observed the relation below for Soy Meal.

Figure 2: Cross Correlation plots of Soybean Returns and Soy Meal Returns

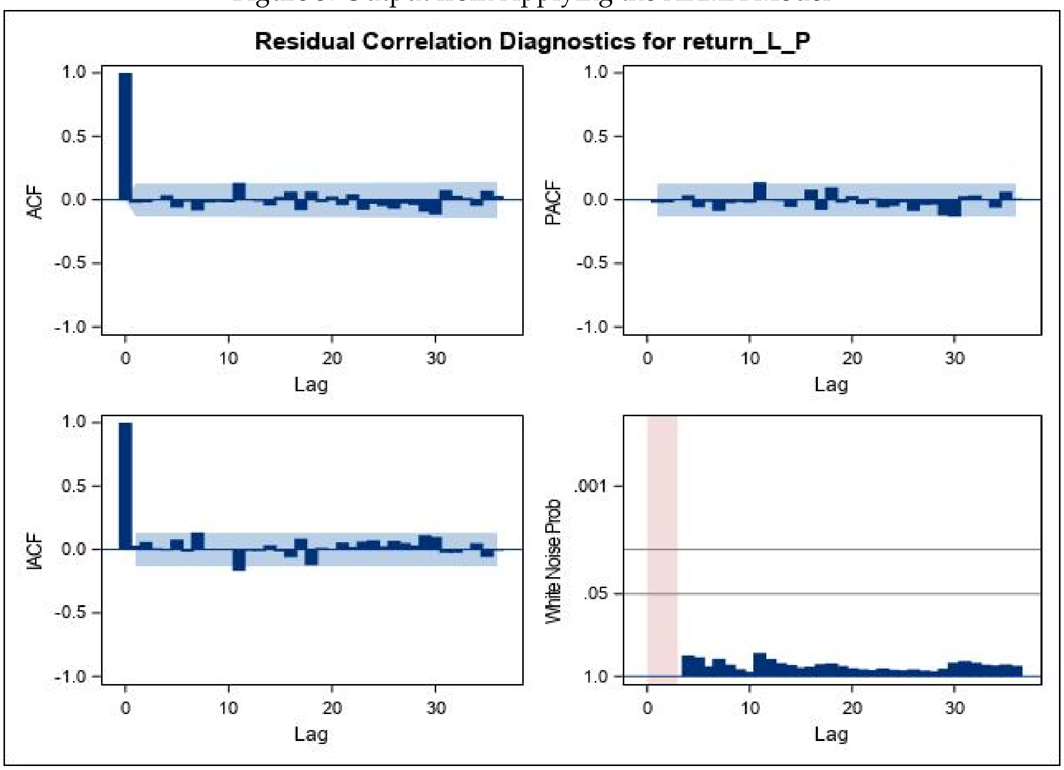


There is high cross-correlation for Soy Meal at lag zero. We can also observe some correlation at lag 9. However, we were not able to fit a transfer function for these two variables. Lean Hogs and Soy Oil also had similar cross correlation hence none of them would suffice as a leading indicator for the monthly data. We expect that more granular data, like returns on an hourly or daily basis might have a relationship that would suit using a transfer function. However, we decided to keep using data on a monthly basis.

3 Approach

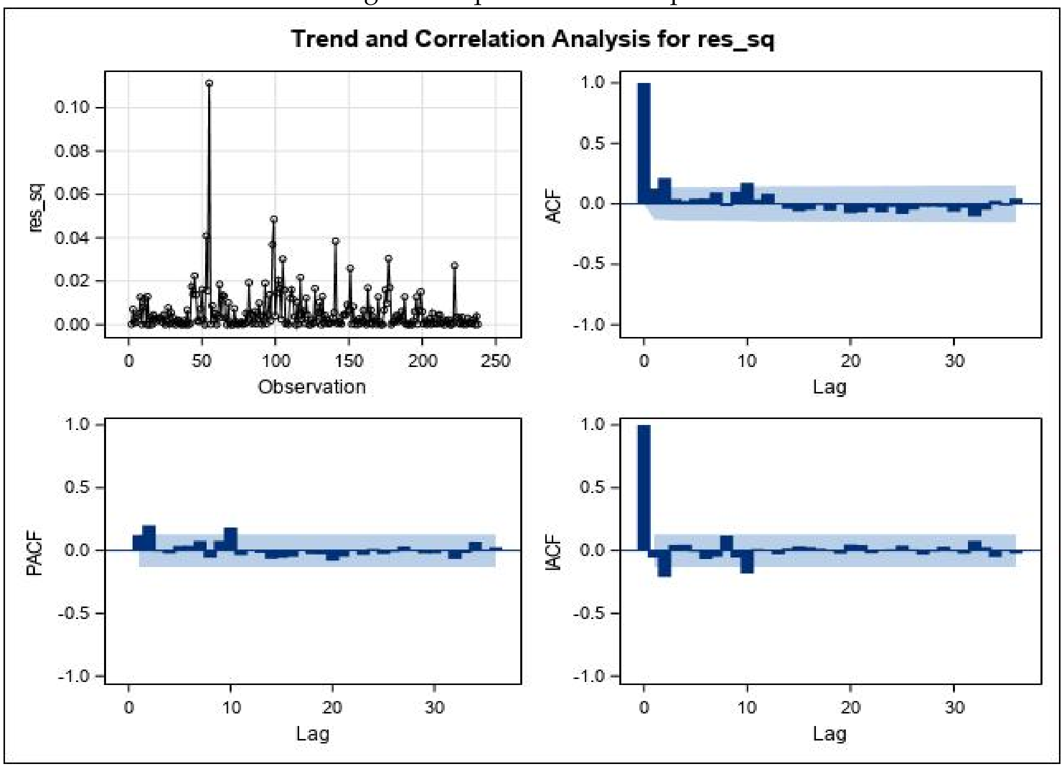
We began our approach by using the ARIMA model on our time series. Using proc arima, we fitted a model to our log returns. Our best model was a ARMA model with p = (2,3) and q = (9). The 9th period lag for the MA, may make sense because the periodic crop cycles. Given production is quite high in both hemispheres, we would not expect specific seasonality corresponding to a single month. In our output, we see that the residuals appear normal as can be seen from the below White Noise Probability results.

Figure 3: Output from Applying the ARMA Model



If we square the residuals, we see volatility clusters and spikes, as seen in the below figure.

Figure 4: Squared Error Outputs



As we can see, clustering and spiking occurring in the plot of the squared residuals. Given these observations and existing papers on volatility modeling of soybean oil futures [4], we determined that we should look into modeling for the non-constant variability within the series.

In many cases, financial time series data exhibit volatility clustering i.e. non-constant conditional variance (heteroscedasticity). An autoregressive conditional heteroscedastic (ARCH process), models the variance at time t as a linear combination of past squared residuals. A generalized ARCH (GARCH) is a generalization of the ARCH process where the variance model looks more like an ARMA model opposed to just an AR model. As explained in [1], the usual approach to GARCH(p,q) models is to model the error in terms of standard white noise $e_t \sim$ N(0,1), as $\epsilon_t = \sqrt{h_t} e_t$ where h_t satisfies the type of recursion used in an ARMA model:

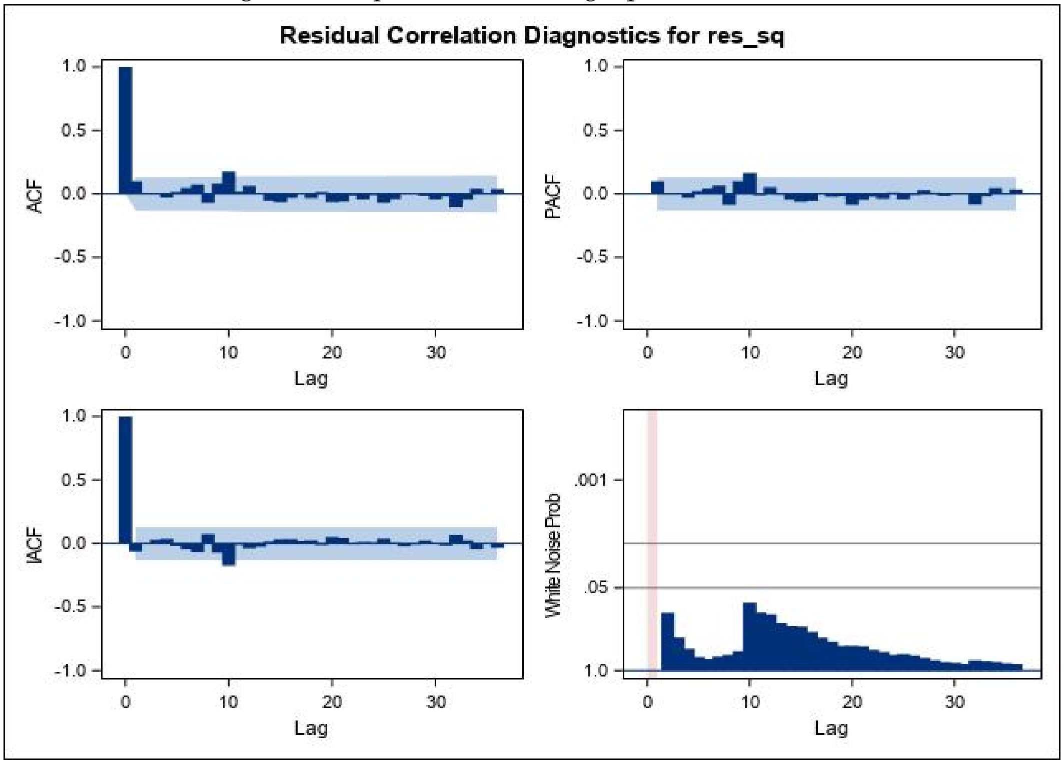
$$h_t = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_i^2 + \sum_{i=1}^{q} \gamma_j h_{t-j} \tag{2}$$

where,

$$\omega = ARCH0$$
$$\alpha_1 = ARCH1, \alpha_2 = ARCH2, ..., \alpha_q = ARCHq \tag{3}$$
$$\gamma_1 = GARCH1, \gamma_2 = GARCH2, ..., \gamma_q = GARCHq$$

By fitting a lag 2 AR to the squared residual series, we can account for all auto-correlated effects and produce a result where the residuals are white noise.

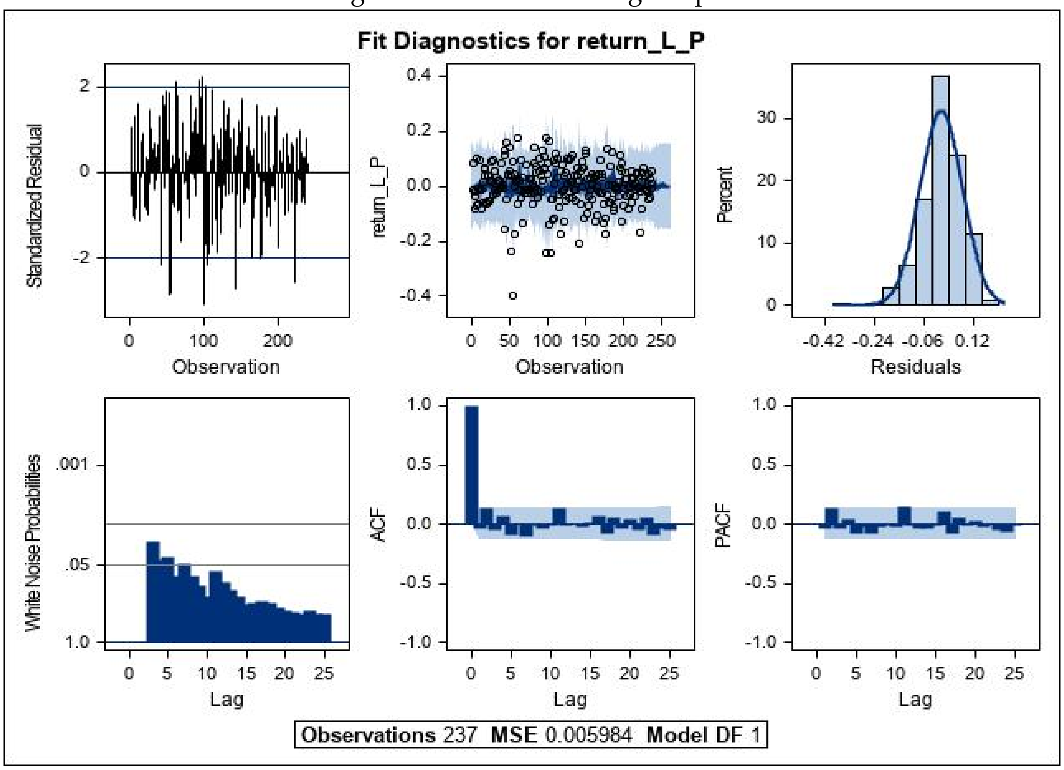
Figure 5: Output from Modeling Squared Error Series



The GARCH model allows us to model the variances as a ARMA, but as we saw from above, only a lag 2 AR was needed to model the squared residuals. Therefore, we can implement an ARCH model where q = (2). The output from proc autoreg for a GARCH(q=(2)) is given below. We also tried to model other GARCH models but adding more p,q terms makes the AR coefficients insignificant.

The proc autoreg fit below has a lag 3,9 AR coefficients and an ARCH lag 2 coefficient. These coefficients are significant at 0.05 threshold.

Figure 6: GARCH Autoreg Output



The parameter estimates given below will help us determine the h_t and the autoregressive error model. The resultant process equation has been mentioned below for convenience.

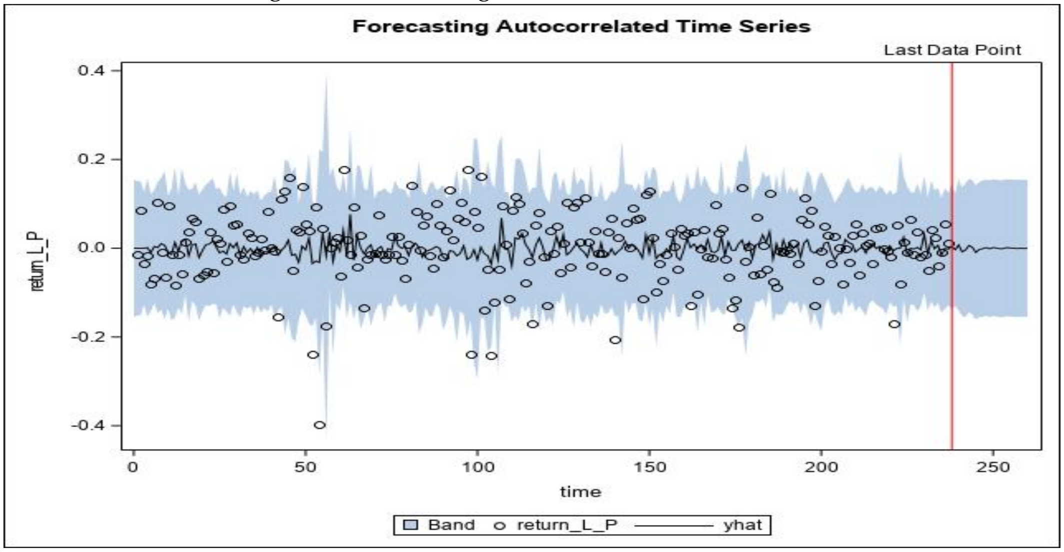
Figure 7: Parameter Estimates

Parameter Estimates					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
AR3	1	0.1078	0.0536	2.01	0.0443
AR9	1	0.1759	0.0560	3.14	0.0017
ARCH0	1	0.004108	0.000524	7.83	<.0001
ARCH2	1	0.3016	0.0910	3.31	0.0009

$$\begin{aligned} h_t &= 0.004108 + 0.3016\epsilon_{t-2}^2 \\ z_t &= 0.1078z_{t-3} + 0.1759z_{t-9} + \epsilon_t^2 \quad \text{where} \quad -\epsilon_t = \sqrt{h_t}e_t \end{aligned} \tag{4}$$

We used proc sgplot to show a comparison between the actual and the fitted values. The next 2 years ahead forecasts are also plotted and can be seen reverting back to zero .

Figure 8: Forecasting Autocorrelated Time Series



4 Conclusion

We observe from the ARIMA fitting that the returns are weakly correlated, but a simple ARIMA fitting does not explain the process very well. We would probably need a leading indicator to better model the returns.

The squared residuals show a pattern of clustered spikes which can be explained with an AR lag 2 process. This led us to model a GARCH(q=(2)) using proc autoreg. We observed that the GARCH model has an AIC (-555.039) smaller than the value we observed with regular ARMA model (-543.36). The forecasts of the GARCH process incorporate the volatility into the estimates which can be noticed from the large confidence intervals of the estimates corresponding to the large movements in the data.

References

[1]

Brocklebank, John C., David A. Dickey, and Bong S. Choi. 2018. SAS for Forecasting Time Series, Third Edition. Cary, NC: SAS Institute Inc.

[2]

Martin, Donald. Class Notes. North Carolina State University. Print

[3]

Roncoroni, Andrea. Fusai, Gianluca. Cummins, Mark. 2015. Handbook of Multi-Commodity Markets and Products, Structuring, Trading and Risk Management. John Wiley & Sons Ltd.

[4]

Sahai. Alok, Kumar. Volatility Modeling and Forecasting Efficiency of GARCH Models on Soy Oil Futures in India and the US.