Time-Series Analysis of 20+ Year Treasury Bond ETF

Alexander Rom
University of San Francisco
December 13, 2024

Abstract:

This paper analyzes the behavior of daily returns of the iShares 20+ Year Treasury Bond ETF (TLT), focusing on the impact of macroeconomic and financial market variables, including market volatility index, S&P500 returns, changes in Fed rates, Treasury bond market yield, and inflation. Models utilized daily data from January 2004 to July 2024, the analysis incorporates advanced time-series models to capture the dynamics of returns in TLT. An ARIMAX model with state-dependent dynamics using a Threshold Autoregressive (TAR) framework reveals the significant role of Federal Funds Rate changes in regime-dependent effects on returns. Additionally, a GARCH-ARIMAX model was employed to address volatility clustering and persistence, highlighting the impact of recent shocks and long-term volatility persistence. Both models demonstrate the critical influence of market volatility and equity returns on the daily performance of TLT. The findings provide a comprehensive understanding of the drivers of long-term Treasury bond returns and underscore the importance of accounting for regime changes and time-varying volatility in modeling.

Keywords:

Treasury Bond ETF Returns, Volatility Clustering, Federal Funds Rate, ARIMAX Model, GARCH Model, Threshold Autoregressive (TAR), Financial Time Series, Forecasting.

JEL Classifications Numbers: C22, E44, G12, G17

1 Introduction

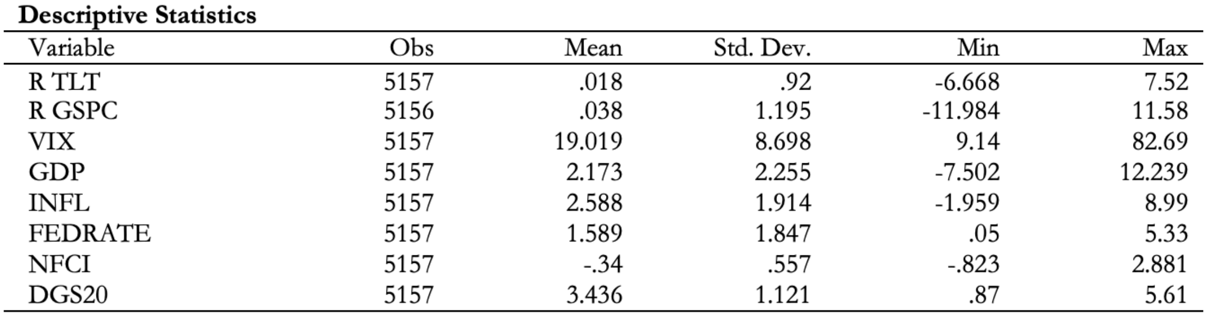
This paper analyzes and models the behavior of daily returns of the iShares 20+ Year Treasury Bond ETF (R_TLT), focusing on the role of macroeconomic and financial market variables in explaining its dynamics. Key factors such as market volatility (VIX), equity returns (R_GSPC), inflation (INFL), the Federal Funds Rate (FEDRATE), and long-term Treasury yields (DGS20) are examined to uncover their relationships with R_TLT. Another critical question addressed is whether changes in the Federal Funds Rate significantly influence the returns of R_TLT. The analysis employs two main models: an ARIMA and ARIMAX framework with state-dependent dynamics through a Threshold Autoregressive (TAR) approach, and a GARCH and ARIMA model with exogenous variables to capture time-varying volatility. Model performance is evaluated based on the Akaike Information Criterion (AIC) and Root Mean Square Forecast Error (RMSFE) to compare different specifications. This comprehensive approach aims to improve understanding of the factors driving long-term Tresurydy bond ETFs' returns while enhancing forecasting accuracy for this critical financial instrument.

2 Data Description

The dataset includes daily observations for all trading days from January 5, 2004 to July 1, 2024 (5156 observations). All data was sourced from Yahoo Finance and the Federal Reserve Economic Data (FRED). Monthly and quarterly data were converted to daily frequency by extending the values across their respective periods to align with the daily trading days.
Core variables are:

- **R_TLT** = iShares 20+ Year Treasury Bond ETF daily returns, (daily percentage)
- **R_GSPC** = S&P 500 daily returns, percentage (daily percentage)
- **VIX** = Index measures the market's expectation of 30-day volatility based on S&P 500 option prices (adjusted close daily)
- **GDP** = GDP growth Percent change from a year ago (quarterly)
- **INFL** = CPI inflation Percent change from a year ago (monthly)
- **FEDRATE** = Federal Funds Effective Rate (monthly)
- **NFCI** = Chicago Fed National Financial Conditions Index (weekly ending Friday)

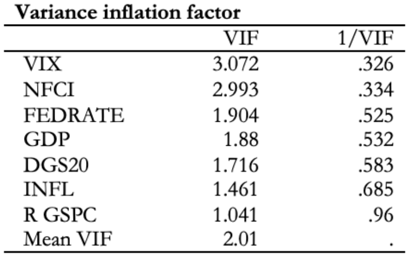
- **DGS20 =** Market Yield on U.S. Treasury Securities at 20-Year Constant Maturity, Quoted on an Investment Basis (daily percentage)



Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
R TLT	5157	.018	.92	-6.668	7.52
R GSPC	5156	.038	1.195	-11.984	11.58
VIX	5157	19.019	8.698	9.14	82.69
GDP	5157	2.173	2.255	-7.502	12.239
INFL	5157	2.588	1.914	-1.959	8.99
FEDRATE	5157	1.589	1.847	.05	5.33
NFCI	5157	-.34	.557	-.823	2.881
DGS20	5157	3.436	1.121	.87	5.61

Our dependent variable exhibits a significant range of values, highlighting that TLT returns can fluctuate considerably within a single day. However, these fluctuations are not as significant as those for GSPC returns, which range from losses exceeding -11% to gains slightly above 11% in a day. It is important to note that due to reasons that variables such as GDD, INFL, and FEDRATE are not reported daily and had to be extended to match daily frequency their mean, minimum, and maximum values correspond to the periods of three reporting periods (month or quarter).

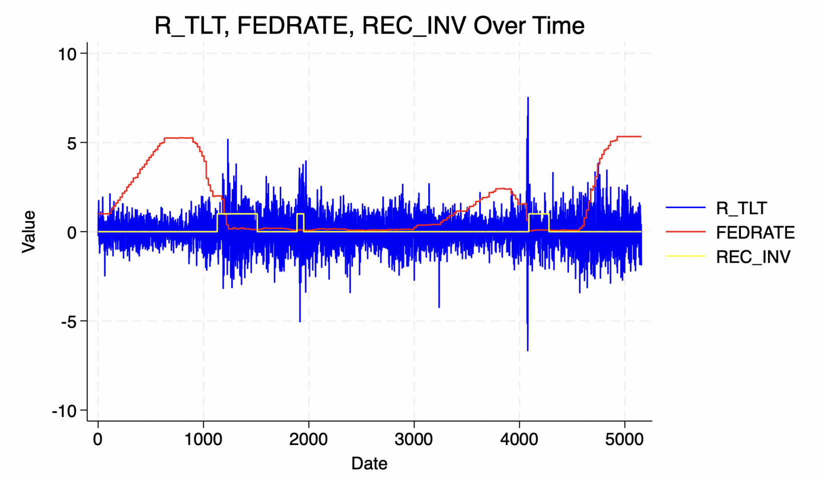


Variance inflation factor

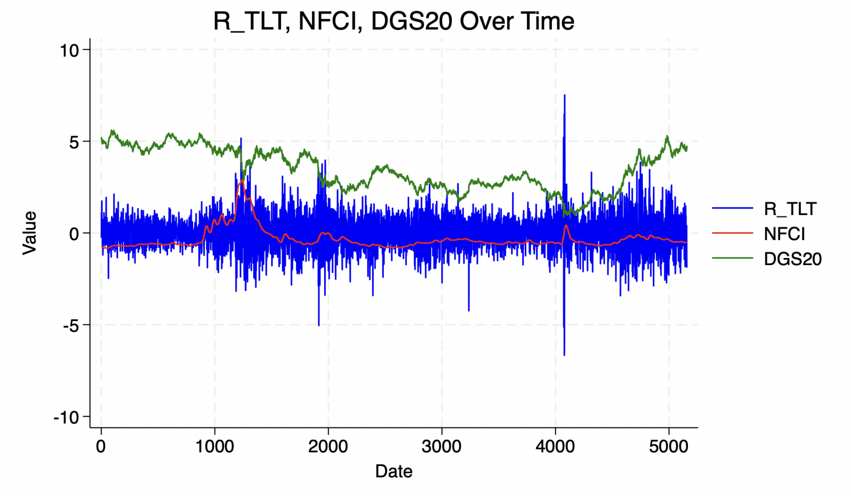
	VIF	1/VIF
VIX	3.072	.326
NFCI	2.993	.334
FEDRATE	1.904	.525
GDP	1.88	.532
DGS20	1.716	.583
INFL	1.461	.685
R GSPC	1.041	.96
Mean VIF	2.01	.

Since the same market and macroeconomic factors influence many variables, it is reasonable to assume they may share similar trends and movements which can cause multicollinearity. However, the Variance Inflation Factor (VIF) was found to be less than 5 for all variables in the model, indicating that multicollinearity is not a concern.

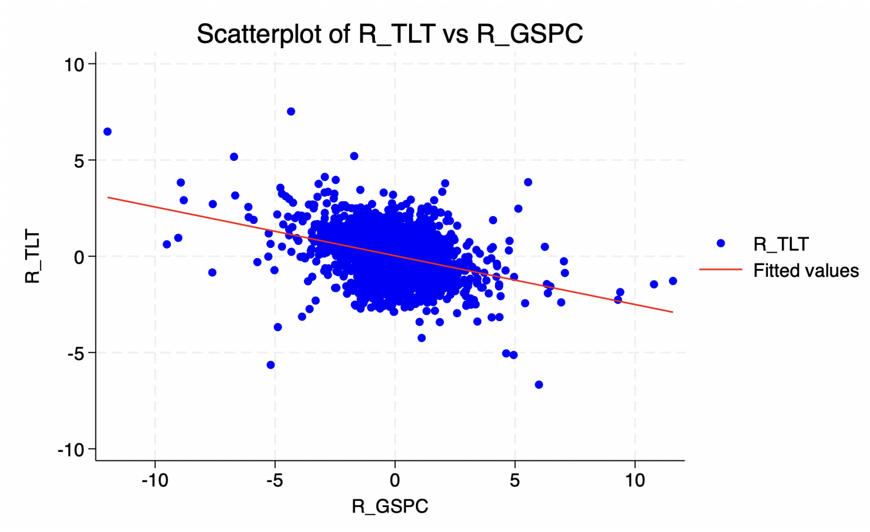
From plot (2.1) it can be seen that higher fluctuations in returns of TLT are in periods when the FED cuts rates and significantly lower returns in periods when rates increase, which is consistent with bond theory the inverse relationship between rates and bond prices. Additionally, on the same graph, it visualized that in periods of recessionary investment when GPD growth is less than 2% TLT sees higher returns which is also consistent with Monterey policy and bond prices in the market. As shown in the next plot (2.2), higher NFCI values, which indicate tighter financial conditions, are associated with higher TLT returns. This aligns with the conception that during periods of financial distress, investors tend to reallocate their assets toward more secure long-term treasuries. Evidence from the plot (2.3) supports that, as it shows that lower returns of GSPC are assisted with higher returns of TLT and vice versa, indicating an inverse relationship between the two returns. Referring back to graph (2.2) and analyzing the movement of DGS20 compared to TLT returns, there is no significant visual evidence or clear pattern to suggest a direct relationship between the two, while theoretically, it should be inverse. Only on three occasions (around dates 1200, 2000, and 4000) do we observe drops in market yields corresponding to higher TLT returns. However, this pattern is not consistent throughout the entire sample period.



(2.1)



(2.2)



(2.3)

Variable	Z(t) statistic	p-value for Z(t)
R TLT	-88.371	0.0000
R GSPC	-93.953	0.0000
VIX	-57.348	0.0000
GDP	-50.739	0.0000
INFL	-50.748	0.0000
FEDRATE	-50.786	0.0000
NFCI	-50.750	0.0000
DGS20	-53.692	0.0000

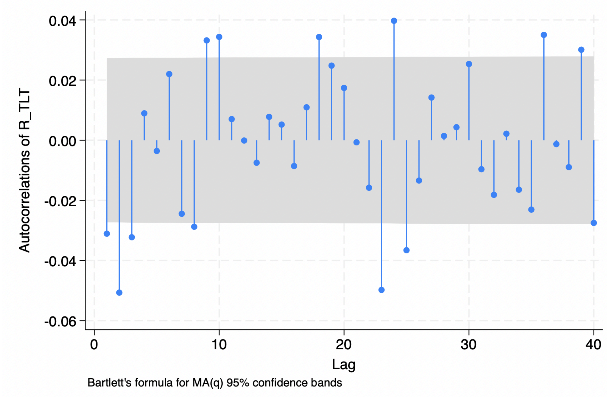
To check for stationarity in series we conduct Augmented Dickey-Fuller Test (ADF). The ADF test checks for the presence of a unit root, where the null hypothesis assumes that the variable is non-stationary. For all variables in the table, the Z(t) test statistic is highly negative, and the corresponding p-values are 0.0000, indicating that the null hypothesis of a unit root is strongly rejected at all conventional significance levels. Therefore, it could be concluded that all variables are stationary, meaning their statistical properties remain constant over time. This allows time-series analysis for these variables without further differencing or transformation.

3 ARMA Model

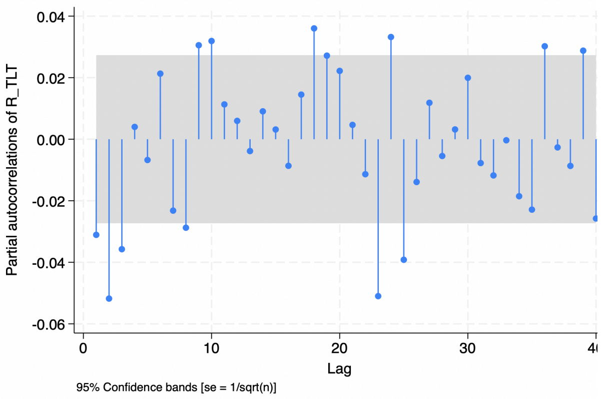
3.1 Specification and Estimation

By analyzing autocorrelation (AC) and Partial Autocorrelation (PAC) from (3.1) and (3.2) we see several lag values outside of 95% confidence bands, with the majority falling within it.

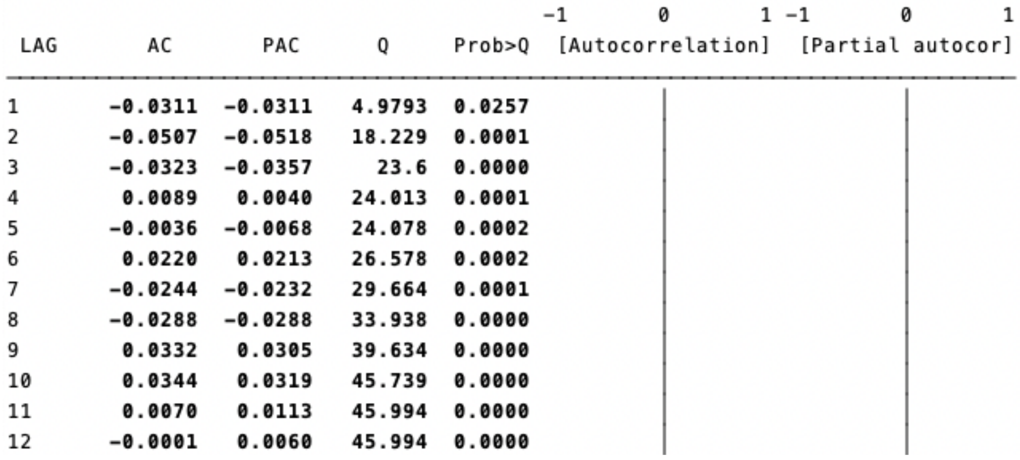
The results do not follow a clear ARIMA process with strong AR or MA components however spikes at the first 3 lags could assume the presence of limited autoregressive structure in the R_TLT data. The correlogram table (3.3) below provides the autocorrelation (AC) and partial autocorrelation (PAC) values for R_TLT, along with Q-statistics and their corresponding p-values. The autocorrelation and partial autocorrelation values at all lags are relatively small, indicating weak serial dependence in returns. However, the Q-statistics are significant for most lags, with low p-values below 0.05. This suggests that while individual lags may not show strong autocorrelation, there is evidence of weak dependence across multiple lags. We see that after the 3rd lag AC and PAC almost fully convert to zero, thus it is relevant to estimate all models of ARIMA up to 3rd lag for both MA and AR components.



(3.1)



(3.2)



```
                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial autocor]
-------------------------------------------------------------------------------
1       -0.0311  -0.0311   4.9793  0.0257          |                  |
2       -0.0507  -0.0518   18.229  0.0001          |                  |
3       -0.0323  -0.0357     23.6  0.0000          |                  |
4        0.0089   0.0040   24.013  0.0001          |                  |
5       -0.0036  -0.0068   24.078  0.0002          |                  |
6        0.0220   0.0213   26.578  0.0002          |                  |
7       -0.0244  -0.0232   29.664  0.0001          |                  |
8       -0.0288  -0.0288   33.938  0.0000          |                  |
9        0.0332   0.0305   39.634  0.0000          |                  |
10       0.0344   0.0319   45.739  0.0000          |                  |
11       0.0070   0.0113   45.994  0.0000          |                  |
12      -0.0001   0.0060   45.994  0.0000          |                  |
```

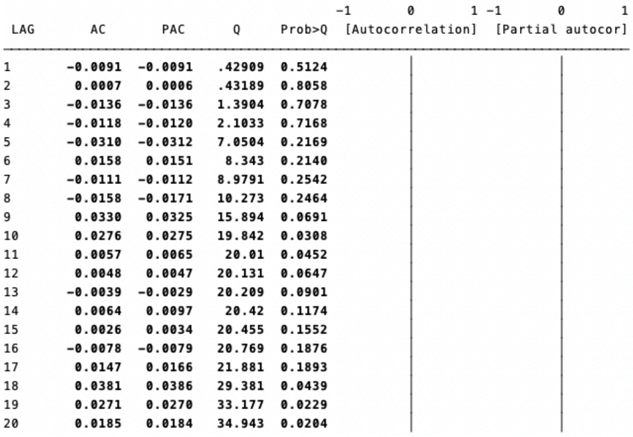
(3.3)

Table (3.4) shows different AR and MA specifications of the ARIMA model for the R_TLT variable. While models with up to 3 lags were tested and compared, only those with up to 2 lags are shown in the table. The optimal model was selected by comparing AIC values, with the lowest AIC corresponding to the ARIMA(2,0,2) model, which has an AIC of 13,765.91. Furthermore, both the AR and MA coefficients for the second lag were found to be statistically significant. However, if we look at lag 1, the AR coefficient is 0.6436, and the MA coefficient is -0.6688. Since these values are close in absolute value but have opposite signs, their effects largely cancel each other out at the first lag and a similar effect is observed for the coefficients at the second lag.

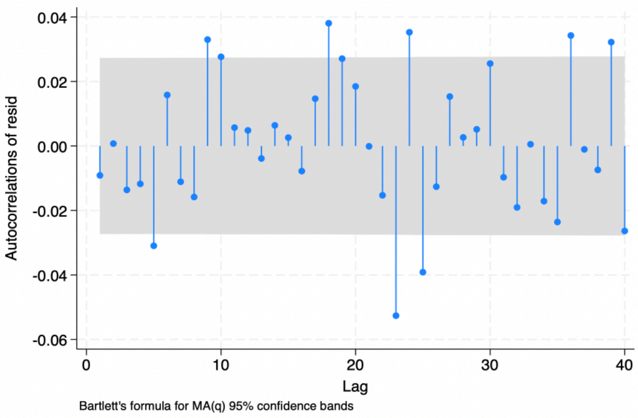
Model ARIMA	***AR***	***P>lzl***	***MA***	***P>lzl***	***AIC***
(0,0,0)	-	-	-	-	13783.32
(1,0,1)	.599645	0.000	-.6475988	0.000	13768.88
(1,0,2)	.3510859	0.016	L1. -.3859 L2. -.0412	L1. 0.008 L2. 0.001	13766
(2,0,2)	L1. .6436 L2. -.6166	L1. 0.000 L2. 0.000	L1. -.6688 L2. .5814	L1. 0.000 L2. 0.000	13765.91

(3.4)

Predicted residuals from the ARIMA(2,0,2) model have a mean equaling 3.16e-06 which is consistent with the white noise behavior assumption. Table (3.5) shows the correlogram output of residuals AC and PAC from the ARIMA(2,0,2) model. AC and PAC values for most lags are close to zero, and many are within the 95% confidence bounds, indicating weak serial dependence. However, a few lags (such as 9,10, 20) have significant p-values suggesting some evidence of autocorrelation at these specific lags, which is not consistent with white noise behavior. Additionally, graph (3.6) supports this evidence as residuals mostly behave like white noise, but there are significant spikes in a few lags hinting at minor autocorrelation that the model might not have fully captured. This suggests that the model requires more specifications and extensions to fully meet white noise assumptions for residuals.



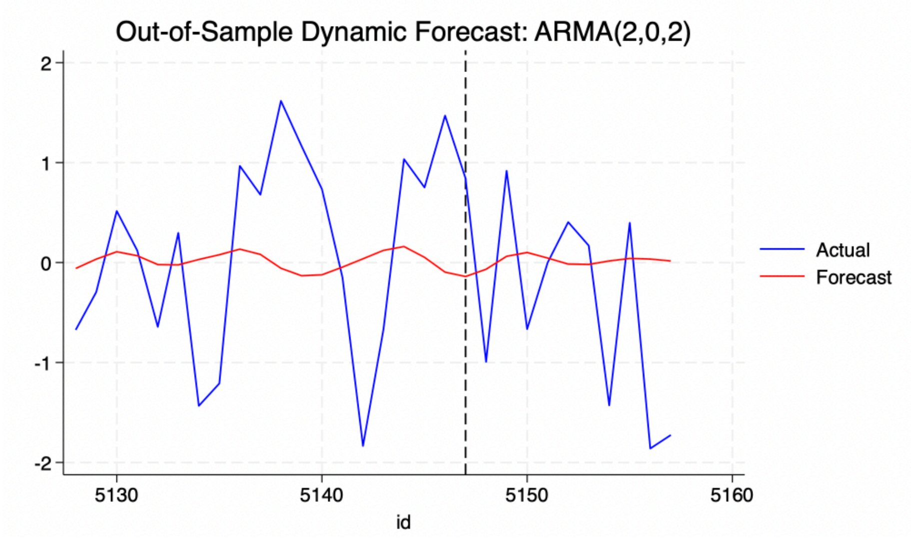
```
                                          -1       0       1 -1       0       1
LAG       AC        PAC        Q     Prob>Q  [Autocorrelation]  [Partial autocor]
-------------------------------------------------------------------------------
1       -0.0091   -0.0091   .42909  0.5124
2        0.0007    0.0006   .43189  0.8058
3       -0.0136   -0.0136   1.3904  0.7078
4       -0.0118   -0.0120   2.1033  0.7168
5       -0.0310   -0.0312   7.0504  0.2169
6        0.0158    0.0151    8.343  0.2140
7       -0.0111   -0.0112   8.9791  0.2542
8       -0.0158   -0.0171   10.273  0.2464
9        0.0330    0.0325   15.894  0.0691
10       0.0276    0.0275   19.842  0.0308
11       0.0057    0.0065    20.01  0.0452
12       0.0048    0.0047   20.131  0.0647
13      -0.0039   -0.0029   20.209  0.0901
14       0.0064    0.0097    20.42  0.1174
15       0.0026    0.0034   20.455  0.1552
16      -0.0078   -0.0079   20.769  0.1876
17       0.0147    0.0166   21.881  0.1893
18       0.0381    0.0386   29.381  0.0439
19       0.0271    0.0270   33.177  0.0229
20       0.0185    0.0184   34.943  0.0204
```



(3.5) *(3.6)*

3.2 Forecast and Evaluation

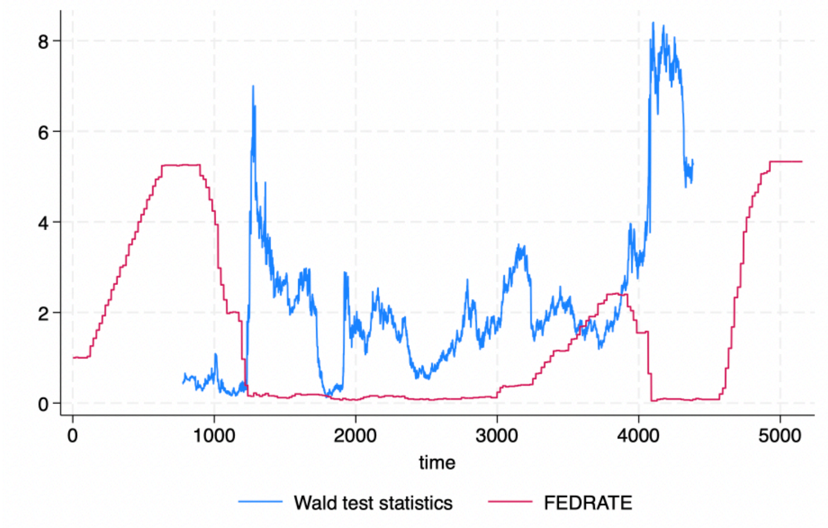
The ex-post forecast for the last 15 observations of the ARIMA(2,0,2) model produced a Root Mean Square Forecast Error (RMSFE) of 1.0598, indicating that, on average, the model's forecasts deviate by 1.06% from the actual values. Graph (3.7) reveals that the predictions significantly underestimate the volatility and are primarily concentrated around zero returns for TLT. This behavior is largely due to model misspecification and the previously mentioned cancelation effect between the AR and MA component coefficients. The ARIMA(2,0,2) model performs poorly in forecasting, despite meeting most of the optimal specification criteria.



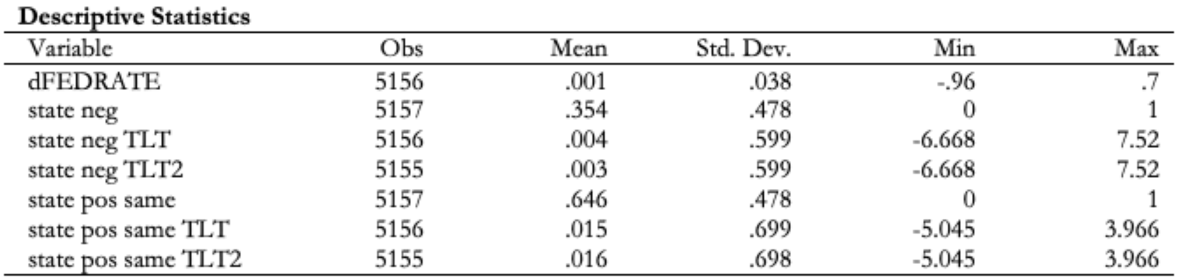
(3.7)

4 ARIMAX(2,0,2) with TAR Model

Several evidence have suggested that the ARIMA model needed improvement, due to omitted variables and the possibility of structural break. Thus we have tested for structural break in the R_TLT variable by applying the Quandt-Andrews test, calculating the Chow F-stat for each day, and seeing all potential breakpoints. At each candidate breakpoint, a Wald test was performed to compare the model's fit with and without the breakpoint. In graph (4.1), both the Wald test statistics and the FEDRATE are plotted, it is evident from the graph that there are several significant spikes in the Wald statistics, which supports the hypothesis that the data has structural breaks multiple times. Since TLT is an ETF based on long-term Treasury bonds, it is highly influenced by the FED rate, which suggests that the regime for TLT may depend on the movements of the FED rate. In the same graph (4.1), it is apparent that sharp drops in the FEDRATE coincide with high Wald statistics, indicating potential structural breaks during these periods. Therefore, the changes in the FEDRATE will be used as a threshold to build a TAR switching model. Specifically, in the table (4.2), the newly generated variables are presented. The variable dFEDRATE represents the change in the FED rate from one period to the next (day to day). The dummy variable state_neg represents periods of decreasing rates, taking the value 1 starting from observations where dFEDRATE is negative and extending for up to 56 observations (the typical maximum time between FED meetings). The extension stops early if a non-zero dFEDRATE (indicating a monetary policy change) is encountered, resetting to 0 for positive values (corresponding to the state_pos_same dummy, which equals 1 during such times) or continuing with 1 for subsequent negative values. Additional variables are created satate_neg_TLT and state_pos_same_TLT which are interaction terms between states and lagged value of R_TLT and interaction for second lag are expressed as state_pos_same_TLT2 state_neg_TLT2.



(4.1)



Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
dFEDRATE	5156	.001	.038	-.96	.7
state neg	5157	.354	.478	0	1
state neg TLT	5156	.004	.599	-6.668	7.52
state neg TLT2	5155	.003	.599	-6.668	7.52
state pos same	5157	.646	.478	0	1
state pos same TLT	5156	.015	.699	-5.045	3.966
state pos same TLT2	5155	.016	.698	-5.045	3.966

(4.2)

After several tests for the ARIMAX model with different independent variables, it was concluded that relevant and with the lowest AIC for the model were independent variables VIX, lag of VIX, R_GSPC with its lag, and INFL. The ARIMAX with TAR regression results in output (4.3) reveal significant insights into the dynamics of R_TLT. Incorporating regime-dependent variables based on changes in the Federal Funds Effective Rate (dFEDRATE), the results demonstrate that during periods of negative dFEDRATE (state_neg), the first lag of R_TLT (state_neg_TLT) positively influences current returns ($p < 0.01$), while the second lag (state_neg_TLT) has a significant negative impact ($p < 0.01$). Conversely, during periods of stable or increasing dFEDRATE (state_pos_same), the effects are negative on R_TLT, with state_pos_same_TLT and state_pos_same_TLT2 both being significant ($p < 0.01$ and $p < 0.05$). Market volatility (VIX) has a positive and significant effect on RTLT($p < 0.01$), reflecting increased demand for long-term Treasuries during uncertain periods. In contrast, R_GSPC negatively influences R_TLT ($p < 0.01$), consistent with the flight-to-safety effect. The significant autoregressive (AR) terms confirm strong persistence in R_TLT, with a positive first lag (AR(L) = 1.933, $p < 0.01$) and a negative second lag (AR(L2) = –0.984, $p < 0.01$). The moving average (MA) terms exhibit short-term smoothing, with significant contributions from both the first and second lags (MA(L) = –1.941, $p<0.01$; MA(L2) = 0.992, $p < 0.01$). These findings emphasize the importance of accounting for structural breaks and threshold dynamics driven by FEDRATE changes when modeling R_TLT, highlighting the interaction between monetary policy regimes and Treasury bond returns. This combined model resulted in an AIC equaling 13148.81 which is a significant improvement compared to ARIMA(2,0,2) with an AIC of 13765.91.

Graph (4.4) illustrates the ex-post forecast for 100 observations which resulted in RMSFE equaling 0.9097872. Compared to the ARIMA model, the results show a significant improvement in the magnitude of fluctuations in R_TLT. However, mismatches in direction remain frequent, indicating persistent challenges in accurately predicting certain variations.

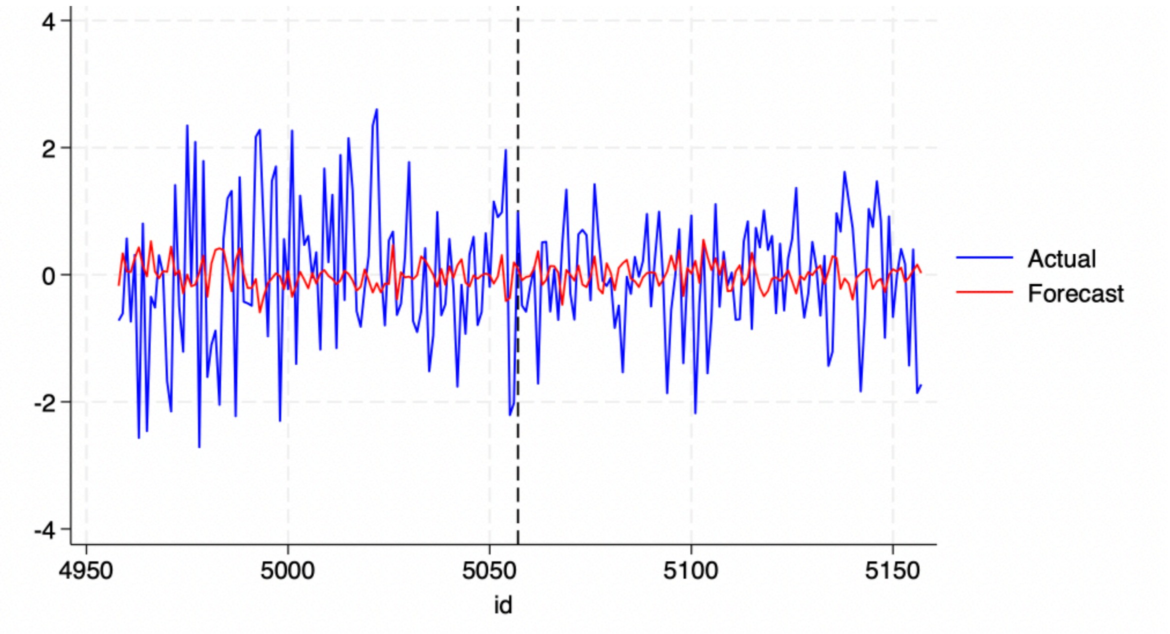
ARIMA regression

R_TLT	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
TAR							
state_neg	.023	.036	0.63	.527	-.048	.094	
state_neg_TLT	.073	.012	5.86	0	.048	.097	***
state_neg_TLT2	-.082	.012	-7.01	0	-.105	-.059	***
state_pos_same	.037	.031	1.21	.228	-.023	.097	
state_pos_same_TLT	-.047	.014	-3.30	.001	-.075	-.019	***
state_pos_same_TLT2	-.03	.015	-2.00	.045	-.06	-.001	**
X-Var							***
VIX	.026	.007	3.86	0	.013	.039	
L	-.024	.006	-3.73	0	-.037	-.011	***
R_GSPC	-.221	.01	-21.90	0	-.241	-.201	***
L	.031	.008	4.10	0	.016	.046	***
INFL	-.014	.005	-2.80	.005	-.025	-.004	***
AR							
L	1.933	.006	299.61	0	1.92	1.945	***
L2	-.984	.007	-151.14	0	-.997	-.971	***
MA							
L	-1.941	.005	-415.11	0	-1.95	-1.932	***
L2	.992	.005	211.05	0	.983	1.002	***
Constant	.864	.006	144.91	0	.852	.875	***

Mean dependent var	0.018	SD dependent var	0.921
Number of obs	5155	Chi-square	1268245.504
Prob > chi2	0.000	Akaike crit. (AIC)	13148.809

****** *p<.01, *** *p<.05, ** *p<.1*

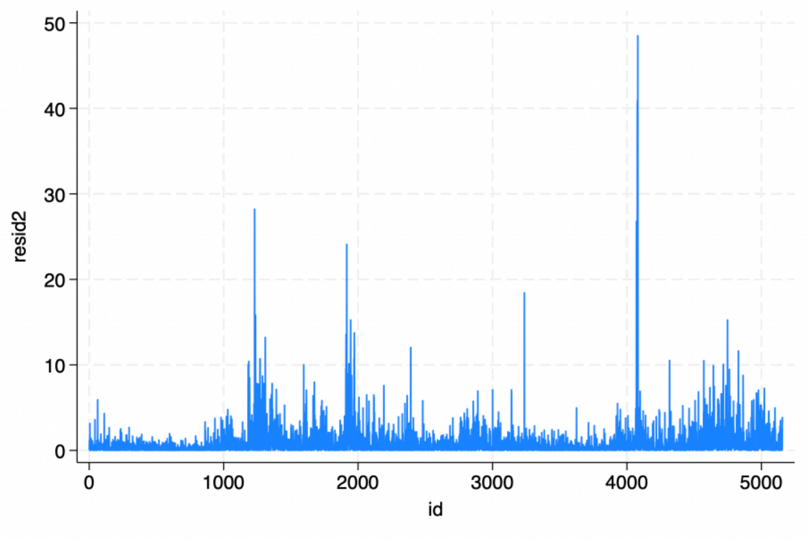
(4.3)



(4.4)

5 GARCH(1,2) with ARIMAX(1,0,3) Model

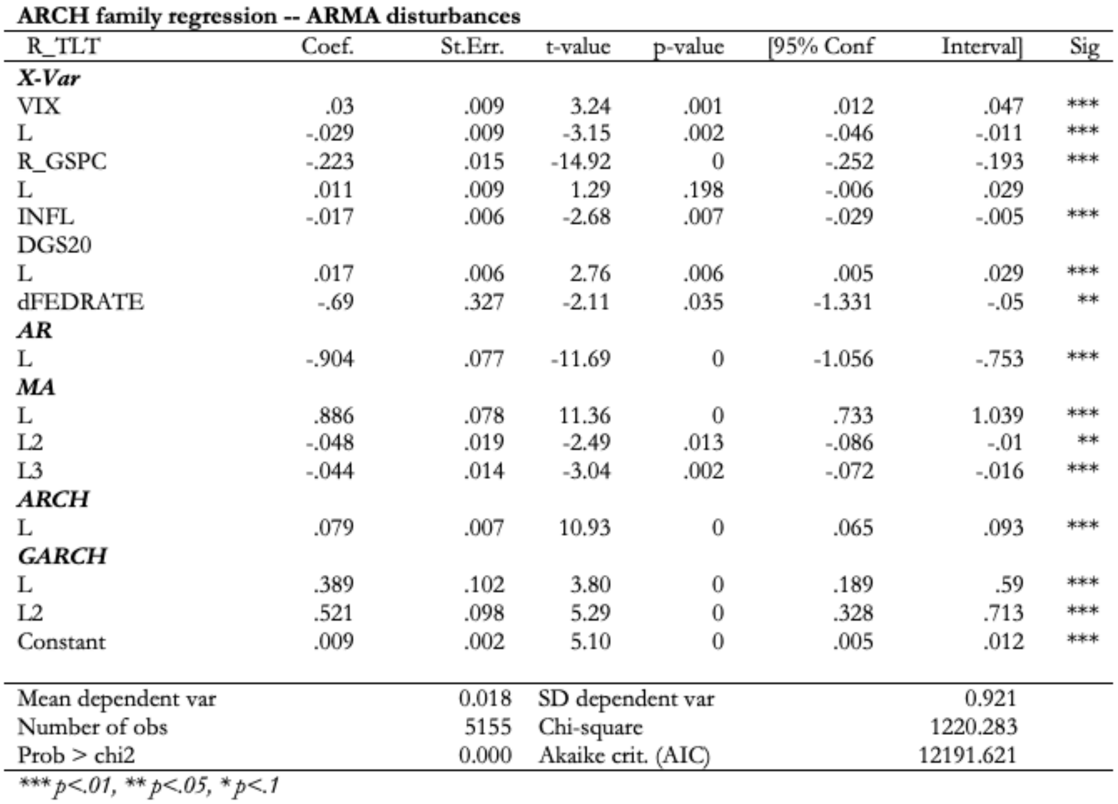
To improve the model, the volatility clustering of R_TLT was investigated and ARCH and GARCH models were applied. The ARCH LM test results showed a highly significant chi-squares statistic of 725.141 with a p-value of 0.0000, indicating strong evidence to reject the null hypothesis of no ARCH effects. This suggests the presence of autoregressive conditional heteroskedasticity (ARCH) in the residuals, confirming volatility clustering in the data. Additionally, in graph (5.1), the squared returns of TLT are plotted, revealing spikes and periods of heightened volatility, this provides evidence of volatility clustering, as periods of high volatility tend to persist over multiple consecutive periods.



(5.1)

These findings influenced the decision to apply the ARCH and GARCH model on top of the previously done but now modified ARIMAX model. After testing all possible combinations of lags for all components optimal model was determined to be ARIMA(1,0,3) with ARCH(1) and GARH(2). Additionally, modifications to the intended variables were made, and optimal were concluded to be VIX and its lag, R_GSPC and its lag, INFL, dFEDRATE, and one lag of DGS20. This model shown in output (4.2) was able to achieve a new and low AIC of 12191.62. In the mean equation, VIX exhibits a positive and statistically significant impact on R_TLT ($p < 0.01$), suggesting again increased demand for long-term Treasuries during periods of heightened market volatility, while the lag of VIX has a small but significant negative effect ($p < 0.01$). R_GSPC has a strong negative influence on R_TLT ($p < 0.01$), consistent with the flight-to-safety behavior, while its lag (L.R_GSPC) shows a marginally significant positive effect (p < 0.05). Both INFL and dFEDRATE are statistically significant (p < 0.01 and p < 0.05, respectively), indicating the role of inflation and monetary policy changes in Treasury bond returns. The first lag of DGS20 is also significant (p < 0.01), reflecting the influence of long-term Treasury yields on returns. The variance equation confirms the presence of volatility clustering. The significant ARCH(1) term (p < 0.01) indicates that recent shocks to the returns, represented by past squared residuals, have an immediate and substantial impact on current volatility. The GARCH terms (GARCH(L) = 0.389, p < 0.01; GARCH(L2) = 0.521, p < 0.01) suggest strong long-term persistence of volatility, where past volatility levels continue to influence current volatility over an extended period. The ARIMA(1,0,3) structure reveals a strong negative autoregressive (AR) component at lag 1 (AR(L) = −0.904, p < 0.01), while the moving average (MA) terms provide significant smoothing effects at lags 1, 2, and 3.

Additionally, slight progress in forecasting efforts was achieved, in graph (5.3) illustrates the ex-post forecast for 100 observations which resulted in RMSFE equaling .90935889. Compared to previous models the biggest progress is seen in more successful prediction of the direction of movement in daily return of TLT.



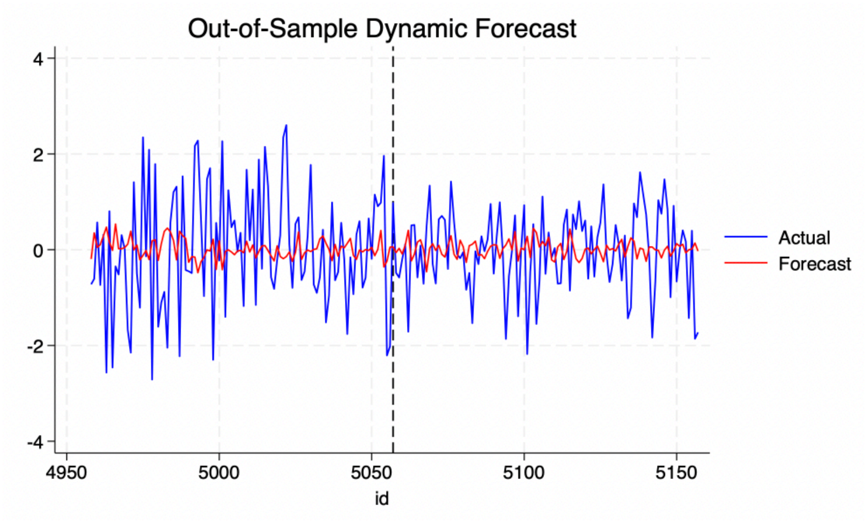
ARCH family regression -- ARMA disturbances

R_TLT	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
X-Var							
VIX	.03	.009	3.24	.001	.012	.047	***
L	-.029	.009	-3.15	.002	-.046	-.011	***
R_GSPC	-.223	.015	-14.92	0	-.252	-.193	***
L	.011	.009	1.29	.198	-.006	.029	
INFL	-.017	.006	-2.68	.007	-.029	-.005	***
DGS20							
L	.017	.006	2.76	.006	.005	.029	***
dFEDRATE	-.69	.327	-2.11	.035	-1.331	-.05	**
AR							
L	-.904	.077	-11.69	0	-1.056	-.753	***
MA							
L	.886	.078	11.36	0	.733	1.039	***
L2	-.048	.019	-2.49	.013	-.086	-.01	**
L3	-.044	.014	-3.04	.002	-.072	-.016	***
ARCH							
L	.079	.007	10.93	0	.065	.093	***
GARCH							
L	.389	.102	3.80	0	.189	.59	***
L2	.521	.098	5.29	0	.328	.713	***
Constant	.009	.002	5.10	0	.005	.012	***

Mean dependent var	0.018	SD dependent var	0.921
Number of obs	5155	Chi-square	1220.283
Prob > chi2	0.000	Akaike crit. (AIC)	12191.621

****** p<.01, ** p<.05, * p<.1*

(4.2)



(4.3)

6 Conclusion

The findings from the two advanced models, ARIMAX with TAR and GARCH(1,2)-ARIMA(1,0,3) with exogenous variables, provide significant insights into the dynamics of R_TLT. The ARIMAX with TAR model revealed the importance of regime-dependent effects, where changes in the Federal Funds Rate (dFEDRATE) significantly impact returns depending on the monetary policy state. During periods of rate cuts, positive influences from lagged returns are offset by subsequent negative effects, while stable or increasing rates have consistently negative impacts. In contrast, the GARCH(1,2)-ARIMA(1,0,3) model highlighted the persistence of volatility clustering, with past shocks and volatility strongly influencing current variance. Both models demonstrated the critical roles of VIX, R_GSPC, INFL, DGS20, and dFEDRATE as key drivers of Treasury bond returns. The variables supported the concept of the flight-to-safety effect as in times of distress in financial markets, low and negative preforms in S&P500 returns (R_GSPC) and high volatility (VIX) return TLT experiences positive returns. Moving to suggestions for further analysis would include additional macroeconomic variables for a broader range of economic cycles. The preference for such assets is heavily influenced by publicly available economic information which is time sensitive, which shapes market sentiment and drives the allocation of assets into long-term Treasury bonds during periods of uncertainty. Furthermore, it would interesting to explore more nonlinear models like regime-switching GARCH or machine learning techniques for improved forecasting accuracy.