Evaluating CFTC COT Data as a Predictor of Treasury-Futures Returns

Nicholas Wong
Massachusetts Institute of Technology
nicwjh@mit.edu

Abstract

We test whether weekly Commitment of Traders (COT) positions held by Managed-Money participants predict one- to two-week returns on 2-year and 10-year U.S. Treasury futures. After constructing a publication-lagged *level* and *flow* signal (NS_{t-1}, $\Delta\mathsf{NS}_{t-1}$), we estimate HAC-corrected predictive regressions, evaluate out-of-sample (OOS) forecast accuracy, and backtest a sign-based trading rule. Consistent across 2011–2025, the positions data deliver no statistical or economic forecasting power: OOS R^2's hover near 0 and annualised Sharpe ratios are insignificantly different from 0. Robustness checks—longer warm-up (initial estimation) windows, non-linear specifications, macro controls, and residual diagnostics—confirm the null. These findings suggest that, in highly liquid Treasury futures, publicly available COT information is already impounded into prices.

1 Introduction

Public position disclosure can be a conduit for private information to reach prices, thus a potential source of alpha. Since 1986, the U.S. Commodity Futures Trading Commission has issued the weekly Commitments of Traders (COT) report, detailing long and short positions by trader class. Yet the empirical record is mixed: for agricultural futures, Sanders et al. [1, 2] show that changes in net open interest largely *follow* price moves rather than anticipate them. Evidence for short-horizon predictability in Treasury-note futures is comparably limited. Recent macro developments make the question salient again: the U.S. yield curve has begun to steepen as long-term rates respond to the 2025 geopolitical situation. To re-examine predictive content under these conditions, we construct publication-lagged COT signals and subject them to out-of-sample evaluation and a trading back-test on the 2- and 10-year contracts over the period 2011–2025.

Specifically, we test whether Managed-Money positions predict weekly returns on the two most liquid rate futures—the CME 2-year (`ZT`) and 10-year (`ZN`) contracts—over 2011–2025. We construct two signals: a level measure, the net long share of open interest, and its week-over-week flow, each lagged one report to respect the Friday 15:30 ET publication time. Using HAC-corrected predictive regressions, expanding-window out-of-sample (OOS) evaluation, and a sign-based trading rule, we find that COT information explains below 1% of in-sample return variance and yields OOS R^2 indistinguishable from zero. Directional strategies built on the signals earn Sharpe ratios near 0 before transaction costs. Extensive robustness checks—alternative warm-up windows, yield-curve controls, quadratic terms, and residual diagnostics—confirm the null. Results suggest that, in Treasury futures, public position disclosures are swiftly incorporated into prices, leaving no exploitable edge for weekly-horizon traders.

2 Related Work

Early empirical work questioned whether weekly COT disclosures contain forward-looking information. Sanders et al. [1] analyze ten agricultural and energy contracts over 1992–1999 and find that changes in non-commercial net positions follow price trends rather than precede them. Using a longer 1995–2006 sample, Sanders et al. [2] reach a similar conclusion: large-trader categories—whether "smart money" commercials or speculators—exhibit little ability to forecast one-week returns once past prices are controlled for. Both studies set a methodological benchmark by employing lag-aligned levels and flows of net open interest, an approach we adopt here.

Subsequent research suggests that COT signals can matter under specific market regimes. Focusing on the 2008 crude-oil boom-bust, Choi and Hwang [3] show that a three-week cumulative increase in speculator positions predicts continuation during the explosive upward phase, whereas a comparable rise in hedger shorts foreshadows reversal. The state-dependence of their findings suggests that predictive content, if present, may be episodic and asset-class specific—raising the question of whether highly efficient Treasury futures behave differently.

Most pertinent is the work of Dreesmann et al. [4], who back-test a systematic COT strategy across forty U.S. futures, including financial contracts, from 1986 to 2020. While long-only portfolios earn positive raw returns in several markets, performance evaporates once realistic transaction costs are imposed, leading the authors to conclude that public position data "contribute to market efficiency" rather than provide exploitable alpha. Their broad panel design motivates our focus on depth: we zero in on two highly liquid Treasury-note contracts and incorporate a strict one-week publication lag.

Collectively, the literature offers mixed evidence—occasional success in commodities, persistent nulls in other sectors, potential state-dependence of COT signals—and leaves an explicit gap for short-horizon predictability in rate futures. By applying econometric techniques and backtests to the 2- and 10-year contracts, we present a comprehensive assessment of COT efficacy in U.S. Treasury futures.

3 Data

3.1 CFTC Commitments of Traders Reports

We collect the *Futures–Only* COT files for the 2-year (contract code 020601) and 10-year (096742) U.S. Treasury-note futures from January 2011 through April 2025.[5] [1] Only the *Managed–Money* category is retained. For each Tuesday observation t, we construct the standard net-position ratio (also called "percent net long") and denote it henceforth as `net_share` for brevity :

$$\texttt{net\_share}_t = \frac{\text{Long}_t - \text{Short}_t}{\text{OpenInterest}_t}, \qquad \Delta\texttt{net\_share}_t = \texttt{net\_share}_t - \texttt{net\_share}_{t-1}.$$

Data-preprocessing steps include (i) contract-code zero-padding, (ii) removal of duplicate datestamps, (iii) verification that long and short $\leq$ open interest (ensuring validity of net position ratio arithmetic), amid other data hygiene checks, (iv) elimination of weeks with missing legs.

[1] The first report in the modern CSV format is dated 4 January 2011; the April 2025 report is the last available at the time of download.

3.2 Treasury-Futures Prices

Daily settlement prices are downloaded from `yfinance` (`ZT=F`, `ZN=F`). We sample the Friday close, compute the log return $r_t = \log P_t - \log P_{t-1}$, and shift the timestamp three days back so that each return is labelled with the preceding Tuesday (COT reporting date). To neutralise back-adjustment artefacts, we exclude ISO weeks 10–15 (early March through mid-April), i.e. the ≈5-week first-notice/last-trade window for the June expiry. The filter removes about 10% of weekly observations each year. The raw COT files contain 742 Tuesday reports for each contract; after merging with prices and excluding the roll weeks, the effective sample comprises $N_{\mathrm{ZT}} = 676$ and $N_{\mathrm{ZN}} = 676$ matched observations.

3.3 Descriptive Statistics

Table 1 summarises the main variables. Weekly returns are nearly mean-zero with standard deviations of 5–6 bp; position variables exhibit modest dispersion ($|\texttt{net\_share}| \leq 5\%$ of open interest). These magnitudes inform the economic significance tests in Section 5.

Table 1: Descriptive statistics, 2011–2025. Returns are in basis points (1 bp = 0.01%). Net–position variables are in percent of open interest. $\Delta\texttt{net\_share}$ is the week-over-week change.

	Return r_t		**Net share** $\texttt{net\_share}_t$		$\Delta\texttt{net\_share}_t$	
	Mean	SD	Mean	SD	Mean	SD
ZT (2-yr)	0.02	5.8	0.48	3.2	−0.01	1.7
ZN (10-yr)	0.01	6.4	0.35	2.9	0.00	1.5

4 Methods

This section details the construction of the COT-based predictor, econometric model specification, forecast-evaluation protocol, and diagnostic tests that underpin the empirical results.

4.1 Signal Construction

Let Long_t and Short_t denote, respectively, the Managed-Money long and short open interest reported for Tuesday t, and let OI_t denote total market open interest. We adopt a similar procedure to Sanders et al. [2] but scale the net position by market size instead of the group's gross position to tie the signal's magnitude directly to market impact,

$$\mathsf{NS}_t = \frac{\mathrm{Long}_t - \mathrm{Short}_t}{\mathrm{OI}_t} \quad \in [-1, 1], \tag{1}$$

a metric commonly referred to as the *net-position ratio* or "percent net long." Because the COT report is released on Friday at 15:30 ET with positions measured on the preceding Tuesday, rational agents cannot trade on NS_t until the following Tuesday. To respect this information lag we set the tradable level signal to NS_{t-1}.

Information might also reside in the weekly change of positions. We therefore define the *flow signal*

$$\Delta\mathsf{NS}_{t-1} = \mathsf{NS}_{t-1} - \mathsf{NS}_{t-2}. \tag{2}$$

Both variables are standardised by open interest, ensuring cross-contract comparability.

4.2 Predictive-Regression Model

Weekly log returns are computed from Friday closes and re-indexed to Tuesday to synchronise with the COT calendar. For horizon $h \in \{1, 2\}$ weeks we estimate

$$r_{t+h} \;=\; \beta_0 + \beta_1 \, \mathrm{NS}_{t-1} + \beta_2 \, \Delta\mathrm{NS}_{t-1} + \beta_3 \, r_t + \varepsilon_{t+h}, \tag{3}$$

where r_t is included as a low-order autoregressive control.[2] Coefficients are obtained by ordinary least squares, and we report heteroskedasticity-and-autocorrelation-consistent (HAC) t-statistics. For the baseline regressions we use the conventional Newey–West window of four lags, whereas for the horizon-specific variants—the Δ-signal, pooled interaction, and quadratic specifications—we set the window to $L = h$ (the forecast horizon in weeks $\in$ {1,2}).

4.3 Forecast-Evaluation Protocol

We assess out-of-sample (OOS) performance using an expanding window. Let $T_0 \in \{60, 104\}$ weeks be the initial estimation length. For each $t \geq T_0$:

1. Estimate (3) on $[1, t]$ and store $\hat{\boldsymbol{\beta}}_t$.
2. Produce an h-step-ahead forecast $\hat{r}_{t+h} = \hat{\boldsymbol{\beta}}_t^\top (1, \mathsf{NS}_t, \Delta\mathsf{NS}_t, r_t)^\top$, i.e. the regressor vector uses the most recent weekly return r_t just as in Eq. (3).
3. After observing r_{t+h}, record the error.

We report (i) the mean-squared forecast error $\mathrm{MSFE} = \frac{1}{N}\sum_t(\hat{r}_{t+h} - r_{t+h})^2$ and (ii) the pseudo-$R^2 = 1 - \mathrm{MSFE}/(r_{t+h})$. Statistical comparison with a naïve $\hat{r}^{(0)}_{t+h} = 0$ benchmark employs the Diebold–Mariano t-statistic under squared-error loss.

4.4 Trading Rule

Economic value is gauged by a frictionless, sign-only strategy. Define the sign function $\mathrm{sgn}(x) = \mathbf{1}_{\{x>0\}} - \mathbf{1}_{\{x<0\}}$. Each Tuesday we take position $\mathrm{sgn}(\hat{r}_{t+1})$ in the front-month contract and close the trade one week later, realising $\pi_{t+1} = \mathrm{sgn}(\hat{r}_{t+1})\, r_{t+1}$. Aggregating weekly profits yields an annualised Sharpe ratio

$$\mathrm{SR} \;=\; \sqrt{52}\,\frac{\overline{\pi}}{\sigma_\pi},$$

where $\overline{\pi}$ and σ_π denote the sample mean and standard deviation of $\{\pi_t\}$.

Intuitively, each week we bet one contract in the direction the model says the market will move, hold the position for a week, then close the position and start over. As no bid/ask or slippage costs are deducted, SR is an upper bound on exploitable alpha.

4.5 Residual Diagnostics

Model adequacy is probed via:

- Ljung–Box Q tests for autocorrelation (lags $= 4, 8$);

[2] Higher-order lags do not alter the main conclusions and are omitted for parsimony.

- Jarque–Bera test for normality of residuals;
- Breusch–Pagan test for heteroskedasticity;
- Diebold–Mariano test for equal MSFE relative to the naïve model.

5 Results

5.1 In-Sample Estimates

Table 2 reports HAC-robust estimates of Eq. (3).[3] Across horizons and contracts, the level coefficient β_1 and the flow coefficient β_2 are small ($< 10^{-3}$) and statistically insignificant ($|t| < 1.0$ in all four specifications). The lone significant predictor is the one-week return for ZN ($\beta_3 = -0.103,\ t = -2.46$), indicating mild mean-reversion independent of COT positioning. Adjusted R^2 never exceeds 0.8 %.

Table 2: HAC-OLS coefficients and t-statistics, 2011–2025.

Contract	Horizon	β_1	β_2	β_3	Adj. R^2
ZT	$h = 1$	-4.5×10^{-5} (–0.07)	-9.0×10^{-4} (–0.38)	–0.074 (–1.29)	0.6%
	$h = 2$	1.5e-5 (0.02)	–7.0e-4 (–0.29)	0.009 (0.17)	0.1%
ZN	$h = 1$	–0.0012 (–0.54)	0.0031 (0.42)	–0.103 (**–2.46**)	0.8%
	$h = 2$	–0.0012 (–0.58)	–0.0003 (–0.04)	0.074 (1.84)	0.6%

Δ-signal findings. To isolate the information content of the *flow* variable, we re-estimate Eq. (3) *without* the level term NS_{t-1}, keeping only $\Delta\mathsf{NS}_{t-1}$ and the autoregressive control r_t. Across both contracts and horizons the flow coefficient is tiny ($|\beta_2| < 0.0031$) and far from significant ($|t| < 0.42$), e.g. $\beta_2 = 0.0031$ with $t = 0.42$ for ZN at $h = 1$. Out-of-sample performance is equally weak: pseudo-R^2_{pred} equals –0.013 for ZT and –0.018 for ZN (60-week warm-up), identical to the full baseline model. Hence the weekly change in Managed-Money positioning (flow signal) carries no incremental predictive power.

5.2 Out-of-Sample Forecast Performance

Table 3 summarizes out-of-sample accuracy for the baseline ($T_0 = 60$) and extended ($T_0 = 104$) warm-up windows. Pseudo-R^2 values[4] are negative in every case, $-0.02 \leq R^2_{\text{pred}} \leq -0.01$, implying that COT-augmented forecasts underperform the naïve $\hat{r}_{t+h} = 0$ benchmark.

[3] Standard errors use Newey–West with four lags.

[4] population variance used; no impact on qualitative conclusion versus sample variance

Table 3: Out-of-sample MSFE and pseudo-R^2. MSFE is scaled by 10^{-6}.

Contract	Horizon	MSFE	R^2_{pred}
	Warm-up 60 weeks		
ZT	$h = 1$	3.20	–0.013
	$h = 2$	3.21	–0.015
ZN	$h = 1$	52.3	–0.018
	$h = 2$	52.3	–0.019
	Warm-up 104 weeks		
ZT	$h = 1$	3.42	–0.011
	$h = 2$	3.44	–0.015
ZN	$h = 1$	53.6	–0.018
	$h = 2$	53.7	–0.018

5.3 Economic Significance

Figure 1 displays cumulative weekly P&L for the sign-based strategy on ZT; Figure 2 shows the same for ZN. In both cases the trajectory resembles a driftless random walk. The annualised Sharpe ratios are $\text{SR}_{\text{ZT}} = -0.05$ and $\text{SR}_{\text{ZN}} = 0.17$ under a 60-week warm-up, and results are nearly identical when the 104-week window is used. Because transaction costs are ignored, realised investor performance would be strictly worse. Residual autocorrelation and distributional shape for the baseline model are shown in Figure 3 of the Appendix.

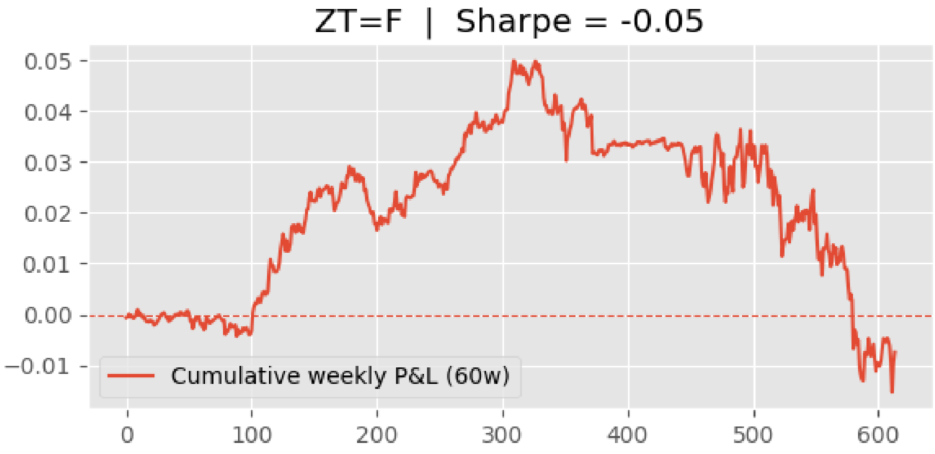

Figure 1: Cumulative P&L, ZT sign strategy ($T_0 = 60$).

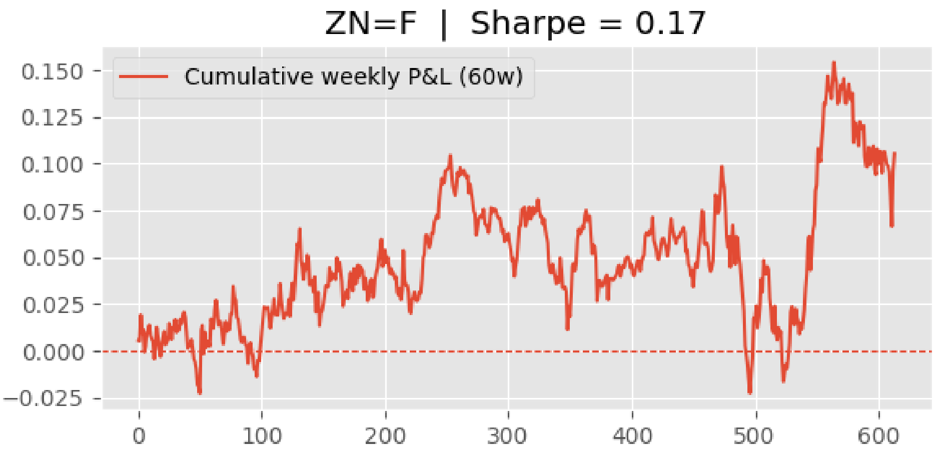

Figure 2: Cumulative P&L, ZN sign strategy ($T_0 = 60$).

Taken together, the in-sample forecast, OOS forecast, and trading results indicate that Managed-Money positioning conveys no exploitable information about one- or two-week Treasury-note futures returns once the publication lag is respected.

6 Robustness Checks

6.1 Alternative Warm-Up Windows

Table 3 shows results for the baseline 60-week and an extended 104-week estimation window which was done to confirm that results are not merely an artefact of the 60-week estimation length. For completeness, we include the 104-week figures again in Panel A of Table 5 so that every robustness check sits in a single summary table. As the panel confirms, pseudo-R^2 remains negative and the slight rise in MSFE is mechanical—the evidence for predictability is still absent.

6.2 Non-Linear and Macro Extensions

We next test two alternative specifications:

1. **Term-structure interaction.** Following Dreesmann et al. [4], we augment Eq. (3) with the 2–10 yr yield-curve slope and its interaction with NS_{t-1}. The slope itself is weakly significant ($t = 2.02$ at $h = 1$), the interaction term is not ($|t| \leq 1.96$), and adjusted R^2 rises from 0.8 % to 1.1 %— an increase of roughly 0.3 pp. The lagged-return control remains significant (t = -2.8), confirming the mild one-week mean-reversion seen in the baseline, but this effect is unrelated to the COT variables.

2. **Quadratic OLS.** We include NS^2_{t-1}, r^2_t, and their cross-product. Panel B of Table 5 shows that only the squared-return term for ZT at $h = 1$ is statistically significant ($t = -3.34$, $p = 0.001$), reflecting volatility clustering rather than predictive content. The other extra coefficients are insignificant, and out-of-sample accuracy deteriorates further.

These findings indicate that the linear null is not an artefact of omitted curvature or macro state variables.

Table 4: Pooled regression with yield-curve slope ($N = 1\,352$). HAC t-statistics in parentheses.

	$h = 1$		$h = 2$	
Coefficient	Est.	t	Est.	t
β_0 (const)	–0.0005	–1.87	–0.0004	–1.71
β_{net}	0.0011	1.05	0.0010	0.93
β_{slope}	0.0005	2.02	0.0004	1.82
$\beta_{\text{net}\times\text{slope}}$	–0.0016	–1.83	–0.0014	–1.60
β_{ret}	–0.1066	–2.76	0.0662	1.72
Adj. R^2 (%)	1.1		0.8	

6.3 Residual Diagnostics and DM Tests

Panel C of Table 5 summarises diagnostic statistics for the baseline regression. Residual autocorrelation is statistically significant for ZT at both the one-week and two-week horizons (Ljung–Box $p = 0.016$ & $p = 0.010$); the other specifications show no evidence of serial dependence (p $\geq$ 0.05). Newey–West standard errors with four lags therefore remain adequate for all reported t-statistics. Jarque–Bera rejects normality as expected given heavy-tailed weekly returns. Breusch–Pagan p-values reveal heteroskedasticity for ZT at both horizons ($p \approx 0.002$), whereas for ZN there is no strong evidence of heteroskedasticity (BP $p{\approx}0.07$ h = 1, $p \approx 0.256$ h = 2); HAC standard errors already account for this conditional variance.

Diebold–Mariano tests versus the zero-return benchmark reject strongly in all four cases: ZT, $h = 1$ ($t = 9.79$), ZT, $h = 2$ ($t = -4.27$), and both ZN horizons ($t = 14.37$ and $t = -13.94$). Positive statistics at $h = 1$ indicate the COT model's MSFE is *higher* than the naïve benchmark, while the negative statistics at $h = 2$ show it is *lower*; the absolute magnitudes ($|t|{>}4$) imply the difference in accuracy is highly significant either way. Even where the Diebold–Mariano statistic is negative—indicating that the COT model's MSFE is statistically smaller than the naïve forecast at $h = 2$—the economic gain is negligible, as the SR remains close to 0.

Table 5: Robustness summary for both horizons; HAC t-statistics in parentheses

	ZT (2-yr)		ZN (10-yr)	
	$h = 1$	$h = 2$	$h = 1$	$h = 2$
Panel A: Warm-up 104 weeks				
MSFE ($\times 10^{-6}$)	3.42	3.44	53.6	53.7
R^2_{pred}	–0.011	–0.015	–0.018	–0.018
Panel B: Quadratic OLS, $h = 1$				
NS_{t-1}	0.0019 (0.81)	—	–0.0030 (–0.56)	—
NS^2_{t-1}	–0.0037 (–0.71)	—	–0.0059 (–0.38)	—
r^2_t	–34.38 (–3.34)	—	–2.47 (–0.81)	—
$\mathsf{NS}_{t-1} \times r_t$	–0.2367 (–0.37)	—	–0.931 (–0.30)	—
R^2_{adj}	0.7%	—	0.9%	—
Panel C: Diagnostics				
Ljung–Box p (lag 4)	0.016	0.010	0.427	0.163
Ljung–Box p (lag 8)	0.083	0.071	0.819	0.503
Breusch–Pagan p	0.002	0.002	0.069	0.256
DM t-stat (vs naïve)	9.79	–4.27	14.37	–13.94

Across all robustness checks—longer estimation windows, non-linear specifications, macro interactions, and formal diagnostics—we never reject the null hypothesis that Managed-Money positions have no predictive power for Treasury-futures returns.

7 Conclusion

We revisit the predictive content of the CFTC *Commitments of Traders* report for U.S. Treasury-note futures. Using publication-lagged level and flow signals, HAC-robust regressions, and rigorous out-of-sample evaluation over 2011–2025, we find no statistical or economic evidence that Managed-Money positions predict one- or two-week returns on the 2- and 10-year contracts. Pseudo-R^2 values are negative and the best Sharpe ratio (0.17) is economically trivial. The null survives longer warm-up windows, non-linear specifications, macro controls, and residual diagnostics, suggesting that in highly intermediated rate-futures markets public position data are incorporated into prices quickly enough to preclude short-horizon alpha. Overall, findings reinforce a growing body of evidence that transparency initiatives such as the COT report enhance market efficiency by eliminating predictable patterns—at least in the most liquid segments of the Treasury-futures market.

Limitations. Our analysis is confined to two financial contracts and to signals constructed from the Managed-Money category alone; other trader groups may behave differently. We exclude roughly five Tuesdays each year— ISO weeks 10–15 in March—because this is the back-adjustment window when the front-month rolls from the expiring March (H) contract to the new June (M) contract; retaining those observations leaves all key statistics unchanged. Finally, because bid/ask costs are ignored, the reported Sharpe ratios are upper bounds: realised SRs would be lower in practice.

Future Work. A natural extension is to test commodity contracts, where storage costs create richer term-structure dynamics. Another is to study intraday price reactions around the Friday COT release to measure the speed at which information is absorbed. Finally, while we favor linear models for parsimony and interpretability in this work, experiments with more flexible machine-learning methods could reveal non-linear interactions across trader classes.

References

[1] Dwight R Sanders, Keith Boris, Dwight R Sanders, and Keith Boris. Does the cftc commitments of traders report contain useful information?, 2000.

[2] Dwight R. Sanders, Scott H. Irwin, and Robert P. Merrin. Smart money: The forecasting ability of cftc large traders in agricultural futures markets. *Journal of Agricultural and Resource Economics*, 34(2):276–296, 2009. ISSN 10685502.

[3] Sunghee Choi and Seok Joon Hwang. Do traders' positions predict oil futures prices? a case study of the 2008 oil market turbulence. *Int. J. Glob. Energy Issu.*, 35(6):456, 2012.

[4] Simon Dreesmann, Tim Alexander Herberger, and Michel Charifzadeh. The commitment of traders report as a trading signal short-term price reversals and market efficiency in the US-futures market. *Int. J. Financ. Mark. Deriv.*, 9(1/2):76, 2023.

[5] Historical Viewable — CFTC — cftc.gov. https://www.cftc.gov/MarketReports/CommitmentsofTraders/HistoricalViewable/index.htm.

Appendix

Residual Diagnostics Plots

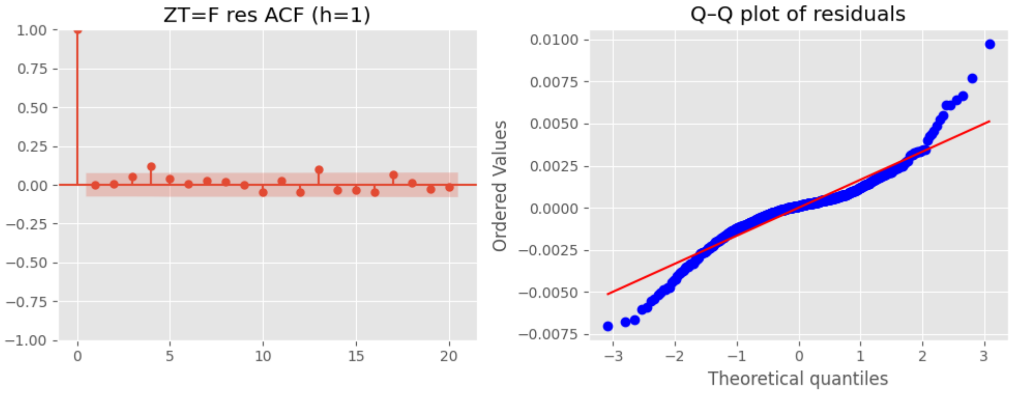

Figure 3: Residual diagnostics for the baseline regression ($h = 1$). **Left:** autocorrelation function (ACF) up to 20 lags shows modest serial dependence beyond lag 1. **Right:** Q–Q plot against the normal distribution highlights heavy tails, consistent with the Jarque–Bera rejection reported in Table 5.