Financial Econometrics Homework 1

for Professor Nour Meddahi

by Hafez, Köhler, Suero - 2024/25

February 14, 2025

1 Descriptive Statistics

a) Sample Statistics

Having a look at some sample statistics, we find that the mean of the returns is very close to zero. Moreover, the standard deviation dominates the mean. Negative skewness indicates the left tail of the distribution is fatter than the right one - that is, extreme negative returns happen more frequently than extreme positive ones. Finally, the high value for kurtosis suggests that the distribution has heavier tails than the normal (whose kurtosis value is 3).

Table 1: Sample statistics of log-returns

Metric	**Returns**
Mean	0.000437
Variance	0.000127
Skewness	-0.816517
Kurtosis	15.792357

When plotting returns of the S&P500 over time, we observe how they fluctuate around zero and undergo periods of high volatility that coincide with major economic shocks e.g. Covid-19.

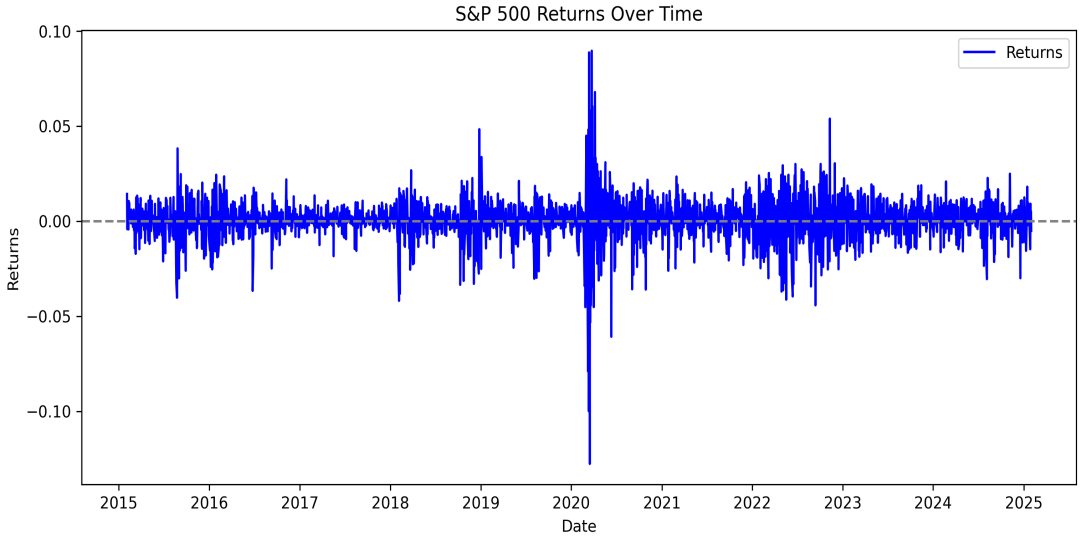

Figure 1: S&P500 log returns over time

Squared returns showcase periods of volatility clustering and persistence, followed by mean rever-

sion. This is a measure of variance, and the fact that it displays correlation with its own past points to models like GARCH as a promising option to capture such volatility dynamics.

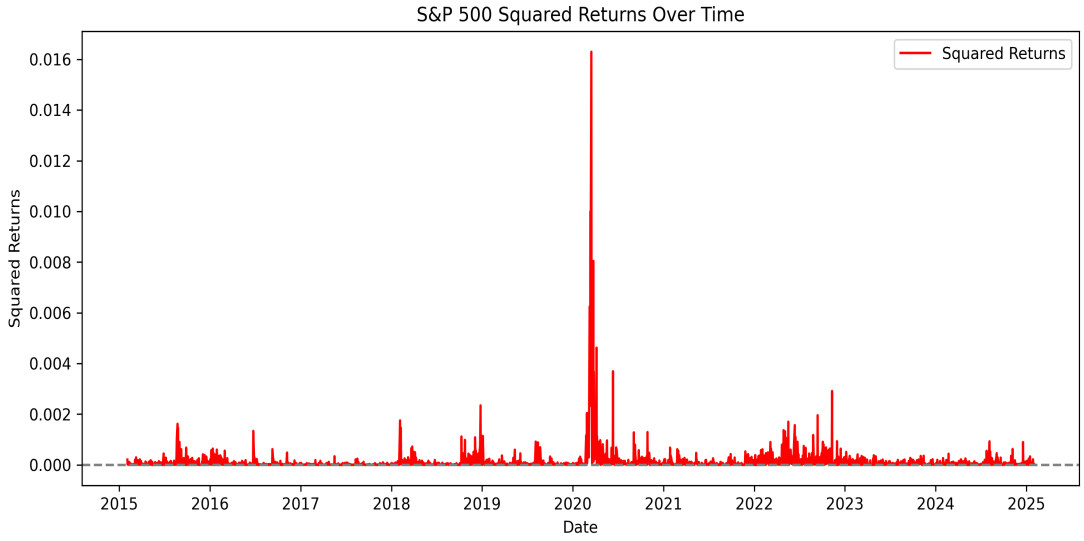

Figure 2: S&P500 squared log returns over time

In the ACF of mean returns we can observe how values revert to the mean. Moreover, they are often low, pointing to low serial correlation.

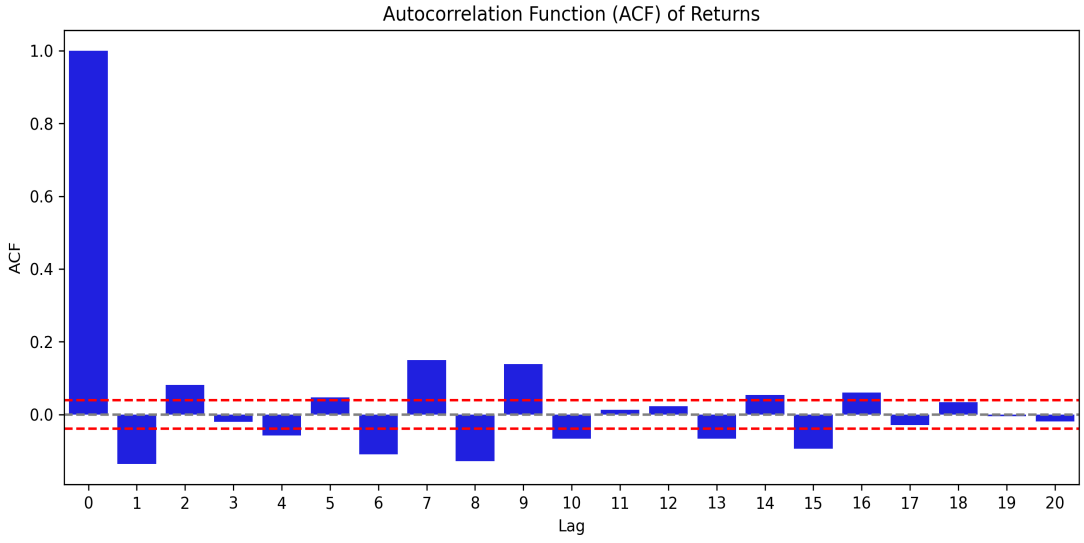

Figure 3: ACF log-returns

However, in the ACF of squared returns we can indeed notice the persistence in volatility, with ACF being positive for all lags and decaying at a lower rate.

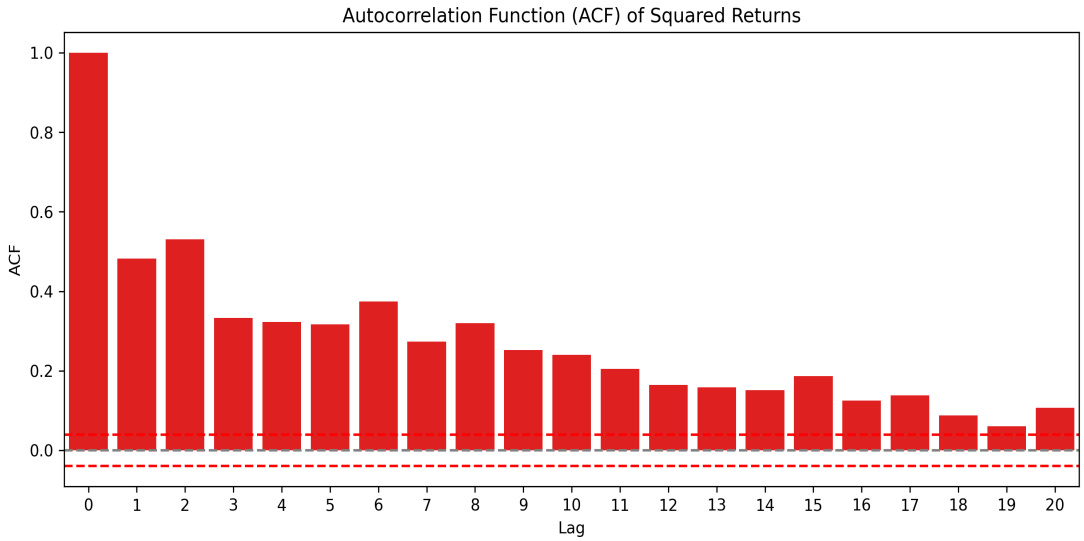

Figure 4: ACF squared log-returns

b) Rolling Window

Below we plot the rolling mean, variance, skewness and kurtosis for the S&P500 chosen period. Choosing an appropriate window size entails evaluating the following trade-off: capturing data variability while preserving some degree of smoothness. For very large rolling window sizes e.g., 3 months, the effect of events such as Covid-19 spills over to neighboring periods. As we decrease it, volatility becomes more concentrated. However, skewness and kurtosis become highly variable. We have decided to strike a balance by presenting our graphs with a rolling window of 60 days.

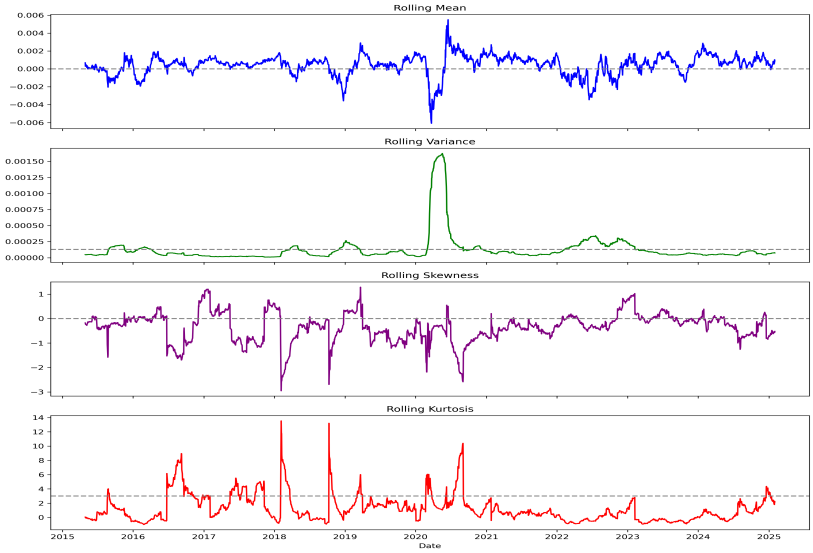

Figure 5: Sample Moments with Rolling Windows

2 GARCH

a) GARCH using simulated data and $z_t \sim N(0,1)$

In this exercise, we consider the following GARCH model:

$$R_t = \mu + \sigma_t z_t, \quad z_t \sim N(0,1),$$

$$\sigma_t^2 = \omega + \alpha(R_{t-1} - \theta\,\sigma_{t-1})^2 + \beta\,\sigma_{t-1}^2.$$

We have two sample sizes, $T = 250$ and $T = 1000$. Moreover, we focus on two designs:

- 1: $\psi = (0,\ 0.01,\ 0.05,\ 0.90,\ 0)$ (no leverage effect)
- 2: $\psi = (0,\ 0.01,\ 0.05,\ 0.6,\ 0.2)$ (with leverage effect)

For each combination of design and sample size, we run 1000 simulations (after discarding an initial burn–in of 250 observations). We estimate the parameters with maximum likelihood for each simulation and compute sample statistics across all the simulations. We start by simulating synthetic GARCH(1,1) time series data. We set certain constraints in the initial parameter values i.e., $\alpha + \beta < 1$ or non-negativity to ensure stationarity, numerical stability and convergence. We initialize σ^2 according to such conditions and generate returns recursively. Then, we define the negative log-likelihood as:

$$\mathcal{L} = -\frac{1}{2}\sum_{t=1}^{n}\left(\log(2\pi) + \log(\sigma_t^2) + \frac{(R_t - \mu)^2}{\sigma_t^2}\right).$$

We maximize the log-likelihood with an optimization algorithm to estimate the parameters for each design and sample size. The results below enable us to evaluate to what extent each specification recovers the true parameters. We observe that as sample size increases, the standard deviation of the estimates decreases and the parameters increasingly converge to the true values.

Parameter	Mean	Std Dev
μ	0.0009	0.0290
ω	0.0321	0.0395
α	0.0527	0.0513
β	0.7228	0.2651
θ	0.0288	3.7653

Table 2: No leverage, $T = 250$, Distribution: normal

Parameter	Mean	Std Dev
μ	0.0001	0.0142
ω	0.0164	0.0173
α	0.0530	0.0238
β	0.8590	0.1163
θ	0.0604	1.0127

Table 3: No leverage, $T = 1000$, Distribution: normal

Parameter	Mean	Std Dev
μ	-0.0004	0.0113
ω	0.0124	0.0073
α	0.0558	0.0590
β	0.4558	0.2863
θ	0.5569	3.0064

Table 4: Leverage, $T = 250$, Distribution: normal

Parameter	Mean	Std Dev
μ	0.0002	0.0054
ω	0.0121	0.0070
α	0.0528	0.0325
β	0.5091	0.2603
θ	0.2870	1.4783

Table 5: Leverage, $T = 1000$, Distribution: normal

Given the large uncertainty surrounding our θ parameter estimation with a standard deviation of 1.4783 we plot the distribution of the parameter estimates (Figure 6). Whilst the distribution is centered around 0.2, the true parameter value, considerable mass is still around 0.2.

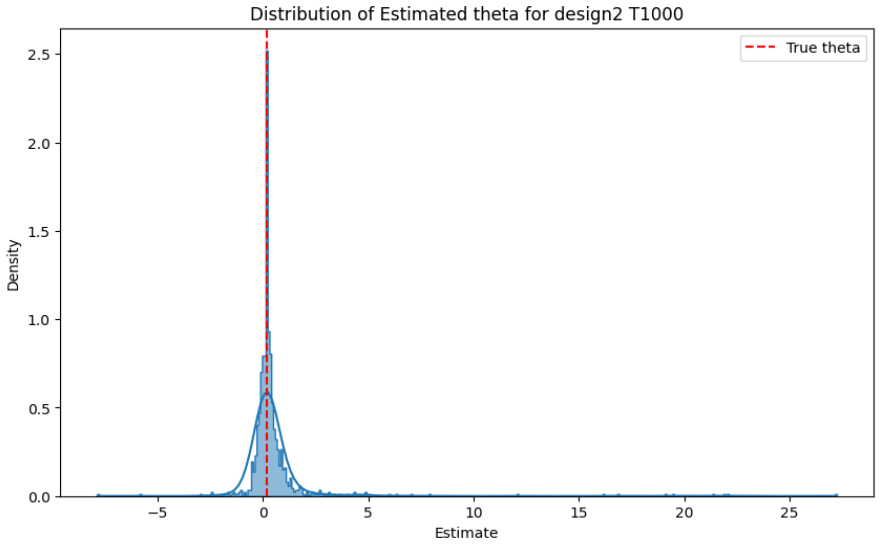

Figure 6: Distribution of estimate θ

b) GARCH using simulated data and $z_t \sim t(10)$

Now we simulate the data for each combination of design ψ and sample size T, but we assume that the innovations are drawn from a standardized $t(10)$ distribution, as opposed to a normal. A standardized $t(10)$ random variable has variance 1 and is defined as

$$t(10) = \frac{X}{\sqrt{Y/10}}, \quad X \sim N(0,1), \quad Y \sim \chi^2(10).$$

Even though asymptotically MLE should yield similar estimates under normal and t-distribution

assumptions, it is preferable not to impose such distributional restrictions. Hence, we perform quasi-maximum likelihood estimation (QMLE) on the simulated data with the assumption $z_t \sim t(10)$ but keep the same likelihood function as in the previous section. QMLE is more robust in the sense that even if the likelihood is misspecified, it provides consistent parameter estimates and hence more reliable inference. In the tables below, we again observe convergence as the sample size increases.

As in the normally distributed simulated data, we observe large standard errors in the estimation of our θ parameter

Parameter	Mean	Std Dev
μ	0.00487006	0.03242044
ω	0.03665555	0.05499183
α	0.0698504	0.07304232
β	0.76314165	0.22173932
θ	-0.12917212	4.75566296

Table 6: No leverage, $T = 250$, QMLE

Parameter	Mean	Std Dev
μ	0.00198074	0.01962976
ω	0.01714407	0.0108924
α	0.06656391	0.02211894
β	0.87872866	0.04577145
θ	-0.02438437	0.27893181

Table 7: No leverage, $T = 1000$, QMLE

Parameter	Mean	Std Dev
μ	-0.00257592	0.01925766
ω	0.0158909	0.01298646
α	0.06942029	0.067069
β	0.3881915	0.31604122
θ	3.45116407	3.79441257

Table 8: Leverage, $T = 250$, QMLE

Parameter	Mean	Std Dev
μ	-0.00032410	0.00925793
ω	0.01314506	0.00451379
α	0.06440765	0.04355851
β	0.53888886	0.16290869
θ	2.35054314	1.8072744

Table 9: Leverage, $T = 1000$, QMLE

As in the normally distributed case, we again observe a high standard deviation for the parameter θ, we omit plotting the distribution as it is large identical with Figure 6, except that we observe "fatter tails" of the distribution and a higher standard deviation compared to the designs simulated with a normal distribution, due to the $T(10)$ distribution.

c) GARCH using daily returns

We now estimate the GARCH model with and without leverage effect by using the S&P 500 daily log-return data. We can observe some significant differences between both models:

Table 10: Model Parameters GARCH with and without Leverage

Parameter	No leverage	Leverage
μ	0.080367	0.047663
ω	0.038799	0.038115
α	0.180367	0.060405
β	0.789102	0.801352
θ	—	0.211094

We perform a likelihood ratio (LR) test to contrast the goodness of fit of both models. The null hypothesis H_0 is that the no-leverage model is sufficient. The alternative hypothesis is that the model with leverage significantly improves the model fit. The test statistic is calculated as:

$$LR = 2\left(\ell_{\text{lev}} - \ell_{\text{nolev}}\right),$$

where ℓ_m is the log-likelihood for a given model. Under H_0, the test statistic $LR \sim \chi^2(1)$ (since

we are only testing one parameter). With a p-value of 0.0000, we reject the null hypothesis at the 1% confidence level. This suggests that the leverage effect is significant and improves the model fit.

Table 11: Likelihood Ratio Test Results

Statistic	**Value**
LR Statistic	61.1069
p-value	0.0000

Finally, we conduct the Ljung-Box test to the residuals so check whether there is any remaining autocorrelation in such. If there is, then the model GARCH (1,1) with leverage is not fully able to capture volatility dynamics. The null hypothesis is that there is no autocorrelation in squared returns while the alternative hypothesis is that there is serial correlation. Since all p-values are essentially zero, we reject the null hypothesis at all lags, with the first lag test statistics being shown below. This suggests that there is significant autocorrelation in the squared residuals, meaning that the GARCH model may not have fully captured the conditional heteroskedasticity in the data.

Ljung-Box Stat	p-value
579.9611	0.0000

Table 12: Ljung-Box Test Results on Squared Standardized Residuals

d)

We now introduce the RiskMetrics variance, which updates the variance in a recursive fashion through the following expression:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)\varepsilon_{t-1}^2, \tag{1}$$

where λ is the decay factor and is set to $\lambda = 0.94$ for daily data, and ε_{t-1} is the innovation or return shock. We compare the GARCH with leverage to the Riskmetrics variance. Both are highly correlated, with correlation coefficient $\rho = 0.79$, indicating that both model capture roughly the same volatility patterns. However, note that a correlation measure is unable to capture differences in scale or levels.

After plotting, we reveal finer dynamics between both models. It seems that Riskmetrics provides a smoothed estimate (lower standard deviation). That is, it could underreact to abrupt spikes and

sudden economic shocks. This is linked to the fact that the decay factor is fixed at $\lambda = 0.94$. On the contrary, GARCH has higher variability. This is due to its flexibility and ability to dynamically adjust to the data. It is more responsive to new information, accommodates periods of heightened market stress and reverts to the mean faster than Riskmetrics. Ultimately, we also find that the GARCH with leverage model has a lower mean-squared error than the Riskmetrics, leading us to prefer the GARCH with leverage model for forecasting volatility.

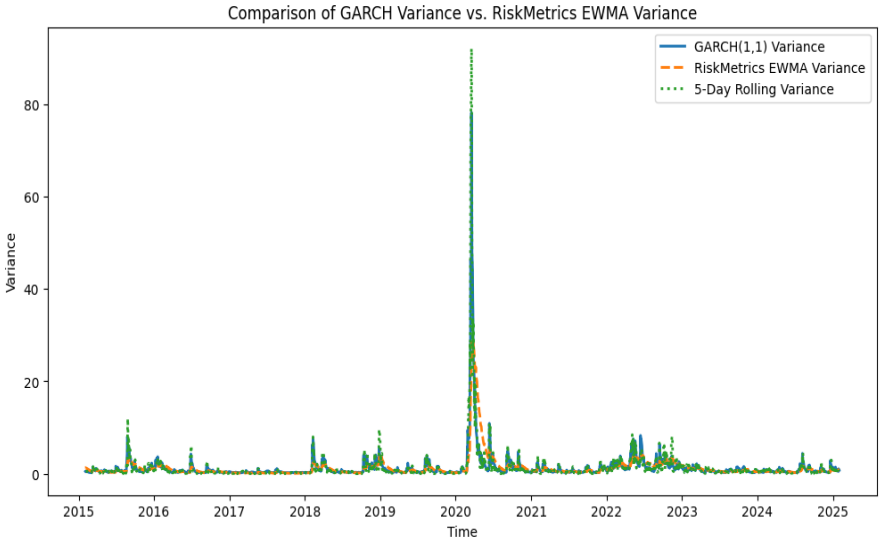

Figure 7: Comparison between GARCH and Riskmetrics variance, and rolling 5-day variance

3 GARCH vs. Riskmetrics

In this question, we compare variance forecasts from the GARCH and RiskMetrics models. Essentially, we will evaluate forecast accuracy for different time horizons by computing the mean squared error (MSE) and formally testing for statistical differences by using the Dielbold-Mariano (DM) test.

We set up a multi-step forcast for horizons $h \in \{1, 5, 10\}$. Moreover, we split the data and keep the last 200 observations as our out-of-sample period. The Riskmetrics variance forecast is:

$$\widehat{Var}_{t,h}^{(RM)} = h \cdot \widehat{Var}_{t,1}^{(RM)},$$

where $\widehat{Var}_{t,1}^{(RM)}$ is the one-step-ahead conditional variance estimate at time t.

For each forecast horizon, we compute the realized variance (RV) as the sum of squared log-returns from our S&P500 data:

$$Var_{t,h} = \sum_{i=1}^{h} R_{t+i}^2.$$

The forecast error for model m at time t and horizon h is defined as

$$E_{t,h}^{(m)} = \widehat{Var}_{t,h}^{(m)} - Var_{t,h}. \tag{2}$$

In the figure below we plot the forecast for each of the models.

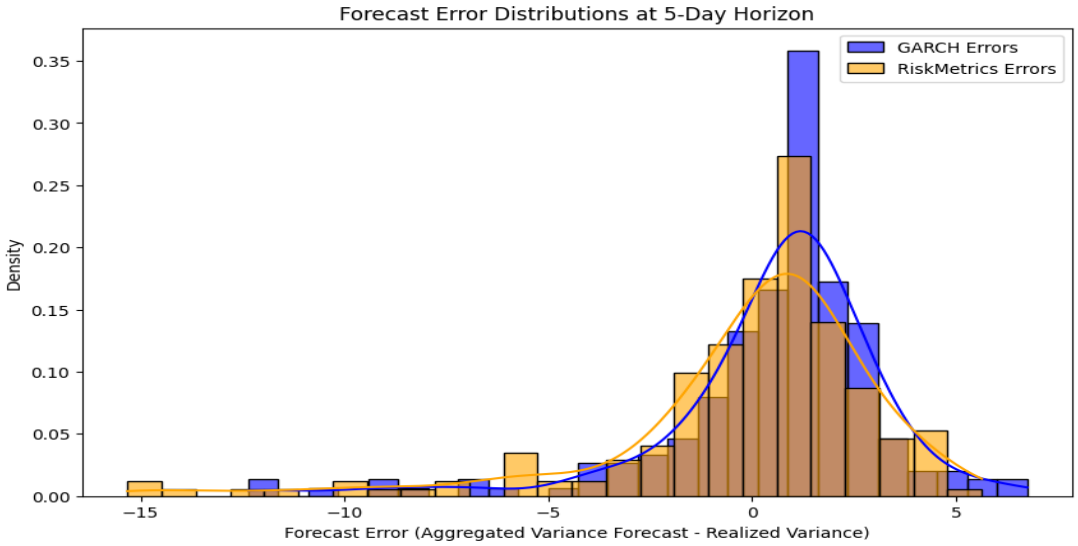

Figure 8: Forecast Error GARCH vs. Riskmetrics

The mean squared error (MSE) for model m at horizon h is then given by

$$\text{MSE}_h^{(m)} = \frac{1}{N} \sum_{t=1}^{N} \left(E_{t,h}^{(m)} \right)^2, \tag{3}$$

where N is the number of forecast origins in the out-of-sample period. We can observe on the table below that GARCH has greater accuracy and lower MSE values across all horizons. This points to its ability to better capture volatility dynamics as opposed to the simple Riskmetrics method. In any case, note that forecast errors increase with longer horizons for both models as expected.

Table 13: Forecast Horizons and Mean Squared Errors (MSE)

Forecast Horizon (days)	GARCH MSE	RiskMetrics MSE
1	1.026423	1.569153
5	8.170472	11.473946
10	29.254021	36.672969

Finally, we use the Diebold-Mariano test to determine whether the forecast accuracy differs significantly between the two models. The test statistic is given by

$$DM_h = \frac{\bar{D}_h}{S_{D,h}/\sqrt{N}}, \tag{4}$$

where

$$D_{t,h} = \left(E_{t,h}^{(1)}\right)^2 - \left(E_{t,h}^{(2)}\right)^2$$

is the difference in squared forecast errors, $\bar{D}_h = \frac{1}{N}\sum_{t=1}^{N} D_{t,h}$ is its sample mean, and $S_{D,h}$ is its sample standard deviation. Under the null hypothesis of equal predictive accuracy, the DM statistic asymptotically follows a standard normal distribution:

$$DM_h \sim \mathcal{N}(0,1).$$

We reject the null at a 5% confidence level, what suggests that the GARCH forecasts are consistently more accurate than those of Riskmetrics. All in all, it seems that GARCH performs better.

Table 14: Diebold-Mariano Test (GARCH vs. RiskMetrics forecasts)

Forecast Horizon (days)	DM Statistic	p-value
1	-2.4431	0.0146
5	-2.6804	0.0074
10	-2.0383	0.0415

4 Value at Risk (VaR)

5% Conditional VaR and Violations

For this question, we examine the Value-at-Risk (VaR) of multiple volatility models for the S&P500 daily log returns:

- GARCH(1,1) (normal and Student-t, without leverage effect)
- GARCH w/ leverage (1,1) (normal and Student-t, with leverage effect)

- IID normal model (benchmark)

All models assume conditional volatility and compute VaR as:

$$\text{VaR}_t = \mu + \sigma_t \, q_{0.05},$$

where $q_{0.05} = \Phi^{-1}(0.05) \approx -1.645$. The conditional standard deviation σ_t is obtained from the filtered variance produced by the GARCH recursion. A VaR violation occurs when the actual return R_t is less than the computed VaR:

$$R_t < \text{VaR}_t.$$

In Table 15 one can find the total number of violations and the violation rate.

Table 15: Summary of VaR Violations (5% VaR)

Model	**Violations**	**Violation %**	**Avg Severity**	**Max Severity**
GARCH (Normal)	133	5.29%	0.680	4.176
GARCH w/ leverage (Normal)	137	5.45%	0.623	4.550
IID Normal	110	4.37%	1.127	10.958
GARCH (Student-t)	139	5.53%	0.694	4.158
GARCH w/ leverage (Student-t)	147	5.84%	0.618	4.765

Violation Rates

- Expected violation rate at 5%: Most models align closely (5.29% for GARCH-normal, 5.45% for Garch w/ leverage-normal, 5.53% for GARCH-t).
- IID normal model underperforms (4.37%): This suggests it underestimates risk, leading to fewer violations than expected.
- Garch w/ leverage (Student-t) has the highest violations (5.84%): This implies it may overestimate risk, resulting in a slightly looser threshold.

Severity of Violations

- IID model has the most extreme risk misspecifications:
  - Average severity: 1.13 → meaning when it fails, it fails hard.
  - Maximum severity: 10.96 → highest single-day loss underestimation.

- GARCH models show lower severity (avg: 0.62-0.69), meaning they better account for tail risk.
- GARCH (Student-t) and Garch w/ leverage (Student-t) reduce tail risk underestimation, making them more reliable for extreme events.

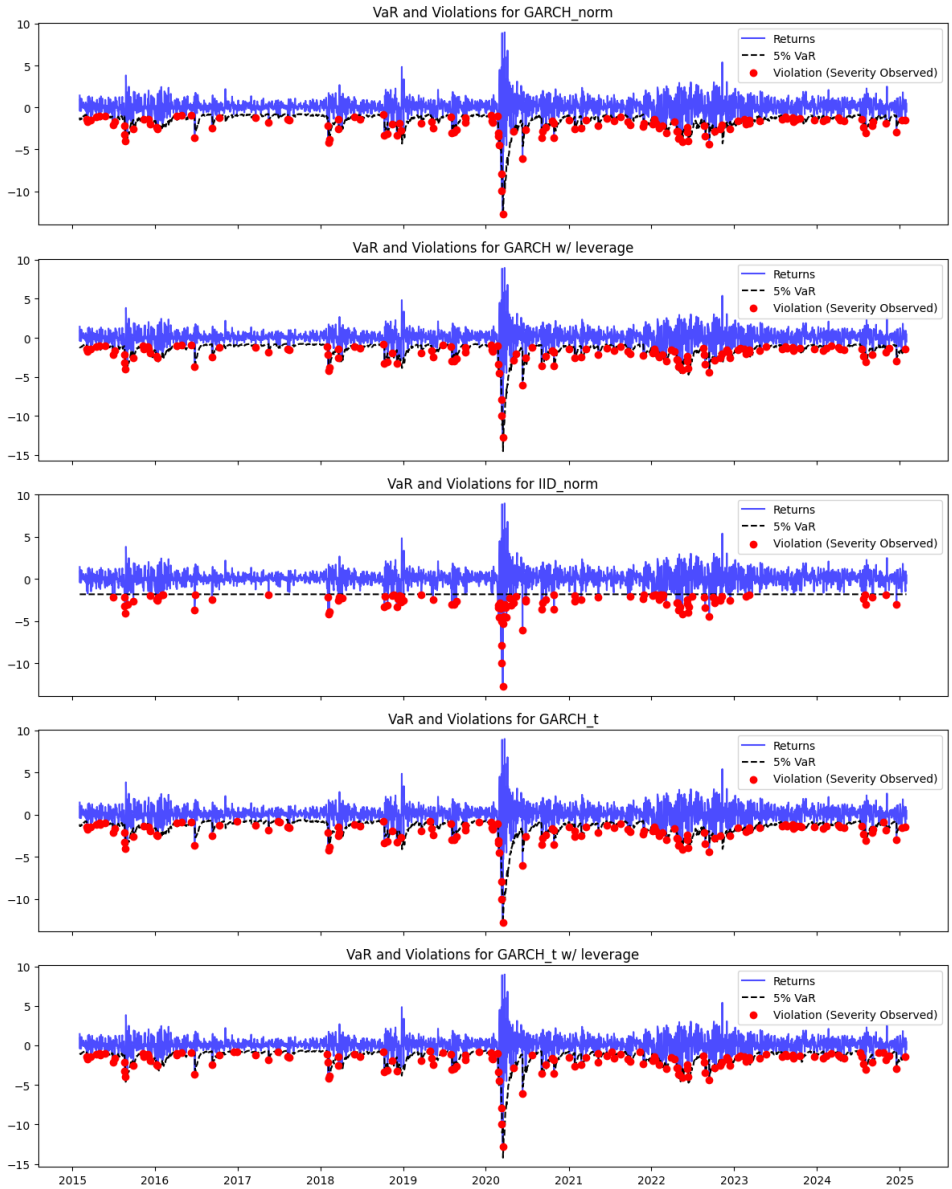

Figure 9: VaR violations

Tests

A well-specified model should produce hit sequences:

$$H_t = \begin{cases} 1, & \text{if } R_t < \text{VaR}_t, \\ 0, & \text{otherwise,} \end{cases}$$

that are independently and identically distributed as Bernoulli(α) with $\alpha = 5\%$. The following tests are conducted:

Unconditional distribution of H_t is Bernoulli(α)

The first test is the unconditional distribution test. Under the null hypothesis, the hit indicators H_t follow a Bernoulli distribution with success probability $\pi = \alpha$. The likelihood function based on a sample of T observations is

$$L(\pi) = \prod_{t=1}^{T} (1-\pi)^{1-H_t} \pi^{H_t} = (1-\pi)^{T_0} \pi^{T_1},$$

where T_0 and T_1 are the number of non-violations (zeros) and violations (ones), respectively.

The maximum likelihood estimator (MLE) of π is

$$\hat{\pi} = \frac{T_1}{T}.$$

Under the null, we have $\pi = \alpha$. The log-likelihood ratio test statistic is then given by

$$LR_{uc} = 2\Big[\log L(\hat{\pi}) - \log L(\alpha)\Big],$$

which, under the null hypothesis, is asymptotically distributed as $\chi^2(1)$.

For that, we use the Kupiec Test (Unconditional Coverage)

- Tests whether the observed violations match the expected 5%.
- Most models pass comfortably ($p > 0.1$), meaning they do not significantly deviate from expected violations.
- GARCH-t with leverage (p = 0.0579) is borderline, meaning it might slightly over-predict risk.

The H_t are i.i.d. and Bernoulli(α)

In the second test, we will check if the hit sequence is serially independent by testing against a first-order homogeneous Markov chain

$$\Pi = \begin{pmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{pmatrix},$$

where

$$\pi_{01} = P(H_t = 1 \mid H_{t-1} = 0) \quad \text{and} \quad \pi_{11} = P(H_t = 1 \mid H_{t-1} = 1).$$

The likelihood is

$$L(\Pi) = (1-\pi_{01})^{T_{00}} \, \pi_{01}^{T_{01}} \, (1-\pi_{11})^{T_{10}} \, \pi_{11}^{T_{11}},$$

where T_{ij} is the number of transitions from state i to state j.

The MLEs for the transition probabilities are

$$\hat{\pi}_{01} = \frac{T_{01}}{T_{00}+T_{01}}, \quad \hat{\pi}_{11} = \frac{T_{11}}{T_{10}+T_{11}}.$$

Under the null hypothesis of independence, the probability of a hit does not depend on the previous outcome; hence, we must have

$$\pi_{01} = \pi_{11} = \pi,$$

with π estimated as $\hat{\pi} = T_1/T$. Denote the corresponding transition matrix under independence by

$$\Pi_0 = \begin{pmatrix} 1-\hat{\pi} & \hat{\pi} \\ 1-\hat{\pi} & \hat{\pi} \end{pmatrix}.$$

The likelihood ratio test statistic for independence is then

$$LR_{ind} = 2\Big[\log L(\hat{\Pi}) - \log L(\Pi_0)\Big] \sim \chi^2(1),$$

under the null.

For that, we use the Christoffersen Test (Conditional Coverage)

- IID normal model fails badly ($p < 0.000007$) $\rightarrow$ It does not capture volatility clustering, making it unreliable in risk management.
- All other models pass comfortably ($p > 0.3$), meaning they handle time-dependency in risk correctly.

Table 16: Backtesting Test Results (5% VaR)

Model	**Kupiec p-value**	**Christoffersen p-value**
GARCH (Normal, no leverage)	0.511	0.452
GARCH (Normal, leverage)	0.310	0.348
IID Normal	0.141	**0.00007**
GARCH (t, no leverage)	0.233	0.396
GARCH (t, leverage)	**0.058**	0.619

Regression-Based Test for Predictability

Testing VaR Violations: Regression on Lagged VIX:

To examine whether market volatility (proxied by the VIX index) influences the occurrence of VaR violations, we estimate the following regression:

$$H_t = \alpha + \gamma \cdot \text{VIX}_{t-1} + \varepsilon_t$$

where H_t is the VaR violation indicator (1 if a violation occurs, 0 otherwise) and VIX_{t-1} is the lagged VIX value.

Table 17: OLS Regression Results: H_t on Lagged VIX

Model	**Const**	**VIX Coeff.**	**p-value**	R^2
GARCH (Normal, no leverage)	0.0529	0.1017	0.068	0.001
GARCH (Normal, leverage)	0.0545	0.1011	0.074	0.001
IID Normal	0.0438	0.0603	0.237	0.001
GARCH (t, no leverage)	0.0553	0.1031	0.070	0.001
GARCH (t, leverage)	0.0585	0.1038	0.076	0.001

Findings and Interpretation

- Across all models, the R^2 values are very low (≈ 0.001), indicating that lagged VIX explains almost none of the variation in H_t.

- The coefficient on VIX is positive, implying that higher market volatility slightly increases the probability of a violation, but the effect is not statistically significant ($p > 0.05$).
- The IID model shows the weakest relationship, suggesting that models accounting for time-varying volatility (GARCH with and without leverage) capture more of the relevant risk dynamics.
- The GARCH-t models, which assume fat-tailed innovations, do not significantly alter the results, implying that tail risk is not driving the relationship.
- The findings suggest that VIX is not a strong predictor of daily VaR violations. The GARCH models may already incorporate most of the relevant risk information.
- IID normal performs the worst → Fails to model extreme events and volatility clustering.
- GARCH models perform well, with the leverage one being more conservative.

5 Predictive Regressions

5.1 a) Predicitive Regression

We estimate the predictive regression model:

$$y_t = \beta x_{t-1} + u_t, \quad x_t = \rho x_{t-1} + v_t,$$

where the innovations (u_t, v_t) follow a bivariate normal distribution with covariance structure:

$$\text{Cov}[[u_t, v_t], [u_t, v_t]] = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}.$$

We simulate the process for different values of ρ and σ_{uv}, each with 1,000 replications. The OLS estimates of β are summarized in the table below.

Discussion

The results show that:

- When $\rho = 0.972$, the bias in the estimated β is positive when $\sigma_{uv} \neq 0$, consistent with Stambaugh's (1999) findings that correlated innovations introduce an upward bias. When $\sigma_{uv} = 0$ we see that the bias is close to zero.

Table 18: OLS Estimates of β Under Different Simulation Parameters

Sample Size (T)	ρ	σ_{uv}	**Mean** β	**Bias**	**Std Dev**
840	0.972	-0.000162	0.286753	0.076753	0.163195
840	0.972	0.000000	0.213983	0.003983	0.149122
840	0.500	-0.000162	0.232342	0.022342	0.510116
840	0.500	0.000000	0.206941	-0.003059	0.513880
840	0.000	-0.000162	0.225952	0.015952	0.594274
840	0.000	0.000000	0.203605	-0.006395	0.568264
1680	0.972	-0.000162	0.251828	0.041828	0.109633
1680	0.972	0.000000	0.207426	-0.002574	0.099582
1680	0.500	-0.000162	0.244054	0.034054	0.363980
1680	0.500	0.000000	0.204671	-0.005329	0.364251
1680	0.000	-0.000162	0.213500	0.003500	0.410459
1680	0.000	0.000000	0.208586	-0.001414	0.399930

- When $\rho = 0.5$, the standard deviation of β estimates increases significantly, highlighting the effect of lower persistence in x_t.
- When $\rho = 0$, the bias reduces and standard deviations remain large, confirming that predictability weakens in absence of persistence.
- Doubling the sample size (from $T = 840$ to $T = 1680$) significantly reduces the bias and standard deviation, confirming that increasing observations improves estimation efficiency.

These results confirm that persistence in x_t and correlation between u_t and v_t affect the reliability of OLS estimates in predictive regressions.

b) Predictive Regressions: Multi-Step Forecasts

5.2 Model simulation

To assess the predictive content of x_t, we estimate two-step and four-step ahead regressions. The standard deviation of the OLS estimator is computed using:

- Naive method: Directly from OLS standard errors.

- Newey-West method: Correcting for autocorrelation in residuals.
- Analytical formula: Derived as:

$$\text{SE}_{\beta,h} = \sqrt{\frac{\sigma_u^2 + h \cdot \sigma_{uv} \cdot \beta + \beta^2 \cdot \sigma_v^2}{T(1-\rho^2)}}$$

Table 19: OLS Estimates of β for Multi-Step Ahead Forecasts

Sample Size	ρ	σ_{uv}	**Horizon**	**Mean OLS** β	**Naive SE**	**Newey-West SE**
840	0.972	-0.000162	2	0.2819	0.1475	0.1455
840	0.972	-0.000162	4	0.2648	0.1477	0.1429
840	0.972	0.000000	2	0.2018	0.1474	0.1466
840	0.972	0.000000	4	0.1888	0.1477	0.1464
840	0.500	-0.000162	2	0.1733	0.5011	0.4961
840	0.500	-0.000162	4	0.0645	0.5018	0.4944
1680	0.972	-0.000162	2	0.2382	0.0994	0.0982
1680	0.972	-0.000162	4	0.2255	0.0995	0.0969
1680	0.000	0.000000	2	-0.0127	0.4076	0.4066
1680	0.000	0.000000	4	-0.0109	0.4078	0.4061

Table 20: Comparison with Analytical Standard Errors

Forecast Horizon	**Newey-West SE**	**Analytical SE**
2-Step Ahead	0.1455	0.0080
4-Step Ahead	0.1429	0.0079

Findings and Interpretation

The results highlight:

- The OLS estimator exhibits bias ($\mathbb{E}[\hat{\beta}] > 0.21$), which aligns with the well-known small-sample bias in predictive regressions when ρ is close to 1.
- The **naive standard errors** tend to overestimate the variability of $\hat{\beta}$, while Newey-West SEs provide a slightly lower but still conservative estimate.

- The analytical SEs are significantly lower than both methods, as they reflect the asymptotic variance without additional correction for serial correlation.
- Increasing the forecast horizon ($h = 4$) leads to a reduction in β, indicating a decay in predictive power over time.
- The impact of persistence (ρ) is substantial: higher persistence results in an upward bias in β, while lower persistence causes β to shrink towards zero.
- When the errors are negatively correlated ($\sigma_{uv} < 0$), the estimated β is higher. Removing this correlation leads to a downward shift in β.
- Standard errors shrink with increasing sample size ($T = 1680$), but the improvement is more noticeable when ρ is high.
- The Newey-West standard errors closely match the naive SEs, but both overestimate the theoretical analytical SE.

Conclusion

The findings confirm that persistence, forecast horizon, and error correlation significantly influence predictive regression estimates. Higher persistence leads to overestimation, longer horizons reduce predictability, and the choice of standard error estimation affects inference reliability.

c) Predictive Regressions: Robert Shiller long horizon regressions

5.3 Long horizon regressions

The data for this study is obtained from Robert Shiller's publicly available dataset, which provides historical U.S. stock market data used in *Irrational Exuberance*. The dataset includes information on stock prices, dividends, earnings, interest rates, and inflation (CPI), allowing us to construct real (inflation-adjusted) variables.

5.3.1 Data Preparation

- The log dividend-price ratio is constructed as:

$$dp_t = \log(\text{Real Dividend}_t) - \log(\text{Real Price}_t) \tag{5}$$

- We initially intended to use the stochastically detrended short-term interest rate, but due to data limitations, we instead used the detrended dividend-price ratio (dp_{short}).

- The detrending follows an AR(1) process:

$$dp_{\text{short},t} = dp_t - (\alpha + \rho dp_{t-1}) \tag{6}$$

  where α and ρ are estimated separately for different sample periods.

5.4 Regression Equations and Methodology

To test for long-horizon return predictability, we estimated the following regressions:

5.4.1 Standard Regression: Log Dividend-Price Ratio

$$R_{t:t+K} = \alpha + \beta dp_t + \varepsilon_t \tag{7}$$

where:

- $R_{t:t+K}$ is the cumulative stock return over K months.
- dp_t is the log dividend-price ratio.
- ε_t is an error term.

5.4.2 Alternative Regression: Detrended dp

$$\sum_{j=1}^{K} \log(R_{t+j}) = \beta^{(K)} \left(\log(Y_t) - \sum_{i=0}^{11} \log(Y_{t-i}) \right) + \eta_{t+K,K} \tag{8}$$

where:

- Y_t represents the real dividend or real price.
- The term $\log(Y_t) - \sum_{i=0}^{11} \log(Y_{t-i})$ removes long-term trends.

- The left-hand side accumulates stock returns over K months.

For both models, we use Newey-West standard errors to correct for autocorrelation and heteroskedasticity.

5.5 Results and Interpretation

5.5.1 Regression of Long-Horizon Returns on dp

Table 21: Regression Results: Long-Horizon Returns on dp

Horizon (months)	**Beta**	**SE**	$\mathbf{R^2}$	**N**
1	0.0008	0.0006	0.0011	1610
3	0.0026	0.0029	0.0012	1610
12	0.0123	0.0225	0.0018	1610
24	0.0282	0.0608	0.0024	1610
36	0.0441	0.1060	0.0028	1610
48	0.0564	0.1540	0.0027	1610

Interpretation:

- Weak predictive power: The R^2 values remain very low, suggesting that the log dividend-price ratio explains only a small fraction of stock return variation.
- Longer horizons slightly increase predictability: The beta coefficients tend to increase as the horizon lengthens, supporting the idea that valuations play a role in predicting long-run returns rather than short-term returns.
- Statistical insignificance: The large standard errors suggest that many coefficients are not statistically different from zero.

5.5.2 Regression of Long-Horizon Returns on $dp_{\mathbf{short}}$

Interpretation:

- Negative coefficients: Unlike the standard regression, the coefficients here are negative, suggesting that removing long-term trends in dp changes its relationship with future returns.

Table 22: Regression Results: Long-Horizon Returns on dp_{short} (Full Sample)

Horizon (months)	**Beta**	**SE**	$\mathbf{R}^2$	**N**
1	-0.0056	0.0020	0.0093	758
3	-0.0171	0.0103	0.0099	760
12	-0.0713	0.0813	0.0108	758
24	-0.1467	0.2225	0.0117	758
36	-0.2369	0.3904	0.0139	758
48	-0.3523	0.5694	0.0176	758

- Predictability remains low: The R^2 values are slightly higher than before but still very low.
- Possible instability: The standard errors are large, especially for longer horizons, indicating that the results may not be stable.

5.6 Predictability Over Different Time Periods

- Early period (1952-1997): Predictability is weak, with low R^2 and positive but small coefficients.
- Late period (2009-2025): The results change dramatically, predictability increases significantly, and the coefficients turn highly positive for longer horizons. This suggests that return predictability may have become stronger in recent decades.

5.7 Conclusion

- Long-horizon predictability is weak: Despite the theoretical motivation, the explanatory power of dp and dp_{short} remains low.
- Detrending changes results: Removing long-term trends reverses the coefficient signs, suggesting that valuation ratios matter more in the long run.
- When we increase the horizon of the regression, we get better predictability and higher R squared for the standard regression, which goes in line with the theory.