ETF Project 2

Kasper Bankler (s255892)

November 2025

Contents

1 Abstract

This project investigates whether it is possible to predict the future geometric average rate of return in the stock market based on data from the past. This will be tested using a multiple linear regression model based on a dataset of 95 exchange-traded funds (ETFs) and their volatility and maximum time under water (maxTuW) as predictors. Statistical analysis found maxTuW to be an insignificant predictor, leading to its removal from the model. The results show that it is not possible to reliably predict future returns based on these parameters.

2 Statistical analysis

a) Descriptive analysis and summary of the dataset

The dataset consists of 4 variables. *ETF* is a categorical variable with the names of the ETFs. *Geo.mean* is a quantitative variable for the geometric average rate of return. *Volatility* and *maxTuW* (Maximum Time under water) are both quantitative risk measures. There are in total 95 ETFs with data from the period 2006-05-05 to 2015-05-08 (454 weeks). The table below shows summary statistics for the three quantitative variables.

Variable	**N**	**Mean**	**Standard Deviation**	**Median**	**0.25 Quantile**	**0.75 Quantile**
Geo_mean	95	0.076904	0.080867	0.082737	0.028712	0.134381
Volatility	95	3.059812	0.879042	3.026361	2.587946	3.675211
maxTuW	95	307.294737	42.767524	324.000000	309.000000	327.000000

Table 1: Summary statistics for each variable

The scatter plots below show the geometric average rate of return against volatility (left) and maxTuW (right). The left plot shows that the volatility is widely scattered with a few outliers in both the positive and negative directions. The right plot shows a tight cluster at around $maxTuW = 300 - 330$ and a smaller cluster at $maxTuW = 170 - 200$.

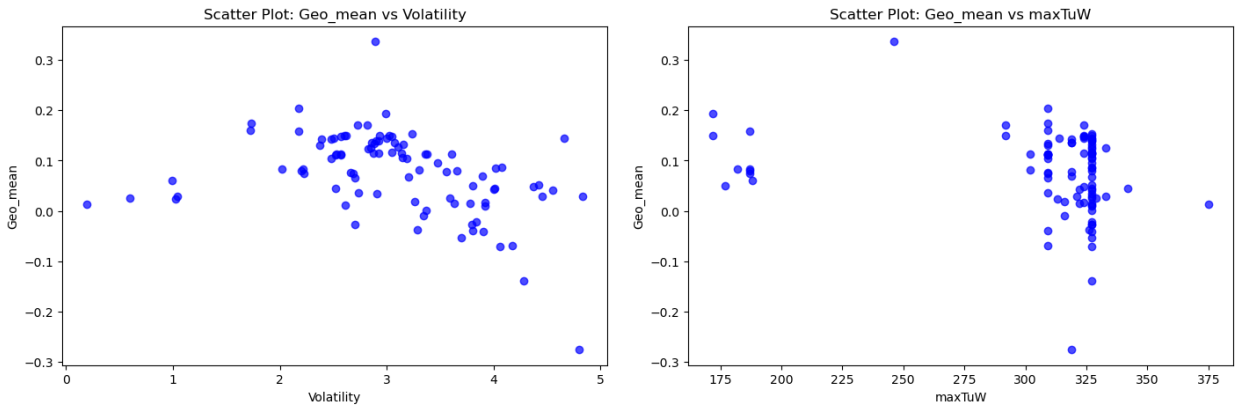


Figure 1: Scatter plots of the geometric average rate of return against volatility and MaxTuW

The boxplots below sshowthat *Geo.mean* (left) and *Volatility* (middle) is relatively symmetrical with only a few outliers. The boxplot of *maxTuW* (right), however, shows an asymmetrical distribution with a lot of outliers.

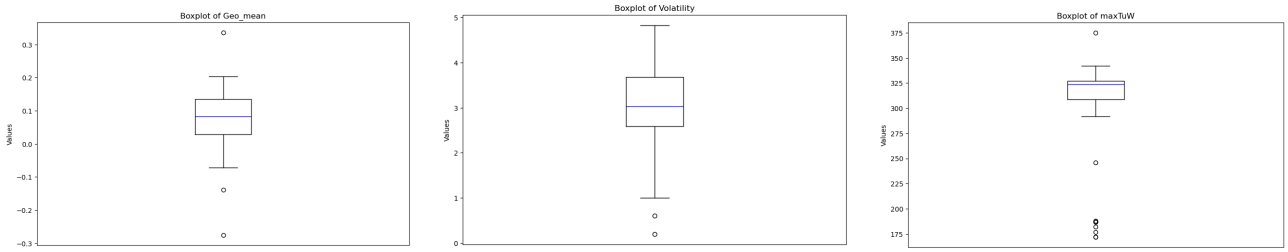


Figure 2: Comparison of boxplots

The histograms below shshowhat *Geo.mea* (left) and *Volatility* (middle) is relatively symmetrical and approximately normal. The histogram of *maxTuW* (right) is asymmetrical with a massive peak on the right. It does not appear to be normal but rather left-skewed.

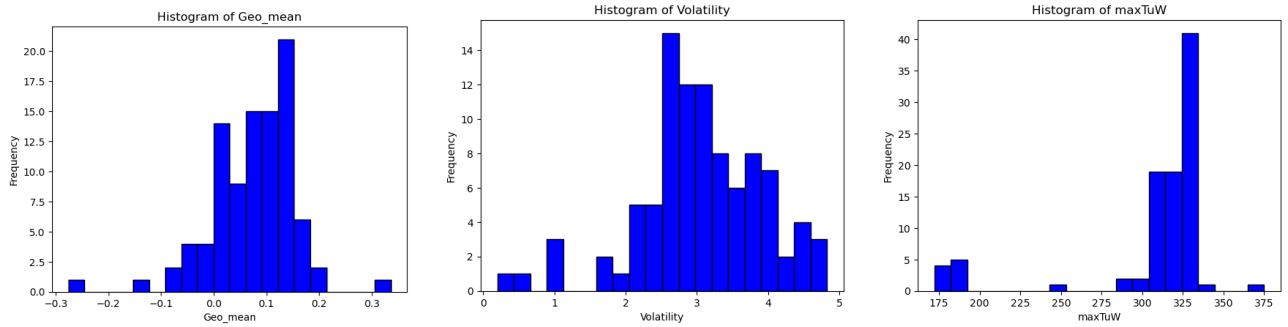


Figure 3: Comparison of histograms

b) Multiple linear regression model

The dataset is modeled using multiple linear regression, as shown below:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

Where Y_i is the geometric average rate of return, $x_{1,i}$ is the volatility, and $x_{2,i}$ is maxTuW. We assume that the residuals (ϵ_i) are independent and identically distributed (i.i.d.) and that they follow a normal distribution with zero mean and a constant variance (σ^2).

c) Parameter estimates of the model

The parameters of the model are estimated using the *ols* function from the Python library *statsmodels.formula.api*. The table below shows the estimated coefficients and the estimated standard deviation. The residual variance is $\hat{\sigma}^2 = 0.005748$. This is calculated in Python with the *mse_resid* method from *statsmodels.formula.api*. It uses $n - (p + 1)$ degrees of freedom. In our case $91 - (2 + 1) = 88$. The explained variation $R^2 = 0.170$ is also given by the ordinary least squares analysis in Python. This shows that 17% of the variance in the dependent variable (Geo_mean) is explained by the independent variables (variance and maxTuW). It means that 83% of the variance in (Geo_mean) is unexplained due to other factors not included in the model.

	Coefficients	**Estimated standard deviation ($\hat{\sigma}$)**
$\hat{\beta}_0$	0.2528	0.058
$\hat{\beta}_1$	-0.0351	0.010
$\hat{\beta}_2$	-0.0002	0.000

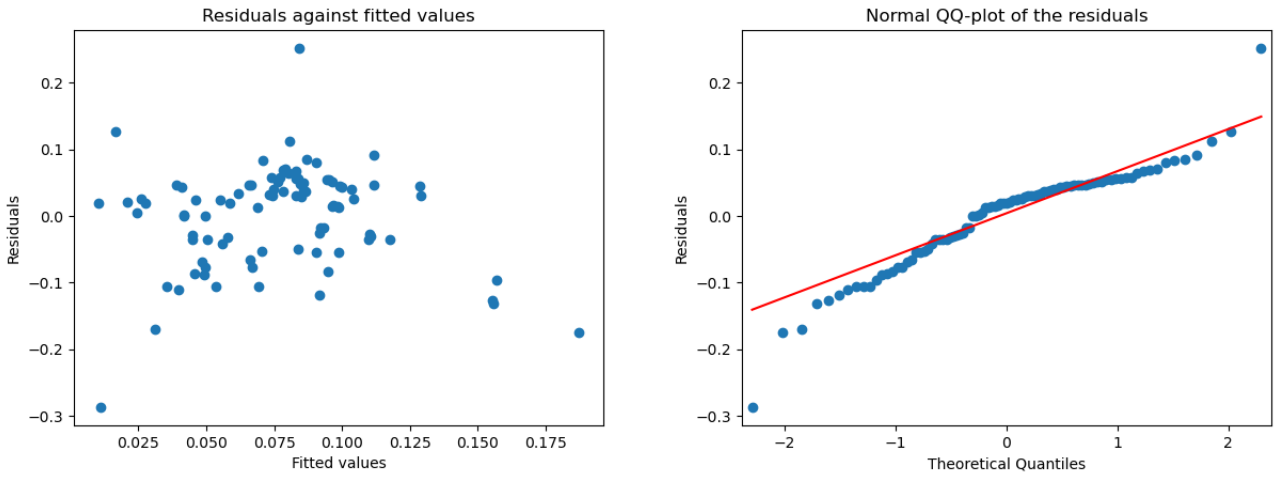
Table 2: Parameter estimates

d) Model validation

We assume independence is satisfied because the data represents a random sample of different ETFs. Therefore one observation does not influence the measurement of another.

Looking at the plot of the residuals as a function of fitted values is constructed (left), there doesn't seem to be any systematic dependence between the fitted values. This suggests that the linear relation assumed by the model is valid. The residuals also appear to have a constant variance as assumed by the model.

To assess the normality assumption, a QQ-plot is made. The QQ-plot (right) shows that the residuals approximately follow a normal distribution. There are, however, some deviations in the tails, with two outliers standing out. According to the central limit theorem (CLT), the distribution of the coefficients will still be approximately normal, as the sample size is large ($n = 91$).



(a) Residuals against fitted values

(b) QQ-plot of the residuals

Figure 4: Residual diagnostics plots

e) Confidence interval for the volatility coefficient

The confidence interval for the volatility coefficient (β_1) is given by the formula:

$$\hat{\beta}_1 \pm t_{1-\alpha/2}\ \hat{\sigma}_{\beta_1}$$

Where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a t-distribution with $n-(p+1)$ degrees of freedom. Inserting numbers and calculating the 95% confidence interval:

$$-0.0351 \pm 1.98729 \cdot 0.010$$
$$\Downarrow$$
$$[-0.0550, -0.0152]$$

This calculation is repeated for the two other regression coefficients. See the table below for the two other confidence intervals.

	Lower	**Upper**
Intercept	0.136617	0.369062
Volatility	-0.054453	-0.015809
maxTuW	-0.000600	0.000159

Table 3: 95% confidence intervals for the regression coefficients

f) Hypothesis test

To test whether β_1 might be -0.06 with significance level $\alpha = 0.05$, the following two hypotheses are formulated:

$$H_0 : \beta_1 = -0.06$$
$$H_1 : \beta_1 \neq -0.06$$

The test statistic is given by the formula:

$$t_{obs,\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$$

Inserting values and calculating t_{obs,β_1}:

$$t_{obs,\beta_1} = \frac{-0.0351-(-0.06)}{0.010} = 2.49$$

Now the evidence against the hypothesis is computed with the formula:

$$p\text{-value}_i = 2P(T > |t_{obs,\beta_i}|)$$

Where T follows a t-distribution with $n - (p + 1)$ degrees of freedom. In our case, that is $91 - (2 + 1) = 88$ degrees of freedom. Inserting values and calculating the p-value:

$$p\text{-value} = 2P(T > |2.49|) = 0.01465$$

We reject hypothesis H_0 since the p-value is less than α ($0.01465 < 0.05$). This means that at a 5% significance level, there is significant evidence to conclude that β_1 is different from -0.06.

g) Reducing the model

To reduce the model, the p-value for the *null hypothesis* is calculated for all three parameters using the *ols* function from the Python library *statsmodels.formula.api*. The three p-values are shown in the table below.

	t	$P > \|t\|$
$\hat{\beta}_0$	4.323	0.000
$\hat{\beta}_1$	-3.613	0.001
$\hat{\beta}_2$	-1.154	0.252

Table 4: t-statistics and p-values

The parameter $\hat{\beta}_2$ is not significant (on a 0.05 significance level). Therefore, this parameter is removed from the model. The model has now been reduced using backward selection. Below is the new multiple linear regression model with maxTuW (β_2) removed:

$$Y_i = \beta_0 + \beta_1 x_{1,i}, \quad \epsilon_i \sim N(0, \sigma^2),$$

Where Y_i is the geometric average rate of return and $x_{1,i}$ is the volatility. We once again assume that the residuals (ϵ_i) are independent and identically distributed (i.i.d.) and that they follow a normal distribution with zero mean and a constant variance (σ^2). Below are the new parameter estimates. The new residual variance is $\hat{\sigma}^2 = 0.0057698$.

	Coefficients	**Standard Error**
$\hat{\beta}_0$	0.1949	0.030
$\hat{\beta}_1$	-0.0382	0.009

Table 5: Parameter estimates for the new model

h) Predictions of the geometric average rate of return

The prediction intervals and the geometric average rate of return (Geo mean) for the model are calculated in Python using the *get_prediction* function from the *statmodels* library. The values are shown in the table below.

	Geo mean (observed)	**Geo mean (predicted)**	**pred_lower**	**pred_upper**	**Percentage Error (%)**
SPY	0.104904	0.100091	-0.052085	0.252266	4.59
IWN	0.066849	0.072449	-0.079323	0.224220	-8.38
AGG	0.024794	0.172013	0.013358	0.330668	-593.77
VAW	0.113346	0.056897	-0.095170	0.208964	49.80

Table 6: Observed vs. predicted values, prediction intervals, and percentage error

The predicted geometric average rate of return for SPY and IWN both fall under 10% error compared to the observed values, indicating a relatively good prediction. The predictions for AGG and VAW, however, show a clear deviation from the observed values with respectively -594.8% and 49.8% error. As seen in Figure 6, the prediction intervals are also extremely wide, indicating a high uncertainty in the model. While the observed values for all four ETFs do fall within these intervals, the ranges are too large to be useful in predicting future values. This concludes that the model is not a useful tool for predicting the geometric average rate of return.

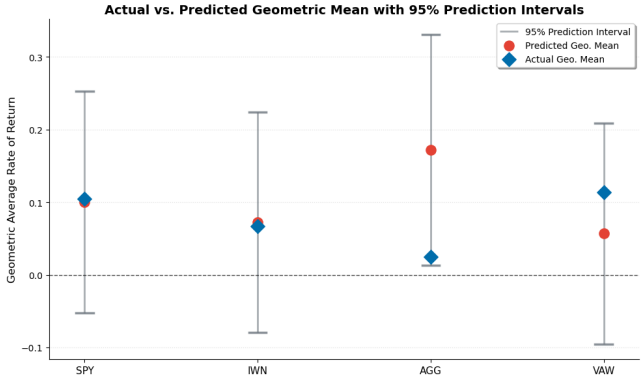


Figure 5: Comparison of actual vs. predicted geometric mean returns, with 95% prediction intervals

3 Appendix

This appendix contains the Python code used for the figures and to perform the calculations in this report. The full Jupyter Notebook is available in:

`etf_analysis2.ipynb`