2018

Factor Testing Report

HUIQI TIAN
MINJIE QIAN

EFT GLOBAL | NYU FRE

Table of Contents

Introduction

Factor analysis can be traced back Ben Graham who laid out the foundation of value investing. While there is a long history of using Factors to support investment decisions, Factor investing has recent exploded in terms of users and research.
ETF Global has been running their multi-factor models live on ETPs since 2012, covering Equity based products that are listed on a US Exchange.
This project aims to test the significance of ETF Global's factor scores in generating alpha. And we used three regression models to do it: Linear regression, Stepwise regression and Lasso regression.

Managing Data

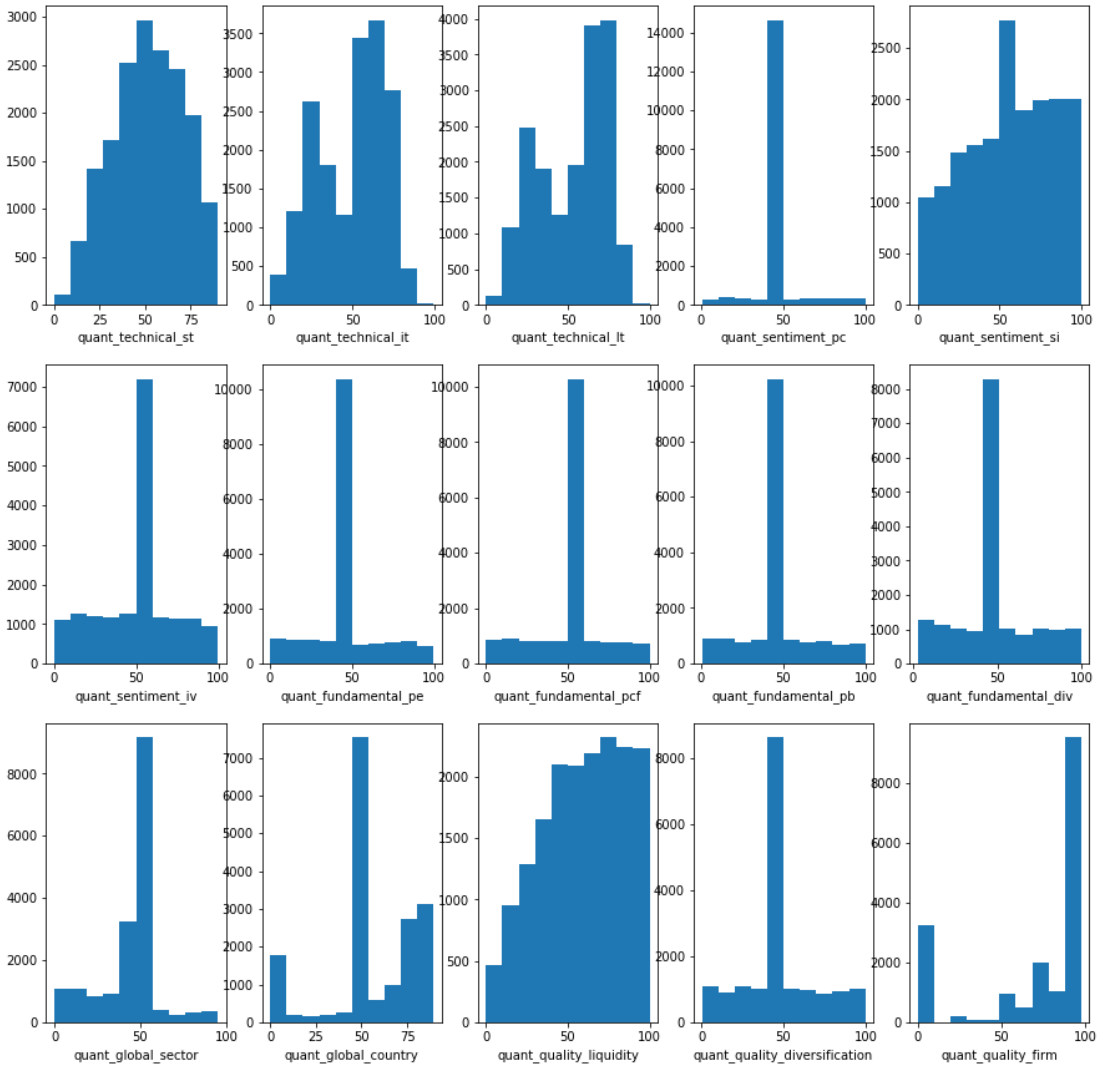
We spent much time on retrieving and cleaning data, while it may seem trivial, but it is quite important to our work later. In this way, it is easier to spot anomalies during the analysis and it helps to build the regression model later. What we eventually want is a csv file that contains the dependent variables and independent variables. Therefore, we did the following steps to achieve it.

The ETF Global company provides the independent variables, which are factors, online and we use FTP to download them to our local file. Then, we read the file into python pandas dataframe and store them as a list. That are the independent variables parts. The data contains technical, sentimental, fundamental, global and quality factor that related to different ETF and could be separated into 21 different features. These factor is ranked in order and convert to 100 scales, which means that the ETF has the best performance in certain factor is scored as 100 at that factor.

For the dependent variables, we use daily return and monthly return calculated by ETF close price automatically downloaded from Yahoo Finance.

Then we concentrate them according to their common dates, and eliminate the outliers and influential point for regression. Also, we test the collinearity between them and eliminated 6 independent variables.

Below is the histogram on the remaining 15 factors from 2012-03-01 to 2012-04-01.



From this histogram we can see that all factors are concentrated around 50. This may due to the calculations behind them, for example, some factors like Put/Call ratio would be expected to have those spikes.

Regression Models

OLS

Linear regression is a linear approach for modeling the relationship between a scalar dependent variable and some independent variables. And we use ordinary least square (OLS) to estimate the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the observed dependent variable in the given dataset and those predicted by the linear function. And then, test the significance of the parameters.

We used two methods to do this part.

One method is that we can use the sklearn package in python to do the linear regression and choose the factors based on the p-value they provided.

Another way is that we can test the significance of beta for each factor by calculating their t-statistics. According to the OLS,

$$\beta = (X^T X)^{-1} X^T y$$

Therefore, we can calculate beta for each factor. And then, according to the central limit theorem, the mean of beta follows the normal distribution with mean of zero, and variance of sigma square over n, which is:

$$\bar{\beta} \sim N(0, \frac{\sigma^2}{n})$$

Thus, the t-statistics is

$$t = \frac{\bar{\beta} - 0}{\frac{\sigma}{\sqrt{n}}}$$

And then get the p-value of t-statistics from python statsmodels

The result shows that 3 technical factors, 2 global factors and 2 quality factors are significant in short term; only two factors are significant in long term. The regression model results are in the appendix 2.

To be specific, in the short term, which means in one months, there are seven factors are significant:
quant_technical_st,
quant_technical_it,
quant_technical_lt,
quant_global_sector,
quant_global_country,
quant_quality_liquidity,
quant_quality_firm

In the long term, which means in one year, there are two factors are significant:
quant_technical_st
quant_sentiment_iv

Stepwise

Stepwise regression is one of the choice to improve the result of the ordinary linear square result. The stepwise regression is used as an automatic method to select efficient independent variables. In this project, a forward stepwise regression based on t-test

were used. The algorithm is that we start with no variables in the model and add one independent variable with minimum p-value each time and repeating this process until none improves the model to a statistically significant extent.

The result shows that 3 technical factors, 2 global factors and 2 quality factors are significant in short term; and there are 3 factors are significant in long term.

To be specific, in the short term, there are 7 factors significant:
quant_technical_st
quant_technical_it
quant_technical_lt
quant_global_sector
quant_quality_liquidity
quant_quality_firm
quant_global_country

In the long term, which means in one year, there are three factors are significant:
quant_technical_st
quant_sentiment_iv
quant_sentiment_si

So the result has been improved by stepwise regression compared to OLS regression. The 'quant_sentiment_si' factor has been recognized as significant in the stepwise regression.

Lasso regression

Lasso (least absolute shrinkage and selection operator) is also a regression analysis method that performs both variable selection and regularization. It selects factors based on L1 regularization. It adds a factor of sum of absolute value of coefficient to the optimization objective. The formula is shown as below:

$$min \frac{1}{2n} \|Xw - y\|_2^2 + \alpha \|w\|_1$$

To find the best alpha of Lasso regression model, Cross Validation is also used to avoid overfitting problem.

Using daily return, the cross validation shows that the best fit alpha is 0.1889 and the result shows that 3 technical factors, 1 sentiment factor, 1 fundamental, 2 global factors and 2 quality factors is significant, which are

quant_technical_st
quant_technical_it
quant_technical_lt
quant_sentiment_iv
quant_fundamental_pb

quant_global_sector

In the long term, the cross validation shows that the best fit alpha is 4.3114，and the result shows in one year, there are only one factor is significant:

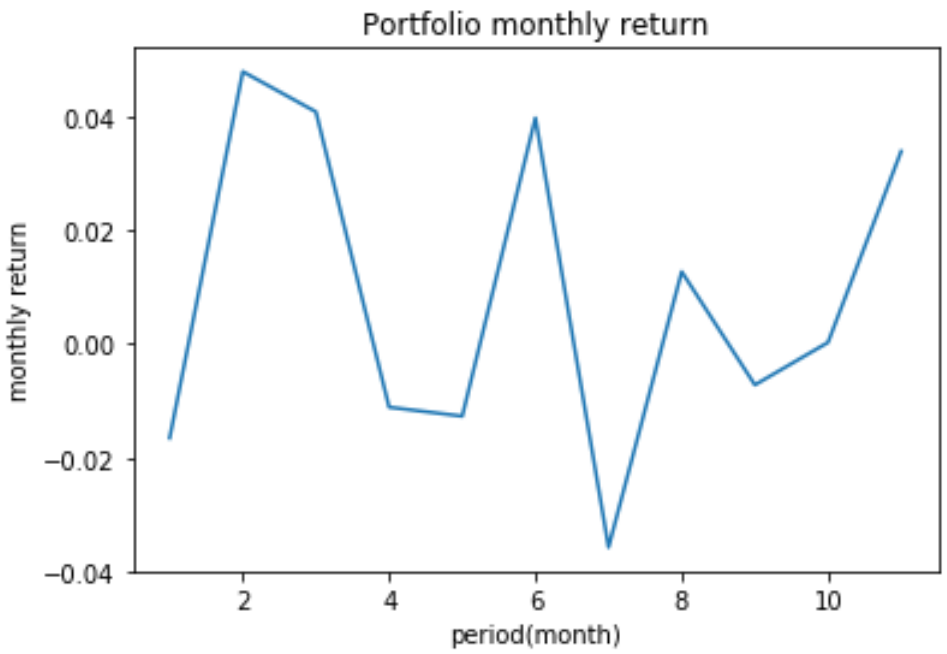
quant_global_sector.

Portfolio Construction

Most regression model shows that technical factor is significant both in short term and long term. Based on that result, we could build the portfolio relate to the technical composite factor. For each end of months, long ETF with 60 highest scores in technical composite factor and short ETF with 60 lowest scores in the same factor.

The annually return of this portfolio from 2013-2014 is 9.15%

Plot of monthly return is showing below:



Conclusion

In conclusion, stepwise regression often works when only few features are significant, Lasso could do better in other cases. The result of stepwise regression shows that it does improve the result from traditional OLS model and the result of Lasso is preferred according to more factor elimination.

According to the result from all three regression models, a market neutral portfolio is

constructed based on the most significant factor, the technical factor, both on short term and long term. This portfolio can be modified by adding or reducing more ETF both in long and short position, which may modify the return and modify the volatility of the return.

Appendix 1

The Code is shown below:

```
#ols
def ordinary_linear(y,x):
    model = sm.OLS(y,x)
    print(model.fit().summary())
    feature = []
    for key,value in model.fit().pvalues.items():
        if value < 0.05 and key != "const":
            feature.append(key)
    return feature

#stepwise regression
#df: dataframe of data
#Y: data of the independent data
#features: the features column of the dataframe
#alpha: setting alpha to limiited the minimum pvalue
def stepwise(df,Y,features,alpha):
    feature_col = list(features)
    length  =  len(feature_col)
    final_feature = []
    for i in range(length):
        pvalue_min = 1
        column_min = ""
        for feature in feature_col:
            temp_feature  = final_feature + [feature]
            x = sm.add_constant(df[temp_feature])
            model = sm.OLS(Y,x)
            pvalue = model.fit().pvalues[i+1]
            if pvalue < pvalue_min and pvalue < alpha:
                pvalue_min = pvalue
                column_min = feature
        if column_min != "":
            feature_col.remove(column_min)
            final_feature.append(column_min)
        else:
            break
    X = sm.add_constant(df[final_feature])
    model =  sm.OLS(Y,X)
    return final_feature
```

```
#Lasso
def Lasso_feature_select(y,x):
    lassocv = LassoCV(fit_intercept=True)
    model = lassocv.fit(x,y)
    print("LASSO feature Selection:")
    print("Best fit alpha under Lasso Cross Validation: %f" %model.alpha_)
    print(np.array(features)[model.coef_[1:] != 0])
    return np.array(features)[model.coef_[1:] != 0]

#ETF portfolio selection
#def etf_pick(date,factor, threshold):
    long_stock = []
    short_stock = []
    file = "data/" + date + ".csv"
    data = pd.read_csv(file)
    order_data = data.sort_values(factor,ascending = False)
    long_stock = order_data["ticker"].head(threshold).tolist()
    short_stock = order_data["ticker"].tail(threshold).tolist()
    return long_stock,short_stock
```

Appendix 2

OLS regression model result:

Short term:

OLS Regression Results

Dep. Variable:	daily_return	R-squared:	0.124
Model:	OLS	Adj. R-squared:	0.123
Method:	Least Squares	F-statistic:	165.6
Date:	Sun, 06 May 2018	Prob (F-statistic):	0.00
Time:	16:25:18	Log-Likelihood:	-21146.
No. Observations:	17539	AIC:	4.232e+04
Df Residuals:	17523	BIC:	4.245e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.5424	0.052	-10.403	0.000	-0.645	-0.440
quant_technical_st	0.0226	0.000	48.138	0.000	0.022	0.024
quant_technical_it	-0.0086	0.001	-11.879	0.000	-0.010	-0.007
quant_technical_lt	-0.0043	0.001	-6.627	0.000	-0.006	-0.003
quant_sentiment_pc	3.981e-05	0.000	0.081	0.936	-0.001	0.001
quant_sentiment_si	8.433e-05	0.000	0.253	0.801	-0.001	0.001
quant_sentiment_iv	0.0003	0.000	1.034	0.301	-0.000	0.001
quant_fundamental_pe	-0.0001	0.000	-0.292	0.770	-0.001	0.001
quant_fundamental_pcf	3.083e-05	0.001	0.061	0.952	-0.001	0.001
quant_fundamental_pb	0.0004	0.001	0.791	0.429	-0.001	0.001
quant_fundamental_div	-4.794e-05	0.000	-0.147	0.883	-0.001	0.001
quant_global_sector	0.0018	0.000	4.544	0.000	0.001	0.003
quant_global_country	-0.0008	0.000	-2.619	0.009	-0.001	-0.000
quant_quality_liquidity	-0.0009	0.000	-2.545	0.011	-0.002	-0.000
quant_quality_diversification	0.0002	0.000	0.754	0.451	-0.000	0.001
quant_quality_firm	0.0005	0.000	2.644	0.008	0.000	0.001

Long term:

OLS Regression Results

Dep. Variable:	monthly_return	R-squared:	0.107
Model:	OLS	Adj. R-squared:	0.096
Method:	Least Squares	F-statistic:	10.15
Date:	Sun, 06 May 2018	Prob (F-statistic):	2.97e-23
Time:	16:38:51	Log-Likelihood:	-1877.0
No. Observations:	1291	AIC:	3786.
Df Residuals:	1275	BIC:	3869.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5192	0.306	1.697	0.090	-0.081	1.120
quant_technical_st	0.0159	0.002	6.376	0.000	0.011	0.021
quant_technical_it	0.0022	0.004	0.586	0.558	-0.005	0.009
quant_technical_lt	-0.0100	0.003	-3.226	0.001	-0.016	-0.004
quant_sentiment_pc	0.0013	0.002	0.538	0.591	-0.003	0.006
quant_sentiment_si	-0.0065	0.002	-3.700	0.000	-0.010	-0.003
quant_sentiment_iv	0.0089	0.001	6.139	0.000	0.006	0.012
quant_fundamental_pe	-0.0010	0.002	-0.506	0.613	-0.005	0.003
quant_fundamental_pcf	-0.0062	0.002	-3.084	0.002	-0.010	-0.002
quant_fundamental_pb	-0.0019	0.002	-0.813	0.417	-0.006	0.003
quant_fundamental_div	-0.0013	0.001	-0.934	0.350	-0.004	0.001
quant_global_sector	-0.0006	0.002	-0.269	0.788	-0.005	0.004
quant_global_country	-0.0080	0.002	-4.501	0.000	-0.011	-0.004
quant_quality_liquidity	0.0023	0.002	1.113	0.266	-0.002	0.006
quant_quality_diversification	-0.0026	0.001	-1.805	0.071	-0.005	0.000
quant_quality_firm	-0.0010	0.001	-1.013	0.311	-0.003	0.001