Zürcher Hochschule für Angewandte Wissenschaften

School of Management and Law

Advanced Quantitative Methods

Fall semester 2022

Building stable and robust models. Identify the relationships with most stable coefficients over time

Presented by:

Luca Bühler (buehlluc)

Eldar Hasani (hasaneld)

Silvano Probst (probssil)

Submitted to:

Tomasz Orpiszewski (orpi)

May 2022

Literature review

That a regression model is suitable for robust modeling of stock returns in global markets has been already shown in the past (Guerard, Xu and Markowitz, 2018, p. 6), but the selected firm-specific variables must meet two criteria: (1) the information must be statically significant and persistent, and (2) statistical techniques could be applied to extract the information (Guerard et al., 2018, p. 4). OLS analysis can falsely provide inflated as well as attenuated estimates of the various coefficients when the variance-covariance of the error terms is unknown (Greene, 1978, p. 461). It is true that the more multicollinear the regressors get, the less predictable the random error in the parameter estimate will be (Greene, 1978, p. 462). Outliers are not errors but are an essential part of the error distribution and must be accounted for (Nolan and Ojeda-Revah, 2013, p. 186 & Lambert and Lindsey, 1999, p. 409). For heavy distributions of the tails, the variances of the error terms can significantly affect the regression coefficients (Nolan and Ojeda-Revah, 2013, p. 186). Lambert and Lindsey (1999, p. 417) therefore believe that modeling the mode as a function of covariates would often be more appropriate in economic studies than other location parameters. However, the former is more difficult to interpret (Lambert and Lindsey (1999, p. 417). When modeling a regression model with time series, non-stationarity is a frequently observed issue (Dimitrova, 2005, p. 15). A nonstationary variable does not show constant means and variances over time and is more likely to correlate with recent lags. The result is inflated R^2 values and T-statistics (Dimitrova, 2005, p. 15). To check this, a Dickey-Fuller test can be performed (Omoruyi and Osaretin, 2015, p. 18). In addition, heteroskedasticity must be omitted, otherwise the error term does not exhibit constant variance (Dimitrova, 2005, p. 17). A stable model is provided by BLUE-estimation ("Best Linear Unbiased Estimator"), where convergence rates must be consistent with respect to the tails, error distribution, and distribution of the independent variables (Nolan and Ojeda-Revah, 2013, p. 187). In the study of stock values, the results of Siew and Nordin (2012, p. 2) showed that regression techniques can be improved if the data are standardized into a common data type. For example, converting real numbers to categorical ordinal data can enhance the results of regression techniques.

The efficient market hypothesis states that all information should already be priced into the market prices and thus the vector of coefficients is likely to be very close to zero (Guerard et al., 2018, p. 20). Prediction of a stock price would be possible when using macroeconomic variables (Omoruyi and Osaretin, 2015, p. 33), but such predictions on future stock prices turn out to be quite difficult (Bhuriya, Kaushal, Sharma and Singh, 2017, p. 510). The exchange rate between two different countries can be considered as a possible dependent variable. While an upward trend in the stock market, in the short run, tends to cause currency depreciation, a weak currency may cause a decline of the stock market (Dimitrova, 2005, p. 21). Gavin (1989) already showed that a booming stock market has a positive effect on aggregate demand. Industrial production, inflation, yield spread between long- and short-term government bonds would significantly explain stock returns (Omoruyi and Osaretin, 2015, p. 21). Industrial production and consumer price index positively influence the stock market, whereas money supply and inflation tend to show a negative relationship (Omoruyi and Osaretin, 2015, p. 21). Fama (1981) already showed the negative relationship between stock price and inflation.

Data description

We have selected ten stocks from ten different companies and four different continents as the data we need to apply to the model. The companies are placed either in Asia, Europe or the USA. All ten stocks are from the "technology" sector. This sector showed a very interesting development due to the great demand in the last ten years. Due to the interest rate hikes, which are likely in the coming months, this sector will also be one of the more interesting in the near future. Furthermore, we have selected only companies that are in an index that includes the largest capitalized companies in the country or continents.

Of these ten companies, the monthly closing prices over the last ten years are used. Thus, we obtain 120 month-end prices of ten different companies. In addition, to be able to test the volatility of the stocks for changes in global economic events, the OLS model requires macroeconomic data. The followings are relevant for this paper: “GDP”, “unemployment rate”, “consumer price index”, “wholesale price index”, “retail sales”, “industrial output”, “import”, “export” and “production capacity utilization”.

The data comes exclusively from Refinitiv and Bloomberg and is pulled directly from their database into our Python-based program. The same process is performed with Excel to check the data for their equality. In a second step, the data is checked on Python. Missing or incorrect data will be changed. After an initial analysis in Python, a database is created. Finally, so that the data can be used efficiently with Python for the model, it is exported and saved to an SQL database after the first analysis.

Share data analysis

After importing all macro as well as stock data, the analysis started. The whole dataset of the selected stocks has no missing data.

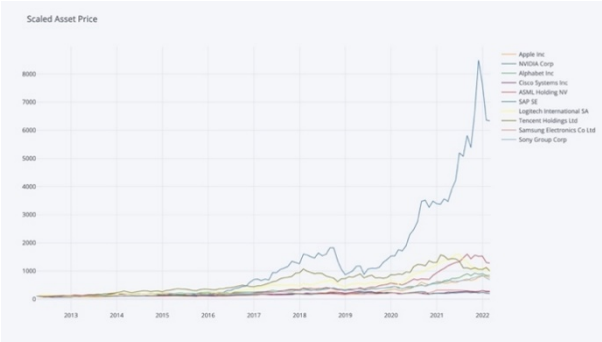

Figure 1 Section 3.2. "asset data normalized"

In a first step we analyzed the original data, i.e. the actual prices, as well as checked for normal distribution and stationarity. In the boxplots, presented under section 3.3, a trend can already be seen over the 10 years for each stock (each stock was presented using 11 boxplots and thus divided into 11 sub-periods). In the "descriptive statistics" section (3.4), it is easy to see that both the mean and the median increase strongly over the years, across all assets. The standard deviations also remain stable until 2020, although they have increased during the "Corona Crisis". The visual representation of all stock prices also reveals that there is no stationarity (p-

value: 0.73 or higher), which is basically normal. In the composition, the trend can be seen as well as some seasonality.

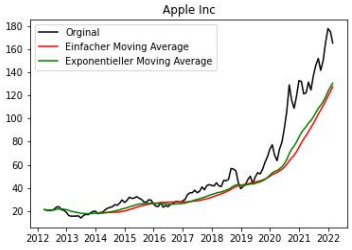

Figure 2 Section 3.5. "visual stationarity (Apple Original)"

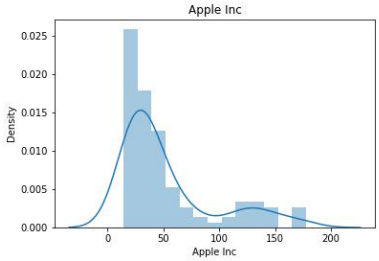

Figure 3 Section 3.6. "normal distribution (Apple Original) "

In a second step, we analyzed the "log returns" of the stocks and checked for normal distribution as well as stationarity. The median, mean and standard deviation are now constant around zero over the years. The visual representation shows that the entire data set of the stocks is stationary and can therefore be used for our model. This is also shown by the Dickey-Fuller test (p-Value: 0.049 or lower). Boxplots were also created for the "log returns" for each stock; now they form a line across the years. In the composition of the "log returns", it is now also apparent that the residuals follow a line and the trends have disappeared. As for the normal distribution of the "log returns", only the daily fluctuations of Nvidia and SAP are not normally distributed (p-Value: 0.023 & 0.000001).

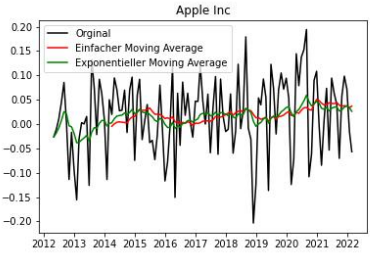

Figure 4 Section 3.10. "Stationarity view (Apple Returns)"

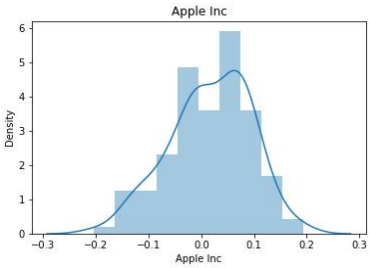

Figure 5 Section 3.11. "normal distribution (Apple Returns) "

Macro Data Analysis

For the macroeconomic variables dataset, in several of them were found some missing values. To fill the missing values, we used the "forward fill" method. The "ffill()" method replaces the "NA" values with the value from the previous row. For the import figures of South Korea, it has missing values from the beginning of our dataset, so the "back fill" method was needed. After these corrections, the entire data set is complete.

In order not to use highly correlated macro data in the OLS model and to finally build a stable model, we checked in section 4.2 the macro data for multicollinearity. For each country or continent, this allowed us to discard some macro data. For example, production capacity utilization was discarded since it correlates strongly negatively with unemployment rates in the

United States (p = -0.64). For the remaining macro data per country/continent, a boxplot was created. The closer together the boxplot is and the fewer outliers, the more stable the macroeconomic variable is.

For a stable model to be built, the macro data must also be stationary. About half of the macroeconomic variables are stationary in the "original", i.e. already from the first download. After converting the macro data to "returns," only the unemployment figures for South Korea remained non-stationary (p-Value: 0.13), which is why they were removed from the data set.

Modeling

In addition to the macro data mentioned above, we also include the lagged return (T-1) as a beta. With this we want to find out whether the price of yesterday has a significant influence on the price of today. When we first check the OLS models, we quickly see from the P-value that most of the macro data is not sufficiently significant. We have defined a minimum significance level of 5%. For this reason, the models were rebuilt by taking out the non-relevant betas.

fit_apple	Apple Inc ~ USGDPF=ECI
fit_nvidia	NVIDIA Corp ~ USGDPF=ECI
fit_alphabet	Alphabet Inc ~ USUNR=ECI
fit_cisco	Cisco Systems Inc ~ Cisco Systems Inc Lag + US...
fit_asml	ASML Holding NV ~ XTIMVA01NLM667S
fit_sap	SAP SE ~ DEGDP=ECI
fit_logitech	Logitech International SA ~ Logitech Internati...
fit_tencent	Tencent Holdings Ltd ~ 1
fit_samsung	Samsung Electronics Co Ltd ~ 1
fit_sony	Sony Group Corp ~ 1
fit_asml_eu	ASML Holding NV ~ EUPPI=ECI
fit_sap_eu	SAP SE ~ EUGDP=ECI
fit_logitech_eu	Logitech International SA ~ Logitech Internati...

Figure 6 Section 5.5. "rebuilding the models"

Section 5.5 now shows the new models with those betas that reach the desired significance level. All Asian assets such as Sony, Samsung and Tencent have been eliminated, as no P-value of 5% was found in the macro data. These three assets are therefore simply not eligible for our further processing. For Logitech, neither the CH data nor the EU data are describing the model, but the additional inserted lag mentioned before. For this reason, Logitech (EU) is omitted and only Logitech CH is used (as they are the same).

After the new selection regarding modeling has been made, the regression models are created and run again in chapter 5.7. All p-values of the betas are now below 5% and therefore enough significant. The highest adjusted R^2 values are found for Apple, Cisco, ASML and SAP_EU, whose values are all above 0.06. The adjusted R^2 value has rather decreased after omitting most of the irrelevant macro data.

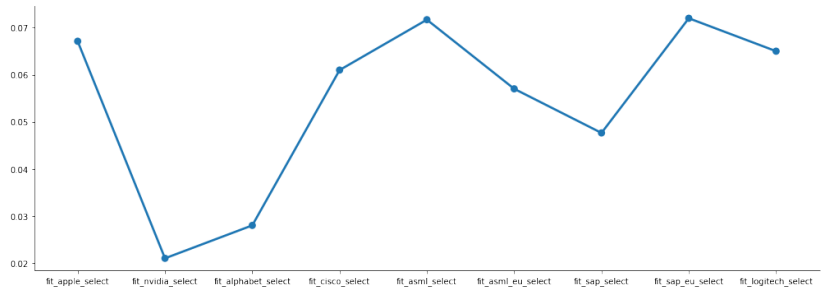

Figure 7 Section 5.8. "adjusted R2"

To check for heteroskedasticity, the residuals of the error terms are examined. In doing so, it quickly becomes apparent from the QQ plots in Section 5.9, that the residuals of Nvidia, ASML (EU) and Logitech are not normally distributed and thus have too many outliers. For this reason, these three assets were excluded for further review. The mathematical calculation by Jarque-Bera gives the final confirmation. After the Breusch-Pagan test, it is now also clear that heteroskedasticity is prevalent in the SAP data and therefore SAP is also excluded completely (CH and EU model). In the end, only four assets are still under review: Apple, Alphabet, Cisco and ASML. No autocorrelation was found for any of the remaining models.

Final selection of the model

Before the most robust model is selected, the last remaining models are checked for their Mean Squared Error (=MSE). It is a quality criterion and is interpreted as the expected squared distance an estimator has from the true value. The smaller this number is, the smaller the bias as well as the variance of the estimator is and thus with high probability also closer to its expected value. In our example, the model of Alphabet Inc shows the lowest MSE with 0.00371, followed by ASML (0.00484), Cisco (0.00491) and finally Apple (0.0060).

As a next step the betas were now checked using a rolling regression model and it was found that basically all betas show stable values over the entire time. The stat model under section 5.10.2 in python also confirms this. A standard rolling window of 60 months (5 years) was used. Finally, the remaining models were tested with the Chow-test. It tests the coefficients of two linear regressions for equality and thus checks for structural breaks in the middle of data stream (after the 60^{th} value out of 120). All models have a p-value of more than 0.05 regarding the Chow statistic, which means that the hypothesis cannot be rejected and that there is no structural break.

After checking the models at all major statistical levels, it was noticed that our last four models do not differ much among each other, so selection was quite difficult. Nevertheless, we propose the model on the Cisco enterprise as the most robust and stable one.

$$Cisco\ Asset = Constant(X0) - Cisco\ Systems\ Inc\ Lag(X1) + USRSL(X2) - USIP(X3)$$

$$Cisco\ Asset = 0.0075 - 0.2093(X1) + 0.0080(X2) - 0.0121(X3)$$

The following reasons were important in our decision:

1. It is the only model, with more than 2 significant betas (lag (T-1), US Retail sales and US Industrial Output). The betas were shown to be stable and robust over the entire run after rolling regression.

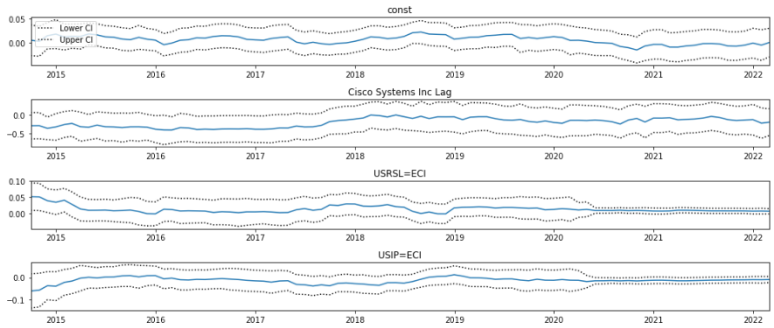

Figure 8 Section 5.10.3. "rolling regression"

2. The adjusted R^2 is around 6.2%, which is the third highest value of all the models reviewed. Moreover, there is no heteroskedasticity and no autocorrelation and no structural break.

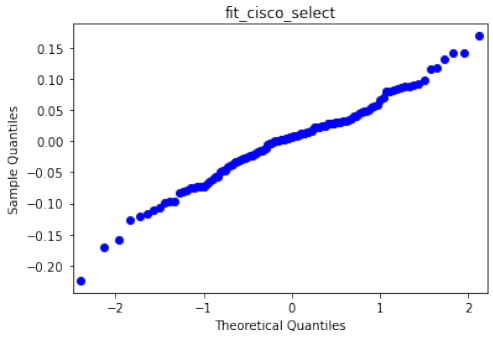

Figure 9 Section 5.9.1. "QQ-Plot"

3. The MSE is quite low at 0.00491, reflecting that the estimator with a high probability is close to the expected value. The MSE was also one of the lowest of all the models reviewed.

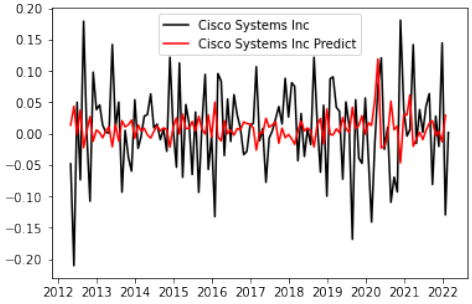

Figure 10 Section 5.9.4. "MSE"

4. The Chow Test shows a value of 0.367 which means no structural break is existing.

Sources

Bhuriya, D., Kaushal, G., Sharma, A. & Singh, U. (2017). Stock market predication using a linear regression. International conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 510-513. doi: 10.1109/ICECA.2017.8212716.
Dimitrova, D. (2005). The Relationship between Exchange Rates and Stock Prices: Studied in a Multivariate Model. Political Economy, 14.
Fama, E. F. (1981). Stock Returns, Real Activity, Inflation and Money. American Economic Review, 71(4), pp. 545-565.
Gavin, M. (1989). The Stock Market and Exchange Rate Dynamics. Journal of International Money and Finance, 8, pp. 181-200.
Greene, V. L. (1978). Aggregate Bias Effects Of Random Error In Multivariate OLS Regression. Political Methodology, 5(4), pp. 461–467. Retrieved from http://www.jstor.org/stable/2579155.
Guerard, J., Xu, G. & Markowitz, H. (2018). A Further Analysis of Robust Regression Modeling in Global Stocks. Retrieved from SSRN: https://ssrn.com/abstract=3190716 or http://dx.doi.org/10.2139/ssrn.3190716.
Lambert, P. & Lindsey, J.K. (1999). Analysing Financial Returns by Using Regression Models Based on Non-Symmetric Stable Distributions. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48, pp. 409-424. Retrieved from https://doi.org/10.1111/1467-9876.00161
Nolan, J. P. & Ojeda-Revah, D. (2013). Linear and nonlinear regression with stable errors. Journal of Econometrics, Volume 172, Issue 2, pp. 186-194.
Omoruyi, A. & Osaretin, A. (2015). THE IMPACT OF MACROECONOMIC VARIABLES ON STOCK MARKET INDEX IN NIGERIA. African Journal of Management Sciences (AJMS), Vol. 1(1), pp. 18-40.
Siew, H. L. & Nordin, M. J. (2012). Regression techniques for the prediction of stock price trend. International Conference on Statistics in Science, Business and Engineering (ICSSBE), pp. 1-5. doi: 10.1109/ICSSBE.2012.6396535.