Prediction of S&P500 Stock Index Using ARIM and Linear Regression

Kefei Guo[1, †], Zifan Jiang[2, †] and Yujian Zhang[3, *, †]

[1]Qingdao No.9 High School, Canada Qingdao Secondary School, Qingdao, Shandong, 266071, China

[2]Department of Mathematics, Xi'an Jiaotong University, Xi'an, Shanxi, 710049, Changsha, China

[3]Department of Mathematics, University of Florida, Gainesville, FL 32603, United States

*Corresponding author: zhangyujian@ufl.edu

†These authors contributed equally

Abstract. This paper mainly establishes a linear model suitable for the volatility of the S&P 500 index and forecasts the S&P 500 index. Firstly, the data set is divided into the training set and test set. After testing a series of data attributes such as the smoothness of the original data series and log series, the original data series and log series of the S&P 500 index weekly data series are modeled based on the ARIMA model. The next step is to check the fit of the model and use ACF and PACF to determine the parameters of two different models to fit the original data series and the log data series, respectively. Based on two different models, the rationality of the model is confirmed by the residual white noise test and various natural maps. By establishing the model and analyzing the residual error of the model, finding out unreasonable fluctuation of the residual error of the model fitting and giving the corresponding explanation combined with the history. Finally, a fitted model to make rough forecasts for the S&P 500 from January 2020 to December 2020. Although this forecasting model cannot predict detailed fluctuations daily, it can still correctly determine whether a stock is going up or down. To sum up, the ARIMA model does not perform well in stock forecasting and it may need to be improved using other methods.

Keywords: S&P500; ARIMA; simple linear regression; LSTM; stock price forecasting.

1. Introduction

1.1. Background

The Standard & Poor's 500 Index, sometimes known as the S&P 500 Index, is a stock index that tracks the 500 publicly traded American firms. The base period index is set to 10 and the base period is defined as 1941–1943. calculated using the weighted average approach, utilizing the base period weighted computation and the number of listed stocks as the weight. The S&P 500 is typically regarded as the perfect stock index futures contract because it differs little from the Dow Jones Industrial Average in terms of sampling size, strength of representation, accuracy, and continuity.

S&P excels in developing impartial benchmarks. The credit rating of Standard & Poor's has garnered considerable interest from international investors due to its impartial analysis and distinctive insights that accurately represent the solvency and willingness of governments, businesses, and other organizations to repay their loans.

S&P is a pioneer in exchange-traded funds and index tracking systems on the stock market (ETFs). Additionally, the company's database standardizes the data of listed companies, making it simple for financial staff to compare data across multiple categories. The internet services offered by Standard & Poor's are an efficient resource for analysts, planners, and investors everywhere.[1].

1.2. Related research

In recent studies, Roman Chuyan and Professor Alexander Melnikov presented a dynamic multifactor linear regression model to describe the S&P500 index, including a broad set of variables employed by practitioners for modeling and forecasting equities, as well as major macroeconomic indicators [2]. In another study, the research group applied the linear regression model based on the trading volumes, which is the number of lots bought and sold which is expressed on daily basis, and successfully construct a linear relationship between the stock price and the trading volume. Their objective is to predict the stock market behavior and assist investors to act with greater certainty and know when to buy and sell at a good price based on the prediction [3].

Some scholars use the ARIMA model for stock prediction. Nowadays, the ARIMA model has been improved to get better predictions. Some researchers use ARIMA to cooperate with other models, like LSTM, SVM, and GARCH. Meizhen Liu and Chunmei Duan conducted a comparative analysis and trend forecast of China's CPI data from 1951 to 2008 based on the ARIMA model [4].

In the study of Jakub Michankow, Pawel Sakowsk,i and Robert Slepaczuk, the LSTM model was used to predict the BTC and S&P500 index, the data from 2013 to 2020 are used[5]. To improve the prediction ability of the LSTM model, they also created a new loss function. They analyzed and described the obtained data, performed sensitivity analysis on the methods and data they used, and finally combined the obtained signals of different frequencies into an effective integrated model. In their hypothesis, the efficiency of the LSTM in algorithmic investment strategies depends strictly on the hyper parameter adjustment procedure, the model construction and the estimation process. In addition, the proper loss function is very important in the modeling estimation process. Moreover, the results depend on the asset class and frequency of use. Finally, we note that the results are not reliable to the original hypothesis [6].

Andrea Leccese's research examines the performance of a deep learning method represented by the S&P 500 index (LSTM). And they found that the LSTM seemed to be effective at predicting the value of the next day, but predicts the value of the following day very similar to the real value of the previous day[7].

1.3. Objection

To better understand and predict the future trend of the stock market, the data of S P500 from 1941 to 1943 need to be plotted. Using these data to find the current trend and situation of these data, and use them to calculate and predict the future changes and trends of the stock market. First, select two models -- ARIMA and linear regression models -- and analyzed the data with two different models. Then, analyzing and reasoning the results very logically, and predicting the future data by using the two models. Finally, comparing forecast data with actual data to test the accuracy of models and analyze the reasons for the difference and ways to improve it. Predicting equity prices can help individuals and businesses understand where and how to invest in the market and minimize the risk of losing money [8].

2. METHODOLOGY

2.1. Stationarity Test

Then, after obtaining a time series, how to test its stationarity? Currently, there are two principal methods for testing the stationarity of time series. The first is the graphical test method, which is to make an intuitive judgment based on the time series chart and the auto-correlation chart. Another way is to construct a test statistic. Testing method. For the image inspection method, we generally draw a time series graph of the time series, as shown in the following figure. If a time series is stationary, then the series should fluctuate randomly around some mean. On the other hand, we can also use the auto-correlation plot to test. The auto-correlation diagram of stationary time series generally increases with the increase of the order, and the auto-correlation system will rapidly decay to close to 0, while the non-stationary time series may have changes that first decrease and then increase or periodically fluctuate.

2.2. Model

2.3.1 Linear regression

Linear regression is a linear model that describes a linear relationship between the input independent variable X_i for i = 1, 2, 3, … and output dependent variable Y. If there is a single input variable X, then it is called simple linear regression. When there is more than one input variable $\{X_1, X_2, \ldots, X_n\}$, the method refers to multi-factor or multi-variable linear regression, as shown in equation (1).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{\mathrm{i}2} + \cdots + \beta_{\mathrm{p}} X_{1\mathrm{p}} + \epsilon_i \qquad (1)$$

where "i" denotes the scores of the "i" and p denotes the total number of predictors. ϵ_i stands for the intercept while all X_i goes to zero, and the coefficients between dependent and independent variables are given by the regression coefficients $\beta_1, \ldots, \beta_p$. The residuals (ϵ_i) are set to be normally distributed by assumption with mean 0 and unknown variance σ^2 [9]. The model of linear regression is simple compared to modern machine learning models and can be explained with mathematical interpretations and formulas to obtain predicted values.

2.3.2 ARIMA model

The ARIMA model, also referred to as the self-regressive moving mean model, was proposed by the American statistician Jenkins and the British statistician Box in the 1970s. This model is used to forecast short-term time series variables because it is really hard to predict a single time series value, but there is a certain law in the overall time series value, and the model of ARIMA is used to express this law in mathematical form. In order to have short-term prediction of time series variables. And by studying its mathematical form, short-term forecasting of the value of time series is carried out [10].

3. Numerical experiment

3.1. Data

We downloaded weekly S&P 500 data from January 1980 to December 2020 from Yingwei Finance, which is nearly 40 years. Use the data from December to December 2020 as the experimental data. In Figure 1, the abscissa is time, and the ordinate is price. **Fig.** 1 below shows the fluctuation curve of the S&P500 data.

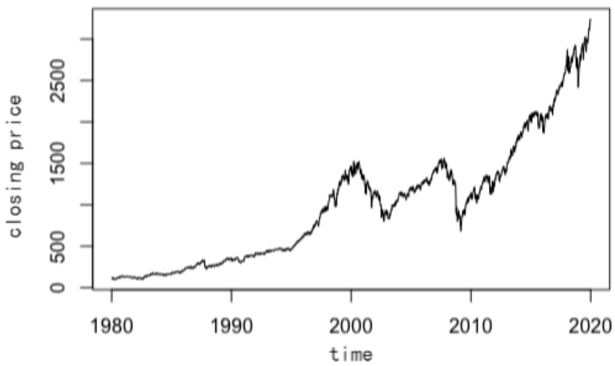

Fig. 1 The raw data of S&P500

Photo credit: Original

3.2. Model prediction

3.2.1 Prediction results based on linear regression

1) Summary of linear regression model using R

In the model, we construct a linear relationship between the closing price and time. The p-value represents a very small value, meaning a statistically significant result, as shown in Table 1.

Table 1. Residuals in linear regression model & p-value test

Residual Analysis				
Min	1Q	Median	3Q	Max
-865.87	-154.72	-31.07	154.21	1044.96
p-value: <2.2e-16				

2) Prediction results

In the prediction **Fig.** 2, the upper plot represents the real closing value with time from 2020.1 to 2020.12. Respectively, the lower straight line represents the prediction data during the period. We can conclude that the S&P500 index shows an increasing trend given that the coefficients have a gradient of 1.145. However, the accuracy needs further research to be improved.

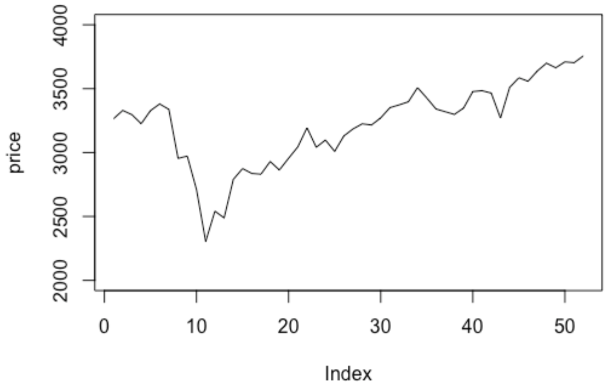

Fig. 2 The prediction data and realistic data of S&P500 from 2020.1 to 2020.12

Photo credit: Original

3.2.2 Prediction results based on ARIMA

1) Logarithmic rate of return

The log rate of return is the difference between the log values of an asset over two time periods, that is, the log rate of return of an asset over more than one period is equal to the sum of the log rate of return over each period. When looking at stock prices, one assumes that the price model follows the Brownian movement, that is, that the logarithmic yield is normally distributed. However, empirical statistics on actual market data show that the log yields of most inventories are not following a normal distribution.Therefore, while analysis of closing prices is often based on stock returns, stock returns can be divided into simple and logarithmic returns. Therefore, the logarithms data is used to set the model. **Fig.** 3 below is the logarithms data of S&P500.

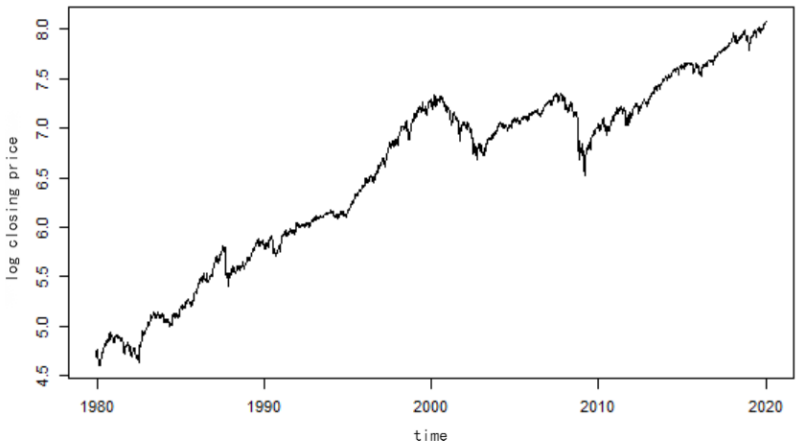

Fig. 3 The logarithmic data of S&P500

Photo credit: Original

2) Logarithmic series trend analysis

Firstly, consider serializing to a fixed sequence. Observing the original sequence image and the logarithmic ally processed image, it can be found that the logarithmic processing alleviates the problem of variance fluctuation to a certain extent. But there is still significant growth in the sequence, which can be visually observed by decomposing the sequence, as shown in **Fig.** 4.

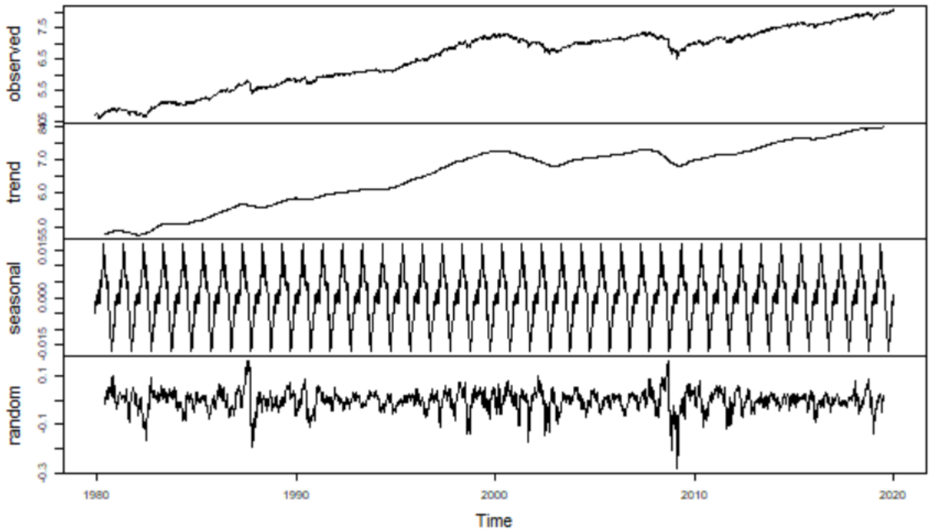

Fig. 4 Decomposition of additive time series

Photo credit: Original

3) Difference and Stationarity Tests for Logarithmic Series

To remove strong growth, differential processing was used, and a series of stationarity tests were performed using the Augmented Dickey-Fuller Test. After the first-order difference, it can be seen that the p-value is 0.01, rejecting the null hypothesis and the series is stationary, as shown in Table 2.

Table 2. Augmented Dickey-Fuller Test

Augmented Dickey-Fuller Test	
Data:	The logarithmic data of S&P500
Dickey-Fuller:	-13.278
Lag order:	12
P-value:	0.01
Alternative hypothesis:	stationary

4) Serial auto-correlation

The ACF and PACF charts do not have visible truncation and tail, and it is difficult to judge the subsequent prescription AR() and MA(). The EACF plot is also difficult to judge the coefficients of the ARIMA model. After several experiments on the model, it can be chosen to use the functions that come with the R language for judgment, as shown in **Fig.** 5-7.

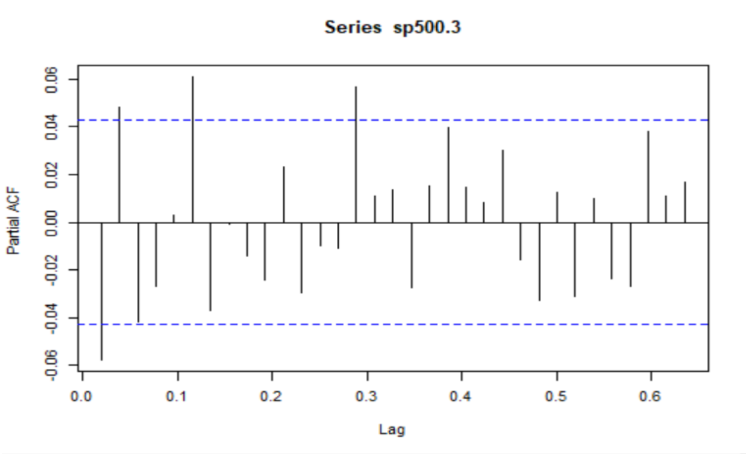

Fig. 5 The ACF graph

Photo credit: Original

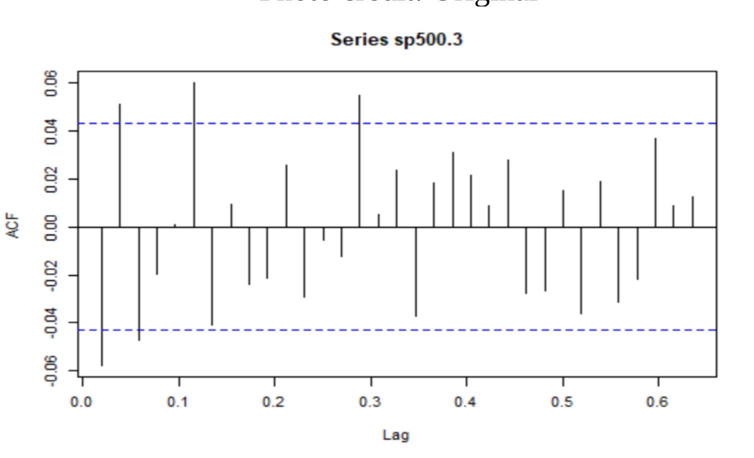

Fig. 6 The PACF graph

Photo credit: Original

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	o	o	x	o	o	o	o	o	o	o	o
1	x	o	x	o	o	x	o	o	o	o	o	o	o	o
2	x	o	x	o	o	o	o	o	o	o	o	o	o	o
3	x	x	x	o	o	o	o	o	o	o	o	o	o	o
4	x	x	x	o	o	o	x	o	o	o	o	o	o	o
5	x	x	x	x	o	o	x	o	o	o	o	o	o	o
6	x	x	x	x	x	x	o	o	o	o	o	o	o	o
7	o	x	x	x	x	o	x	o	o	o	o	o	o	o

Fig. 7 The EACF graph

Photo credit: Original

5) Finding the Optimal ARIMA Model Based on AIC Criterion

Table 3. auto.arima() in ARIMA model

ARIMA(1,1,1)(1,0,0)[52] with drift				
Coefficients				
	ar1	ar2	ar3	ar4
	-0.7265	0.6694	-0.0224	0.0016
s . e .	0.1210	0.1302	0.0221	0.0005
Sigma^2 estimated as 0.0004976:		Log likelihood=4974.94		
AIC = -9939.89	AICC = -9939.86		BIC = -9911.67	

From Table 3, we can see that the model given by auto.Arima() is an ARIMA model with a seasonal difference with a drift term. In the R language, the meaning of the seasonal difference is to add a lag of 4th order to the AR () and MA() models, but, since we choose weekly data, the seasonal difference is an extra The fourth-order difference term of, which is essentially a monthly difference, aims to eliminate the monthly periodicity.

6) Residual Analysis

According to Table 4, Logarithmic series models cannot be identified as pure white noise. However, when the relative error is consistent, a large amount of absolute error occurs. Thus, the arch Test should be considered to look for other amounts of information in the residuals.

Table 4. Residual White Noise Test

Box-Ljung test		
Data:el		
X-square = 67.146	df = 522	P-value = 0.07702

The LM test is significant no matter how many periods the series lag, that is, there is an arch effect in the residual series of the model. To judge from the figure below, it can be seen that GARCH (1,1) can be used to fit the residuals, as shown in **Fig.** 8.

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x o o o o o o x x o o  o  o  o
2 x x o o o o o x x o o  o  o  o
3 x o x o o o o x x o o  o  o  o
4 x x x x o o o o x x x  o  o  o
5 x x x x x o o o x o o  o  o  o
6 x x o x x x o o x o o  o  o  o
7 x x x x x x o o x o o  o  o  o
```

Fig. 8 The EACF Pragh

Photo credit: Original

7) Build the GARCH model

All coefficients are significant. The results of the test show that the residuals of the Garch model are non-normal, and the square of the residuals is a white noise sequence, as shown in Table 5.

Table 5. The model of GARCH (1,1)

Model:GARCH(1,1)				
Residuals:				
min	1q	median	3q	max
-7.2800	-0.4953	0.1696	0.7276	3.6873
Coefficients:	Estimate	Std.Error	T-value	Pr(>|t|)
a0	0.214745	0.101712	2.111	0.0347
a1	0.163012	0.012919	12.618	<2e-16
a1	0.859259	0.009722	88.382	<2e-16

3.3. Discussion

Our forecast data has a large deviation in the first half of 2020. We speculate that it may be caused by the unexpected and sudden impact of the new crown epidemic in 2020, and the fluctuation of the S&P 500 index data is abnormal.

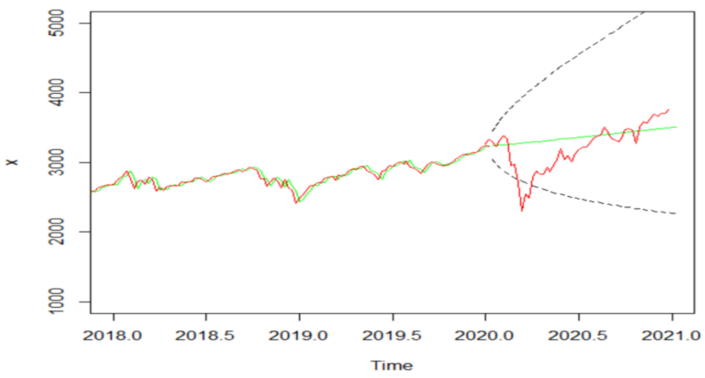

Fig. 9 Prediction data using the ARIMA model

Photo credit: Original

From **Fig.** 9, the predictive effect of another model using the original data sequence is worse than that of the logarithmic model, but the model using the original data sequence is a pure ARIMA model (1, 1). There are no seasonal differences and drift terms, which are simpler than models based on logarithmic series.

On the other hand, the prediction data of the simple linear regression model obtains similar results as the ARIMA model. The overall trend of the S&P 500 index shows an increasing tendency in the future stock market. The model a applies simple linear regression approach, meaning that only one independent variable is used in the calculation. Recent researchers tend to use multivariable linear regression analysis other than simple linear regression to obtain more accurate predictions in further research.

4. Conclusion

4.1. Summary

Our team analyzed and calculated S P500 data from 1941 to 1943. Based on this information, two different models were used in analysis: ARIMA and linear regression model to verify that our curves were roughly the same as the real data by comparing the data and images analyzed with the real data. These two models to predict image trends and data for S P500 after 1943. The two models that were used predict roughly the same result - the overall S P500 index will show an upward trend in the future stock market.

4.2. Limitations & Future outlooks

In comparison with real data, some errors were in the model predictions. First, founding that many different social and environmental factors, such as the COVID-19 pandemic and the economic downturn, can affect the trend of the stock market. And that's when our predictions are going to be way off. Second, our model can only predict the general trend, and there is no way to accurately predict the trend daily. Therefore, the trend predicted by our model is consistent with reality, but it will deviate when the accuracy is daily.

References

[1] "Introduction to Financial Data Analysis: Based on R Language" by Ruey S. Tsay

[2] Meln ikov, Alexander & Chuyan, Roman. (2012). On linear regression analysis for modeling and forecasting of S&P500 Index. Risk Management (Russian). 2-6.

[3] Gharehchopogh, Farhad Soleimanian, Tahmineh Haddadi Bonab and Seyyed Reza Khaze. “A LINEAR REGRESSION APPROACH TO PREDICTION OF STOCK MARKET TRADING VOLUME: A CASE STUDY.” (2013).

[4] M. Liu and C. Duan, “The Analysis of China CPI Trend Forecast Based on ARIMA Model,” Proceedings of the 2018 International Conference on Management, Economics, Education and Social Sciences (MEESS 2018). Atlantis Press, 2018.

[5] He S Y. 2003. Applied time series analysis. Peking University Press.

[6] Michańków, Jakub, Paweł Sakowski, and Robert Ślepaczuk. 2022. "LSTM in Algorithmic Investment Strategies on BTC and S&P500 Index" *Sensors* 22, no. 3: 917. https://doi.org/10.3390/s22030917

[7] Bluesky Capital, and Andrea Leccese. “Machine Learning in Finance: Why You Should Not Use LSTM’s to Predict the Stock Market.” Bluesky Capital, 13 Jan. 2019, https://www.blueskycapitalmanagement.com/machine-learning-in-finance-why-you-should-not-use-lstms-to-predict-the-stock-market/.

[8] Liu Song, Zhang Shuai. An empirical study on stock price prediction using ARIMA model. Economic Research Guide, 2021.

[9] S. Ravikumar and P. Saraf, “Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms,” 2020 International Conference for Emerging Technology (INCET). IEEE, Jun. 2020. doi: 10.1109/incet49848.2020.9154061.

[10] Bergh, D.v.d., Clyde, M.A., Gupta, A.R.K.N. et al. A tutorial on Bayesian multi-model linear regression with BAS and JASP. Behav Res 53, 2351–2371 (2021). https://doi.org/10.3758/s13428-021-01552-2.