Finding a Prediciton Model for Short-term Stock Price Movement through Logisitic Regression with Ordinal Outcomes

STAT 835

Tyler Corcoran

Department of Biostatistics and Data Science
University of Kansas, USA
May 12, 2021

Contents

I. Abstract

Stock price movement can be chaotic and subject to unclear market signals and over reactions causing difficulty in short-term trading and investing, and is the primary reason most financial literature for individuals suggest only viewing the market in a long term perspective. However, hedge funds, portfolio managers, and day traders are able to realize large short-term financial gains regularly. It's well known that professional traders use a combination of subjective assessment of the market via technical indicators and sentiment in addition to macro level influences that affect global markets. We attempt to show that a simple, objective, decisioning system underlies all of this and attempt to capture this in an ordinal cumulative logistic regression model. We were unable to capture a full model that was significant, but were able to discern important variables that contribute significantly more than others. We show that these variables have a strong inverse influence on market changes over a short term perspective. The significance of our research attempts help build a more comprehensive understanding of simple, but powerful effects that can act as primary decisioning mechanisms on their own, or in a less naive model as is popular through decision trees, machine learning, and other advanced model training procedures.

II. Introduction

Stock price prediction algorithms have been of great interest for corporate and personal gains since the early days of stock market trading. There have been many methods attempted in this process using everything form machine learning, to basic linear regression analysis. With the ease of access to large volumes of data aggregated in real-time currently most of these methods use highly advance training procedures on models that adjust frequently with incoming data. The data read by these algorithms include fundamental analysis, technical analysis, news, insider actions, dark pools (though difficult to ascertain in a real-time world), and others.

Here we attempt to use a subset of the available technical indicators in a less complicated procedure than most modern algorithms to determine if there are any underlying data points that can predict Stock Price % Change over a week. We hope to fit a model that provides the probability of stock price movement with reasonable certainty. Reasonable certainty is a fairly loose measurement but we've observed that some of the most sophisticated models have a hit percentage of ~60%. That is, using the sophisticated model to guide trading decisions, they are able to win on trades about 60% of the time. Therefore, if we can find a subset of predictors, under specific criteria, that allow for a response probability near 60% we would consider this analysis a success.

The main categories of data used here are those which are accessible without unreasonable costs and time, and are universally accepted as key indicators of stock price movement by market participants and traders. The categories include technical analysis, and dark pool data. However, this is not a comprehensive sample of all key factors that contribute to a global market that's affected by macro level politics, currency, pandemics, and unkown insider data that all influence stock price movement. Because of this, it is most important to us to find a very small subset of predictors that can be relied on as an underlying calculation of the macro level influence of stock prices.

A. Basics of Technical Analysis

Technical analysis encompasses all parameters that are calculated via the available statistics around a stock such as Open, High, Low, and Close price; volume, notional value, order entry times, float, % float, short float, % short float, market capitalization, and others data points that are available to a market participant and observer. From these statistics there are hundreds of studies and technical indicators that are used to create dependent functions indicating propensity for performance around a stock. The basic indicators/studies of most interest here are listed in Appendix 1.

B. Basics of Dark Pools

Public exchanges are where the majority of stock transactions occur and the transaction data is available in real-time to the informed market participant. In addition there are dark exchanges or dark pools, where transactions are not available to the public and the data reporting can be delayed for up to two weeks based on current regulations. However, even though this data is delayed and private, the transactions do still occur in the market in real-time. There are many competing theories concerning the fairness of dark pools to the

overall market, hedge funds, and retail investors and traders, which is beyond the scope of this paper. However, in general it's assumed that dark pools allow for price discovery when there is a low amount of information available, therefore making the cost of information high. Because of this it's assumed that during periods of high information costs (highly anticipated earnings, high amounts insider trading, very low volatility, etc.) market participants will use dark pools for purposes of "price discovery".

Additionally, the dark pools are used during periods where there are not high information costs in order to hide trade transactions from the public, High Frequency Trading Systems (HFTs), and other algorithms. The objective is for the Dark Pool participant to buy or sell a large number of shares without causing a rippling effect exacerbated by other market participants who would normally use that information printed in Time & Sales to initiate a buy/sell. It's also assumed that the majority of dark pool transactions are executed by hedge funds and similar entities with extremely large asset values.

III. Methods

Our general research hypothesis is that a key subset of technical indicators can reasonably predict the outcome of a % change in stock price movement over a week, based on the levels of the explanatory variables collected concerning the prior week. Therefore, we approached the problem through a logistic regression framework. The predictors of interest contain both continuous and categorical structure. For this reason it seems reasonable to fit a linear model, however because the potential variables of interest in our study can be classified as Technical Indicators, where the underlying parameters all rely on price, volume, and timing, a logistic regression approach may help improve the power of analysis. Additionally, it's actually more useful to have thresholds to act upon for categories of response. This would allow an opportunity to better separate the trading strategy for stock, options, and futures setups (long and short positions) based on the category of outcome with highest/lowest probability response.

Therefore we separated the continuous response into 5 ordinal categories as listed in 1 for % change in stock price. This approach is supported by Agretsi (2013), where it's stated, that if you can reasonably foresee an ordinary linear model with some explanatory variables describing the data in a linear fashion, then the same variables should apply to just as reasonably to the case of discrete outcomes using the cumulative logit model. Additionally, power is increased in an ordinal test, therefore we deemed an ordinal logistic regression approach as being valid.

Our research objective is ultimately interested in discovering a model that can provide outcome probabilities for stock price % change under specific criteria. Through this discovery we are also hoping to identify the key predictor variables that signal market behavior more than others. This information can be used for further research since we expect that a method that stops at logistic regression, like this one, is likely naive for a comprehensive understanding of underlying effects of stock price movement. The classification of effects in logistic regression is very convenient approach, as we can retrieve the odds ratios for certain parameters easily and their extreme values have at least some interpretability in a model that exhibits poor fit. We used RStudio and Microsoft Excel for all data calculations and

model building, primarily relying on vglm() in the VGAM package in R.

The cumulative logit model relies on a proportional odds structure in the cumulative logit model as follows.

$$logit[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \ j = 1, ..., c-1$$

where for the response variable Y, the cumulative probability for outcome category j is

$$P(Y \leq j) = \pi_1 + ... + \pi_j, j = 1, .., c.$$

Cumulative probabilities reflect the ordering with,

$$P(Y \leq 1) \leq P(Y \leq 2) \leq ... \leq P(Y \leq c) = 1$$

A purposeful selection procedure as outlined by Hosmer et al. (2013, Chapter 4) was used to assess potential models of interest. Due to the potential multicolinearity of our technical indicators, and their sensitivity to outcome in logistic regression, we used a backwards selection process to help reduce the chances of this. We compared fit statistics including, deviance residuals, AIC, pseudo-R^2, AUC/ROC, and response plots to determine significance. The thresholds for significance were set at $\alpha = 0.05$, based on convention.

Additionally, a test for proportional odds structure of the cumulative logit was performed for Model 1. We compared this model to two other potential models, Model 2 and Model 3, to test for improved fit and structure due to an inability to robustly test the proportional odds structure Model 1, assuring no violation. The model of best fit was determined by heavy preference to parsimony, minimized standard errors, confidence intervals, and pre-established assessments of power relating to logistic regression types with data structured similar to ours.

A. Data Gathering

Stock price data and technical analysis was collected from My Barchart Premier (barchart, 2021), and FlowAlgo.com (FlowAlgo LLC, 2021). We first reviewed the highest average volume stocks over the previous 52 weeks, from May 2019 to May 2020 for ETFs, Mega-Caps, Large-Caps, and Mid-Cap stocks and selected the top three from each market capitalization category. Then pulled all data for the respective Stock over the past 3 years, May 2018 to May 2021. Some of the technical analysis metrics needed to be calculated manually as it was not available from Barchart. Additionally all dark pool data was retrieved from FlowAlgo.com under the same time intervals for each stock, there was also some manual calculation for this data.

After all data was collected it became apparent that some of the stocks we selected had recent IPOs, and or were strong momentum stocks that contained very skewed data points where the stock value surged over a 1 week to 2 month period and then retraced 50-80%. In order to create a robust model we then narrowed our selection of potential stocks to the most highly traded overall by volume. Due to time constraints and time of analysis for each stock symbol we decided to only use the SPDR S&P 500 Trust ETF (ticker symbol: SPY) for our analysis. While we would like to use this same research approach on other stocks

later, this is reasonable preliminary choice as this stock is an indexed ETF that references the S&P 500, one of the largest indexes in the world and a frequent point of comparison for minimum achievable yield for portfolio mangers, hedge funds, and signs of economic productivity. Therefore a model that could reasonably fit this data would potentially have application elsewhere in the market.

B. Data Calculations

Technical analysis data that was not readily available from Barchart was calculated using formulas and methods provided by reputable financial literature found readily online. These calculated parameters included Relative Volume, Relative VWAP, and Average-ATR. Average-ATR was calculated as the average of the SMA and SMA/EMA adjusted ATR, as a conclusion couldn't be made on which smoothing procedure would be best. We concluded that attempting to fit both methods in the same model would cause too much colienarity, for a data set that is already subject to many other related parameters. Additionally, it's reasonable to assume that there is not a drastically significant difference between the SMA-ATR and EMA-ATR smoothing when were able to invoke asymptotic convergence through a large sample size, which is a necessary premise due to a large number of initial predictor variables. The ordinal response variable was discretized at pre-specified cut points described in Table 1. These cut points were determined from experience and confirmed with other colleagues as being significant take-action thresholds for different trade setup opportunities if predicted by an adequate model.

Table 1: Ordinal Categorization for Response Variable, Stock Price Open-Close Change.

j	Interval	Variable Parameter Name
1	$-1.8\% < \%SP_{\triangle}^{W} \leq -0.9\%$	dn_2
2	$-1.8\% < \%SP_{\triangle}^{W} \leq -0.9\%$	dn_1
3	$-0.9\% < \%SP_{\triangle}^{W} < 0.9\%$	flat
4	$0.9\% \leq \%SP_{\triangle}^{W} < 1.8\%$	up_1
5	$\%SP_{\triangle}^{W} \geq 1.8\%$	up_2

The weekly aggregated Relative Volume Signal was categorized in a similar fashion, where if the relative volume was above, between, or below a threshold it was categorized as Up, Flat, Down. Relative volume values of 100% would indicate that the stock is trading at normal volume levels for the time frame of interest.These values are a very conservative estimate of trading signals for stocks for market participants, as relative volume can be high as 400-500% during strong momentum, but rarely drops below 75%, and typically will be between 150-200% during a normal uptrend.

Our potential variables of fit for the model are listed below. These are all identified as key technical indicators that influence market dynamics and trader psychology therefore provide an opportunity to best understand behavior cues of market participants in a model, supported by our experience and research from Bitvai and Cohn (2014, pp. 195–197).

Table 2: Thresholds for Previous Relative Volume Signal Variable

Factor	Threshold	Variable Parameter Name
1	%Vol^W_{rel} <80%	down
2	80% <%Vol^W_{rel} \leq 150%	flat
3	%Vol^W_{rel} \geq 150%	up

Table 3: All variables at time t reference values observed at t-1 due to model structure

Variable	Variable Name
Relative Strength Index (RSI) - 15 day	prev_RSI_15day
Relative Average Volume -Weekly (Referenced to 3 Month Average Volume)	p_rel_vol_wk_3mo
Open-Close % Change - Weekly	prev_OC_per_change
Average True Range (ATR) - 15 day	prev_rel_ATR_15day_avg
Low Price Relativity to Volume Weighted Average Price (VWAP) - Weekly	prev_rel_low_VWAP_week
Count of Dark Pool Transactions - Weekly	prev_dp_count
Volume of Dark Pool Shares Transacted - Weekly (100K)	prev_dp_volume_100K

C. Data Cleaning

After collecting our data we reduced our sample size slightly due to the technicality in how some technical indicators are calculated. Several of these indicators rely on previous history in their equations, therefore there was a maximum of 9 weeks of data for some of these indicators where null values were present. Since there was only minor reduction in sample size, removing these 9 data points compared to dealing with null values we restricted our data set to roughly May 2021 to July 2018, giving us n = 204.

D. Additional Statistical Model Considerations

1. Normality

The univariate distributions for all variables produced seemingly dissimilar results comparatively, with many distributions appearing non-normal from the histograms below. However, many methods related to asset returns in the field of economics are considered to be normal even though extreme observations have a tendency to give slightly fatter tails than the normality would imply (Yiu, T., 2020). This is partly due to the simplicity and broad applicability of modeling under the normal, but is also supported by the central limit theorem and asymptotic convergences over large periods of time that can be used to transform the definition of a valuable asset–one that takes value that move from lower lows to higher highs over a long period of time.

From the univariate observations on our response variable, there's an increasing trend towards more Stock Pice % Change $\geq 1.8\%$ (up_2) observations, we posit that this is normal for any stock that is increasing in value over a long period of time, because successive time points of observation for a company that is increasing in value or market share would indicate higher and higher stock prices, though they are still subject to short-term macro level market

corrections. Additionally, knowing that price is directly influenced by the quantity of shares transacted it makes sense that technical indicators would also tend to be left skewed over long terms. As we can see from the histograms showing that RSI, and Relative VWAP, are trending towards higher values which are considered as price trend increase signals as they trend towards higher values as well. Regarding Open-Close Price Percent Change, the data appear to be normally distributed which is further supported by an understanding healthy market corrections that suggest systematic and predictable stock price percent correction and improvements based on cyclical trends and market participant behavior.

The logistic regression model lends itself well to this data regardless, as the primarily relevant distribution is that of the x_i to Y in which a binomial distribution is present. Therefore making logistic more robust and broader in scope (Agretsi, 2013).

IV. Results

A. Selection and GOF

We were unable to find a statistically significant logistic regression model. We began with the cumulative logit using an ordinal response with proportional odds. The saturated model we began the backwards selection with had 7 covariates. Through roughly 20 steps checking for interactions and main effects, our selection yielded a reduced model with 2 covariates, the previous Relative Volume Signal and previous Percent Change (p_rel_vol_signal and prev_OC_per_change). This model was statistically improved from our previous stepwise model via a reduced AIC value, and a likelihood ratio test. Additionally, this model, with Residual Deviance = 585.8, df = 785 compared to the null model of with no explanatory variables with Residual Deviance = 598.48 on 788 degrees of freedom shows an improved fit.

However, the overall significance values for the model are far from ideal to determine an overall GOF. To confirm this we checked the AUC and found a value of 0.59 indicating that our model was moderately better than random chance, but we also found a $pseudo - R^2 \approx -2.1\%$ indicating that our effects very weakly capture the data in whole. Plotting the model with predicted values did not help yield clarity on significance as roughly half the vales at set points of comparison had absolute differences matching corresponding plots for different levels of response and relative volume signal level, while the other half did not, as in Figure 1.

We had difficulty checking the proportional odds assumption through several different methods in vglm(), polr(), Brant's test, entropy and a few other alternative methods. The Brandt test via brandt() in R gave us a moderate chance of non-proportional odds and we ended up relying on the plots of predicted probabilities to judge parallel slope effects, that show there may be a moderate effect of non-proportionality at a varying level of volume signal, but we are not overly confident about this.

Table 4: Brant Test for Proportional Odds Assumption

Test for	X2	df	probability
Omnibus	19.11	9	0.02
prev_OC_per_change	3.07	3	0.38
p_rel_vol_signalflat	3.6	3	0.31
p_rel_vol_signalup	0.53	3	0.91

H0: Parallel Regression Assumption holds

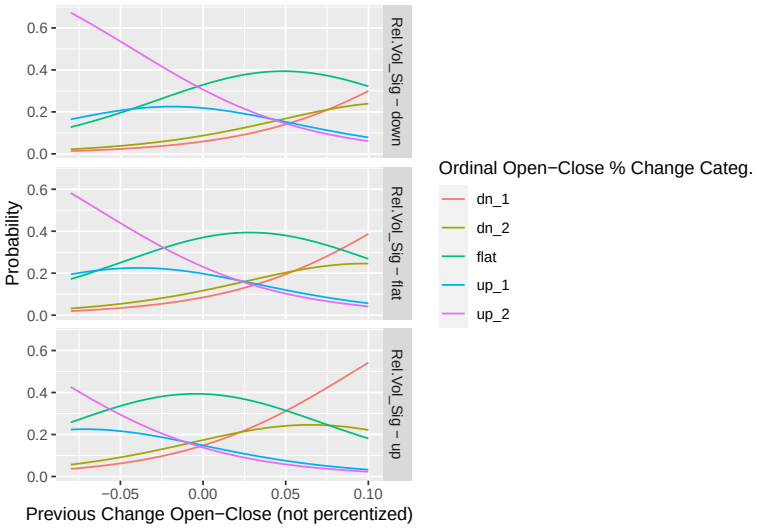

Figure 1: Comparison of Predicted Values at Different Levels

Due to lack of clarity and findings from plots we fit the same proportional odds model under the partial proportional odds assumption. The results were slightly improved based on AIC criterion and LRT statistics however standard errors increased overall for parameters of interest (Table 6). And the partial proportional model has significantly more parameters violating considerations of parsimony. We additionally considered a baseline reference multinomial logit and assessed goodness of fit. It has improved the total model deviance (Table 9), but increased standard errors for nearly all parameters and comparison groups (Table 7).

The confidence intervals for these baseline comparisons for all Open-Close % change parameters were extremely large causing difficulty in realistic interpretation beyond signficance. Even under this model with the most improved LRT, and AIC, the $pseudo - R^2 \approx -5.1\%$, which is twice as large as we found in the original model (≈ 2.1, but still indicates a very poor lack of fit. All model output and LRT significance tests are provided below.

Table 5: Proportional Odds Cummulative Logit Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-2.76890	0.44140	-6.27304	0.00000
(Intercept):2	-1.76461	0.40581	-4.34832	0.00001
(Intercept):3	-0.09932	0.38456	-0.25827	0.79620
(Intercept):4	0.81710	0.39119	2.08876	0.03673
p_rel_vol_signalflat	0.38921	0.41017	0.94888	0.34268
p_rel_vol_signalup	1.01707	0.46640	2.18067	0.02921
prev_OC_per_change	19.20793	5.97618	3.21408	0.00131

Table 6: Partial Proportional Odds Cummulative Logit Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-17.46357	675.85796	-0.02584	0.97939
(Intercept):2	-1.76482	0.55842	-3.16039	0.00158
(Intercept):3	0.12604	0.42770	0.29469	0.76823
(Intercept):4	0.70146	0.46129	1.52065	0.12835
p_rel_vol_signalflat:1	15.08841	675.85803	0.02232	0.98219
p_rel_vol_signalflat:2	0.25338	0.59871	0.42321	0.67214
p_rel_vol_signalflat:3	0.11259	0.46311	0.24313	0.80791
p_rel_vol_signalflat:4	0.78280	0.51715	1.51367	0.13011
p_rel_vol_signalup:1	15.96364	675.85806	0.02362	0.98116
p_rel_vol_signalup:2	1.25257	0.63669	1.96733	0.04915
p_rel_vol_signalup:3	0.75317	0.53632	1.40435	0.16021
p_rel_vol_signalup:4	0.48235	0.57471	0.83930	0.40130
prev_OC_per_change	19.07304	6.07365	3.14030	0.00169

Table 7: Baseline Logit Odds Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-15.97546	583.19351	-0.02739	0.97815
(Intercept):2	-0.81916	0.64311	-1.27375	0.20275
(Intercept):3	0.08692	0.51231	0.16966	0.86527
(Intercept):4	-1.05830	0.70278	-1.50588	0.13210
p_rel_vol_signalflat:1	15.25518	583.19362	0.02616	0.97913
p_rel_vol_signalflat:2	0.20047	0.72803	0.27535	0.78305
p_rel_vol_signalflat:3	0.65578	0.57361	1.14325	0.25294
p_rel_vol_signalflat:4	1.30552	0.75363	1.73229	0.08322
p_rel_vol_signalup:1	15.86171	583.19371	0.02720	0.97830
p_rel_vol_signalup:2	0.71931	0.80336	0.89537	0.37059
p_rel_vol_signalup:3	0.38873	0.66807	0.58186	0.56066

	Estimate	Std. Error	z value	Pr(>|z|)
p_rel_vol_signalup:4	-0.13183	0.97274	-0.13553	0.89220
prev_OC_per_change:1	32.67930	12.42244	2.63067	0.00852
prev_OC_per_change:2	30.03054	11.97118	2.50857	0.01212
prev_OC_per_change:3	21.35664	9.66823	2.20895	0.02718
prev_OC_per_change:4	25.82085	12.20564	2.11549	0.03439

Table 8: H_0: Proportional Odds Compared, vs. Partial Proportional Odds

Likelihood ratio test

Model 1: f_Or_close_change_v_prev ~p_rel_vol_signal + prev_OC_per_change
Model 2: f_Or_close_change_v_prev ~p_rel_vol_signal + prev_OC_per_change

#Df LogLik Df Chisq Pr(>Chisq)
1 785 -293
2 779 -285 -6 16.5 0.011 *
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 9: H_0: Proportional Odds Compared, vs. Baseline Logit Odds

Likelihood ratio test

Model 1: f_Or_close_change_v_prev ~p_rel_vol_signal + prev_OC_per_change
Model 2: f_close_change_v_prev ~p_rel_vol_signal + prev_OC_per_change

#Df LogLik Df Chisq Pr(>Chisq)
1 785 -293
2 776 -284 -9 17.6 0.04 *
—
Signif. codes 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

B. Interpretation

Despite all of our models have significant lack of fit, our ordinal model with proportional odds has the least number of parameters, and a minor reduction in comparative signficance overall, so interpretation value for the parameter's estimates to give us a sense of the summary of effects at extremes. The model indicates a $\theta_{Previous\ \%SP^{W}_{\triangle}} = 1798.7$, suggesting for one unit increase in the previous week's stock price % change, the chance of response nearer the outcome category Y = 1 is at least 1798 times higher than for a one unit decrease, based on the low C.I. estimate (Table 9). Additionally, the $\theta_{Vol^{W}_{Rel}} = 1.1$, suggesting that the response being nearer the outcome category Y = 1 is at least 1.1 times greater for a relative volume "up" signal versus that of a "down" signal (Table 9).

Practically, this could be interpreted that a very large increase (positive) in the previous week's % change price would reduce the chance of the following week having a large increase in % change price compared to a large previous week's % chance, all else held constant. And, the chance of having a large increase (positive) in % change price for the following week, would be less for the up signal group versus down signal group of the prior week.

Even more simply, it would appear that if during the previous week the stock price had a large % increase there's a stronger chance the following week will be down versus the previous week seeing a large decrease in % change. For the relative volume signal, if the previous week gave an up versus a down signal, the following week will have a reduced chance of a large increase in % price change. Though we should keep in mind this is a general assumption that needs be taken with some caution due to significant lack of fit.

Table 10: Proportional Odds Cummulative Logit Model - Wald Confidence Intervals

Model 1: Proportional Odds Cummulative Logit Model			
Predictors	Odds Ratios	CI (2.5% - 97.5%)	p
(Intercept) * 1	0.06	0.03 – 0.15	<0.001
(Intercept) * 2	0.17	0.08 – 0.38	<0.001
(Intercept) * 3	0.91	0.43 – 1.92	0.796
(Intercept) * 4	2.26	1.05 – 4.87	0.037
p_rel_vol_signal [flat]	1.48	0.66 – 3.30	0.343
p_rel_vol_signal [up]	2.77	1.11 – 6.90	0.029
prev_OC_per_change	219733432.07	1798.72 – 26842855096314.90	0.001
Observations	198		

The estimated probabilities for the predicted model are listed below.

Table 11: $\hat{P}(Y = j)$ at Fixed Levels - Low Previous Volume Signal (down) & Min. Previous % Open-Close $\triangle$

	$P(Y = j)$				
j	1	2	3	4	5
Ordinal Category	dn_1	dn_2	$flat$	up_1	up_2
	0.0125	0.0209	0.1213	0.1593	0.6858

Table 12: $\hat{P}(Y = j)$ at Fixed Proportions - Low Previous Volume Signal (down) & Max. Previous % Open-Close $\triangle$

	$P(Y = j)$				
j	1	2	3	4	5
Ordinal Category	dn_1	dn_2	$flat$	up_1	up_2
	0.3317	0.2437	0.3021	0.0696	0.0529

Table 13: $\hat{P}(Y = j)$ at Fixed Proportions - High Previous Volume Signal (down) & Min. Previous % Open-Close $\triangle$

	$P(Y = j)$				
j	1	2	3	4	5
Ordinal Category	dn_1	dn_2	$flat$	up_1	up_2
	0.0339	0.0535	0.2488	0.2226	0.4412

Table 14: $\hat{P}(Y = j)$ at Fixed Proportions - High Previous Volume Signal (down) & Min. Previous % Open-Close $\triangle$

	$P(Y = j)$				
j	1	2	3	4	5
Ordinal Category	dn_1	dn_2	$flat$	up_1	up_2
	0.5786	0.2108	0.1626	0.0283	0.0198

V. Conclusion and Discussion

Despite not being able to find an overall model of significance, we improved the understanding of significant technical indicators in prediction of % stock price change. Out of our 7 initial considerations, we determined 2 of these to most relevant when considered at their extremes, Previous Day Relative Volume Signal, and Previous Open-Close % change. The outcomes of these input variables had an inverse relationship with their input parameters. Which makes sense in an market oscillation perspective–short term gains tend to correct themselves, at least temporarily.

We believe it may be worth changing the thresholds for our ordinal variables and retesting the model for improved significance. Additionally, a shorter period of time, or a longer period of time may improve the predictability as one week may be just long enough to caputure the extremes of upward and downward volatility. It may also we worth testing a linear model to determine if continuous outputs would be more helpful–it may be that our thresholds, while helpful for take-action signals, may have cut points that don't adequately capture the spectrum. Alternatively, clustering could potentially be explored. It would also be helpful to look at a larger period of the data sample (~10 years), with a consideration for time series due to natural market cycles. We hope this research serves as a good stepping off point for building a more comprehensive model using advanced algorithmic procedures.

VI. Refrences

[1]: barchart. (2021, May 1). Barchart.com | Commodity, Stock, and Currency Quotes, Charts, News & Analysis. Barchart.Com. https://www.barchart.com.

[2]. FlowAlgo LLC. (2021, May 1). FlowAlgo :: Realtime Option Flow, Unusual Option Activity, Darkpool Flows. FlowAlgo. https://FlowAlgo.com

[3]. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[4]. Thomas W. Yee (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

[5]. Thomas W. Yee and C. J. Wild (1996). Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B, 58(3), 481-493.

[6]. An Introduction to Categorical Data Analysis (2018). , 3rd Edition (Wiley Series in Probability and Statistics) (3rd ed.). Wiley.

[7]. Hosmer, D., Lemeshow, S., & Strudivant, R. (2013, April 1). Applied Logistic Regression, 3rd Edition.

[8]. Sasani, A., & Tibado, S. (2020). Stock Market Prediction using Hidden Markov Model and Neural Network. International Journal of Engineering and Applied Sciences (IJEAS), 7(4). https://doi.org/10.31873/ijeas.7.04.07

[9]. Contributors to Wikimedia projects. (2021, March 8). LaTeX/Floats, Figures and Captions. Wikibooks, Open Books for an Open World. https://en.wikibooks.org/wiki/LaTeX/Floats,_Figures_and_Captions

[10]. Bitvai, Z., & Cohn, T. (2014). Day trading profit maximization with multi-task learning and technical analysis. Machine Learning, 101(1–3), 187–209. https://doi.org/10.1007/s10994-014-5480-x

[11] Yiu, T. (2020, March 29). Are Stock Returns Normally Distributed? - Towards Data Science. Medium. https://towardsdatascience.com/are-stock-returns-normally-distributed-e0388d71267e

V. Appendix

A. Reference on Indicators of Interest

1. Relative Volume

Volume for specified period compared to an average for n number of specified periods. Higher Relative Volumes indicate strong price movement (up or down) and lower relative volumes indicate weak price movement (consolidation).

2. Relative Strength Index (RSI)

A technical indicator taking continuous values for a specified period that acts as an Overbought or Oversold indicator, however it is extremely rare for this indicator to give values outside the range of $[20, 90]$. Typically accepted values for Overbought are RSI ≥ 70, and Oversold RSI ≤ 30, Consolidation and Lateral price are seen for value inside those thresholds. One would Sell when RSI indicates Overbought, and Buy when RSI indicates Oversold.

3. Average True Range (ATR)

A measure of volatility and expected movement based on the average range of High Price minus Low Price for a specified period. Typically, as ATR trends towards zero one would expect a larger amount of volatility in the near term. And as ATR trends towards higher values, away from zero, one would expect less volatility in the near term.

4. Volume Weighted Average Price (VWAP)

A volume weighted average price that is used as an a indicator of the price comparative to volume of shares traded. Typically when a stock is below VWAP it would indicate a good buying opportunity, and when it is above VWAP it would indicate a less opportunistic buy opportunity but may indicate strong upward price direction in near term. It's well assumed that most Hedge Funds and High Frequency Trading Algorithims (HFTs) use a price crossing from below VWAP to above VWAP as a buy signal

B. Appendix: R-code

```
# load packages

library(knitr)
library(epiDisplay)
library(psych)
library(formatR)
library(stargazer)
library(xtable)
library(ggplot2)
library(emmeans)
library(rquery)
library(readxl)
library(nlme)
library(MASS)
library(Hmisc)
library(vcd)
library(scatterplot3d)
library(Hmisc)
library(rgl)
library(Matrix)
library(dplyr)
library(car)
library(moments)
library(PerformanceAnalytics)
library(olsrr)
library(tidyverse)
library(caret)
library(magick)
library(sjPlot)
library(sjmisc)
library(sjlabelled)
library(binom)
library(VGAM)
library(stats)




spy_data <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Documents/Tyler/KUM
    header = TRUE)


knitr::opts_chunk$set(echo = TRUE)
```

```
options(digits = 5, width = 60, xtable.comment = FALSE)
opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
out_type <- knitr::opts_knit$get("rmarkdown.pandoc.to")
p_rel_vol_signal <- relevel(factor(spy_data$p_rel_vol_signal),
    ref = "down")
# levels(p_rel_vol_signal)
f_Or_close_change_v_prev <- ordered(spy_data$close_change_v_prev)
# levels(f_Or_close_change_v_prev)
fit_spy_11 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
fit_spy_0 <- vglm(f_Or_close_change_v_prev ~ 1, family = cumulative(parallel = TRUE),
    data = spy_data)

# Proportional compared to no interactions summary(fit_spy_0)
# summary(fit_spy_11) (psuedoRSquared <- 1 -
# (598.4781/585.8445))
library(VGAM)
# lrtest_vglm(fit_spy_0, fit_spy_11) extractAIC(fit_spy_0)
# extractAIC(fit_spy_11)

# Proportional compared to no interactions summary(fit_spy_0)
# summary(fit_spy_11) (psuedoRSquared <- 1 -
# (598.4781/585.8445)) lrtest_vglm(fit_spy_0, fit_spy_11)
# extractAIC(fit_spy_0) extractAIC(fit_spy_11)

# Check validity of model checking the validity of the model
# in total using the ROC and AUC #### Test for overall
# goodness of fit
library(pROC)
require(NHANES)
rocplot1 <- multiclass.roc(spy_data$close_change_v_prev ~ fitted(fit_spy_11),
    data = spy_data)
# rocplot1 auc(rocplot1) # AUC has a minimum of 0.50 which
# indicates that model has no better prediciton then simple
# guessing

##### Checking Proportional Odds prop_test_fit_spy_11 <-
##### polr(factor(close_change_v_prev) ~ prev_OC_per_change +
##### p_rel_vol_signal, data=spy_data, method = 'logistic', Hess
##### = TRUE) library(MASS) library(brant)
##### kable(brant(prop_test_fit_spy_11), caption = 'Brant Test
##### for Proportional Odds Assump.')
```

```
# See if variance in relative volume is problematic Calcuate
# the entropy.  Convert counts to relative frequences (pi_i)
# and sum then divide by pi_i*log(pi)
relvolsig_freq <- ftable(spy_data$p_rel_vol_signal)/198
ss_sq <- sum(relvolsig_freq[1]^2 + relvolsig_freq[2]^2 + relvolsig_freq[3]^2)
# This is equivalent to R value so

entro <- sum((relvolsig_freq[1] * log(relvolsig_freq[1]^-1)) +
    relvolsig_freq[2] * log(relvolsig_freq[2]^-1) + relvolsig_freq[3] *
    log(relvolsig_freq[3]^-1))
# Says that ins't entropy isn't bad, b/c it's near maximum
# level at log(4) indicating that variance is minimal. Woudl
# be minimal at 0 if variance is maximal


# Will do a proportional odds assessment using both the
# continous and categorical variable
# https://rpubs.com/shreejit16/625642
library(rms)
arth.po2 <- lrm(close_change_v_prev ~ p_rel_vol_signal + prev_OC_per_change,
    data = spy_data)

### Plotting the predicted values with fitted response Visually
### display predicted probabilities ####
### https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/
### min(spy_data$prev_OC_per_change)
### max(spy_data$prev_OC_per_change)

down_data <- data.frame(p_rel_vol_signal = rep("down", each = 80),
    prev_OC_per_change = rep(seq(from = -0.08, to = 0.1, length.out = 80)))

flat_data <- data.frame(p_rel_vol_signal = rep("flat", each = 80),
    prev_OC_per_change = rep(seq(from = -0.08, to = 0.1, length.out = 80)))

up_data <- data.frame(p_rel_vol_signal = rep("up", each = 80),
    prev_OC_per_change = rep(seq(from = -0.08, to = 0.1, length.out = 80)))

new_signal_vec <- rbind(down_data, flat_data, up_data)
# nrow(new_signal_vec)

newdata_spy2 <- cbind(new_signal_vec, predictvglm(fit_spy_11,
    newdata = new_signal_vec, type = ("response")))

## Create data frame for the use in ggplot
```

```
library(reshape)
library(reshape2)
lnewdat2 <- melt(newdata_spy2, id.vars = c("p_rel_vol_signal",
    "prev_OC_per_change"), variable.name = "close_change_v_prev",
    value.name = "Probability", variable.factor = FALSE)
# head(lnewdat2)

# Plot for probablites with all 3 volume signals as facet
appender <- function(string, suffix = "Rel.Vol_Sig - ") paste0(suffix,
    string)
ggplot(lnewdat2, aes(x = prev_OC_per_change, y = Probability,
    colour = close_change_v_prev)) + geom_line() + facet_grid(p_rel_vol_signal ~
    ., labeller = as_labeller(appender, multi_line = TRUE)) +
    labs(y = "Probability", x = "Previous Change Open-Close (not percentized)",
        col = "Ordinal Open-Close % Change Categ.")


# p_rel_vol_signal <-
# relevel(factor(spy_data$p_rel_vol_signal), ref = 'down')
# levels(p_rel_vol_signal)
f_close_change_v_prev <- factor(spy_data$close_change_v_prev)
# levels(f_Or_close_change_v_prev) Cummulative odds model
fit_spy_11 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)

# Partial Proportional Odds model
fit_spy_22 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = FALSE ~
    p_rel_vol_signal), data = spy_data)

# Baseline Refrence Model - Log odds Multinomial
fit_spy_23 <- vglm(f_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = "multinomial", data = spy_data)

ctable11 <- coef(summary(fit_spy_11))
kable(ctable11, caption = "Proportional Odds Cummulative Logit Model")
ctable22 <- coef(summary(fit_spy_22))
kable(ctable22, caption = "Partial Proportional Odds Cummulative Logit Model")
ctable23 <- coef(summary(fit_spy_23))
kable(ctable23, caption = "Baseline Logit Odds Model")
# Confidence intervals for prop odds structure, copied these
# into a table. sjPlot::tab_model(fit_spy_11) Must reassure
# that the levels don't change b/c multinomial requres
```

```r
# objects to be listed as factors
p_rel_vol_signal <- relevel(p_rel_vol_signal, ref = "down")
f_close_change_v_prev <- factor(spy_data$close_change_v_prev)

# Interpreting the effects of the model Predict max open to
# close to min open to close at min rel volume signal - For
# outcome Y=5
pred_minOC_rldn1 <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "down",
    prev_OC_per_change = min(spy_data$prev_OC_per_change)), type = "response")
pred_maxOC_rldn <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "down",
    prev_OC_per_change = max(spy_data$prev_OC_per_change)), type = "response")

pred_minOC_rlup <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "up",
    prev_OC_per_change = min(spy_data$prev_OC_per_change)), type = "response")
pred_maxOC_rlup <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "up",
    prev_OC_per_change = max(spy_data$prev_OC_per_change)), type = "response")

# citation() citation(package = 'VGAM')

options(width = 60, xtable.comment = FALSE)


########## CODE USED FOR MODEL BUILDING, EXPLORATION AND ANALYSIS -
########## NOT NECCESSARY IN REPORT Check for Cut Point Adequcy
########## Quantiles

quantile(spy_data$OC_per_change)
sort(spy_data$OC_per_change)
nrow(spy_data)

library(mmand)
threshold(spy_data$OC_per_change, level, method = c("kmeans"),
    binarise = TRUE)
k2 <- kmeans(spy_data$OC_per_change, centers = 5, nstart = 25)
k2
oc_mat <- data_frame(spy_data$OC_per_change)


# Brant test
# https://stats.stackexchange.com/questions/58772/brant-test-in-r
# THIS WORKS WITH polr() function. You want all pvlaues to be
# greater than 0.05 to be parraell slope assumption
library(brant)
brant(model, by.var = F)
```

```r
print(fit_spy_11)


#### USED THIS TO SET THE LEVELS 5.11.21
p_rel_vol_signal <- relevel(p_rel_vol_signal, ref = "down")
levels(p_rel_vol_signal)
f_Or_close_change_v_prev <- ordered(spy_data$close_change_v_prev)
levels(f_Or_close_change_v_prev)

# Checking multicolinearity of vars
sus_vars_matrix <- as.matrix(cbind(spy_data$prev_RSI_15day, spy_data$p_rel_vol_wk_3mo,
    spy_data$prev_OC_per_change, spy_data$prev_rel_ATR_15day_avg,
    spy_data$prev_rel_low_VWAP_week, spy_data$prev_dp_count,
    spy_data$prev_dp_volume_100K))
cor(sus_vars_matrix)


fit_spy_11_polr <- polr(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, method = "logistic", data = spy_data,
    Hess = TRUE)
summary(fit_spy_11_polr)
summary(fit_spy_11)

library(gtsummary)


## Use backwards selection model because of high
## multicolinearity potential due to technical indicators
## being calculated from similar underlying factors


# Will do a proportional odds assessment using both the
# continous and categorical variable
# https://rpubs.com/shreejit16/625642
library(rms)
arth.po2 <- lrm(close_change_v_prev ~ p_rel_vol_signal + prev_OC_per_change,
    data = spy_data)
arth.po2

# See if variance in relative volume is problematic Calcuate
# the entropy.  Convert counts to relative frequences (pi_i)
# and sum then divide by pi_i*log(pi)
relvolsig_freq <- ftable(spy_data$p_rel_vol_signal)/198
ss_sq <- sum(relvolsig_freq[1]^2 + relvolsig_freq[2]^2 + relvolsig_freq[3]^2)
```

```
ss_sq  # This is equivalent to R value so

entro <- sum((relvolsig_freq[1] * log(relvolsig_freq[1]^-1)) +
    relvolsig_freq[2] * log(relvolsig_freq[2]^-1) + relvolsig_freq[3] *
    log(relvolsig_freq[3]^-1))
entro  # Says that ins't entropy isn't bad, b/c it's near maximum level at log(4) indi

# sjPlot::tab_model(fit_spy_11,show.r2 = TRUE, show.se =
# TRUE, show.aic = TRUE, show.dev = TRUE, show.ci = FALSE,
# show.stat = TRUE,transform = NULL, show.obs = TRUE,
# df.method = 'wald', string.est = 'Estimate', title =
# 'Proportional Odds Model', string.p = 'Wald P-Value')


# First check the main effects for signficance for inclusion
# in potenital main effects model. Want to see P-vals < 0.20
# for inclusion or other criteria that would indicate
# suspicion of them being important

#####################################################################
library(VGAM)
# Model with no parameters
fit_spy_0 <- vglm(f_Or_close_change_v_prev ~ 1, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_0)

# Models with 1 parameter RSI param
fit_spy_1 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_1)
# signficant at 0.065

lrtest_vglm(fit_spy_0, fit_spy_1)  #not very signficant 0.08
# Borderline signficant so test with other tests - b/c wald
# test says 0.065 from summary and LRT says 0.08
extractAIC(fit_spy_0)
extractAIC(fit_spy_1)
### AIC is reduced here, we probably want to include it

## Relative Volume paramater
fit_spy_2 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_2)
# semi - ignficant at 0.1
```

```r
extractAIC(fit_spy_0)
extractAIC(fit_spy_2)

lrtest_vglm(fit_spy_0, fit_spy_2)  #not signficant 0.22 -- likely due to factor referen

## Relative Volume paramater -- as a numeric value instead of
## category. Is not more better than categorical var.
fit_spy_2_n <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_wk_3mo,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_2_n)
# not signficant at 0.516


lrtest_vglm(fit_spy_0, fit_spy_2_n)  #not signficant 0.51 -- likely due to factor refer

## Open to close change
fit_spy_3 <- vglm(f_Or_close_change_v_prev ~ prev_OC_per_change,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_3)
# signficant at 0.0054

lrtest_vglm(fit_spy_0, fit_spy_3)  #signficant 0.0077

extractAIC(fit_spy_0)
extractAIC(fit_spy_3)

## Relative Average True Range
fit_spy_4 <- vglm(f_Or_close_change_v_prev ~ prev_rel_ATR_15day_avg,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_4)
# NOT SIGNIFICANT at 0.77

lrtest_vglm(fit_spy_0, fit_spy_4)  # not significant 0.72

extractAIC(fit_spy_0)
extractAIC(fit_spy_4)

## Relative VWAP-week from low price
fit_spy_5 <- vglm(f_Or_close_change_v_prev ~ prev_rel_low_VWAP_week,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_5)
# NOT SIGNIFICANT at 0.82

lrtest_vglm(fit_spy_0, fit_spy_5)  # not significant 0.83
```

```r
## Dark Pool order Count
fit_spy_6 <- vglm(f_Or_close_change_v_prev ~ prev_dp_count, family = cumulative(parallel
    data = spy_data)
summary(fit_spy_6)
# NOT SIGNIFICANT at 0.868

lrtest_vglm(fit_spy_0, fit_spy_6)  # not significant 0.88

## Dark Pool share volume
fit_spy_7 <- vglm(f_Or_close_change_v_prev ~ prev_dp_volume_100K,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_7)
# NOT SIGNIFICANT at 0.978

lrtest_vglm(fit_spy_0, fit_spy_7)  # not significant 0.98


########################################################################## Full model without i
########################################################################## are known important
########################################################################## of being relevant as
########################################################################## all significant main
fit_spy_8 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day +
    p_rel_vol_signal + prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_8)

## Open Close percent change removed
fit_spy_9 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day +
    p_rel_vol_signal, family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_9)

### Likelikehood ratio test
lrtest_vglm(fit_spy_8, fit_spy_9)  #Significant at 0.022 -> significant difference exis

extractAIC(fit_spy_8)
extractAIC(fit_spy_9)

## Relative volume signal removed
fit_spy_10 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_10)

### Likelikehood ratio test
```

```r
lrtest_vglm(fit_spy_8, fit_spy_10)  #Significant at 0.03 -> significant difference exis

extractAIC(fit_spy_8)
extractAIC(fit_spy_10)

## Relative Strength Index removed
fit_spy_11 <- vglm(f_0r_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_11)

### Likelikehood ratio test
lrtest_vglm(fit_spy_8, fit_spy_11)  #Not significant at 0.17 --> reduced model may be b

#### Further Check with AIC
extractAIC(fit_spy_8)
extractAIC(fit_spy_11)
##### AIC is slight reduced in the reduced model compared to the
##### fuller model, additionally the degrees of freedom is
##### reduced by 1 which is signficant consdiering we're trending
##### towards a model with 2-3 degrees of freedom in total. This
##### allows us to effectively reduce the number of paramters in
##### the model by 25-30%, moving towards a more parsimonous
##### model.

########################################################################## Add in any vars that
########################################################################## contribution in the
fit_spy_12 <- vglm(f_0r_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change + prev_rel_low_VWAP_week, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_12)

lrtest_vglm(fit_spy_11, fit_spy_12)  # not significant - do not include this var

# Add dark pool order count in
fit_spy_13 <- vglm(f_0r_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change + prev_dp_count, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_13)

lrtest_vglm(fit_spy_11, fit_spy_13)  # not significant - do not include this var

# Add dark pool volume in
fit_spy_14 <- vglm(f_0r_close_change_v_prev ~ p_rel_vol_signal +
```

```
    prev_OC_per_change + prev_dp_volume_100K, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_14)

lrtest_vglm(fit_spy_11, fit_spy_14)  # not significant - do not include this var

######################################################################## Continue backwards e
######################################################################## reduced model earlie
######################################################################## fullest model so far
fit_spy_11 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_11)

## Open to close percent change removed
fit_spy_15 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_15)

lrtest_vglm(fit_spy_11, fit_spy_15)  ##Significant at 0.0019 -> significant difference

## Relative Volume signal removed
fit_spy_16 <- vglm(f_Or_close_change_v_prev ~ prev_OC_per_change,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_16)

lrtest_vglm(fit_spy_11, fit_spy_16)  # not very significant, not advised to remove ter
extractAIC(fit_spy_11)
extractAIC(fit_spy_16)

## Empty model - all covariates removed
fit_spy_17 <- vglm(f_Or_close_change_v_prev ~ 1, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_17)
smod17 <- summary(fit_spy_17)

# Refrence model becomes one where RSI was removed
lrtest_vglm(fit_spy_11, fit_spy_17)  # significant at 0.0055 not advised have a model u

######################################################################## Check for plausible
######################################################################## far Refrence model
fit_spy_11 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
```

```r
summary(fit_spy_11)

# Model with interactions from covariates in refrence model
fit_spy_18 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change + p_rel_vol_signal * prev_OC_per_change,
    family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_18)

lrtest_vglm(fit_spy_11, fit_spy_18)  # Not significant at 0.2, therefore reduced model
extractAIC(fit_spy_11)
extractAIC(fit_spy_18)

# Suspected interaction of RSI with covariates based on 1
# paratmer model signficance from earlier. Must include RSI
# as a main effect to include in interactions
fit_spy_19 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day +
    p_rel_vol_signal + prev_OC_per_change + p_rel_vol_signal *
    prev_RSI_15day + prev_OC_per_change * prev_RSI_15day, family = cumulative(parallel =
    data = spy_data)
summary(fit_spy_19)

lrtest_vglm(fit_spy_11, fit_spy_19)  #not signficant at 0.14, try reduced interactions
extractAIC(fit_spy_11)
extractAIC(fit_spy_19)

## Reduced interactions - no Open Close Change * RSI
## interaction
fit_spy_20 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day +
    p_rel_vol_signal + prev_OC_per_change + p_rel_vol_signal *
    prev_RSI_15day, family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_20)

lrtest_vglm(fit_spy_19, fit_spy_20)  # not signficant at 0.12, try other reduced intera
extractAIC(fit_spy_19)
extractAIC(fit_spy_20)

lrtest_vglm(fit_spy_11, fit_spy_20)  #Not signficant, reduced model 11 better

## Reduced interactions - no Relative volume * RSI interaction
fit_spy_21 <- vglm(f_Or_close_change_v_prev ~ prev_RSI_15day +
    p_rel_vol_signal + prev_OC_per_change + prev_OC_per_change *
    prev_RSI_15day, family = cumulative(parallel = TRUE), data = spy_data)
summary(fit_spy_21)
```

```r
lrtest_vglm(fit_spy_19, fit_spy_21)  # not signficant at 0.12, however AIC improved wi
extractAIC(fit_spy_19)
extractAIC(fit_spy_21)

# Non-interaction model compared to most signficant reduced
# interaction model
lrtest_vglm(fit_spy_11, fit_spy_21)  # not signficant at 0.21, and AIC improved with le
extractAIC(fit_spy_11)
extractAIC(fit_spy_21)

########################################################################## Final Model
fit_spy_11 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_11)

# Check the fit of model with chi-sq statistic of deviance
1 - pchisq(585.84, 785)

########################################################################## Check validity of mo
########################################################################## in total using the R
########################################################################## goodness of fit
library(pROC)
require(NHANES)
rocplot1 <- multiclass.roc(spy_data$close_change_v_prev ~ fitted(fit_spy_11),
    data = spy_data)
rocplot1
auc(rocplot1)  # AUC has a minimum of 0.50 which indicates that model has no better pr

# Calcuate ROC curve by hand

fitted_fit_spy_11_vals <- fittedvlm(fit_spy_11)

library(ggplot2)  ## For plotting
library(caret)  ## For model fitting and evaluation
library(visreg)  ## For visualizing regression models
library(plotROC)  ## For constructing ROC curves
library(mgcv)  ## For fitting GAM models
library(kernlab)  ## Contains an example dataset
library(glmnet)
df1 <- data.frame(Status = spy_data$close_change_v_prev, Prob = fitted_fit_spy_11_vals)
df1

# plot.roc(rocplot1, legacy.axes = TRUE) #Doesnt work via
```

```r
# categorical response
auc(rocplot1)  # AUC has a minimum of 0.50 which indicates that model has no better pr

#### Get predicted vglm values ####
library(VGAM)
predictvglm(fit_spy_11, type = ("response"))

# plot(f_Or_close_change_v_prev ~ p_rel_vol_signal +
# prev_OC_per_change, data = spy_data, center = TRUE, main =
# 'smart prediction') lines(predictvglm(fit_spy_11, data =
# spy_data, type = 'response'))
# points(predictvglm(fit_spy_11, spy_data, type =
# 'response'), type = 'b', col = 2)


#### Visually display predicted probabilities ####
#### https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

max(spy_data$prev_OC_per_change)
min(spy_data$prev_OC_per_change)
mean(spy_data$prev_OC_per_change)
quantile(spy_data$prev_OC_per_change)
count(spy_data, prev_OC_per_change)

##### Using new generated vectors that are random######

### Doesn't work for random numbers #####
myFun_char <- function(n = 5000) {
    a <- do.call(paste0, replicate(1, sample(c("down", "flat",
        "up"), n, TRUE), FALSE))
    paste0(a)
}
new_signal_vec <- myFun_char(240)

new_OC_change_vec <- runif(240, -0.07, 0.09)
new_OC_change_vec

new_data_vec <- rbind(new_signal_vec, new_OC_change_vec)
new_data_vec_df <- data.frame(new_data_vec)
colnames(new_data_vec_df) <- c("prev_OC_per_change", "p_rel_vol_signal")

###### ################################### ###### Used to create
###### data vectors for predicted probablities
min(spy_data$prev_OC_per_change)
```

```r
max(spy_data$prev_OC_per_change)

down_data <- data.frame(p_rel_vol_signal = rep("down", each = 80),
    prev_OC_per_change = rep(seq(from = -0.08, to = 0.1, length.out = 80)))

flat_data <- data.frame(p_rel_vol_signal = rep("flat", each = 80),
    prev_OC_per_change = rep(seq(from = -0.08, to = 0.1, length.out = 80)))

up_data <- data.frame(p_rel_vol_signal = rep("up", each = 80),
    prev_OC_per_change = rep(seq(from = -0.08, to = 0.1, length.out = 80)))

new_signal_vec <- rbind(down_data, flat_data, up_data)
nrow(new_signal_vec)

predictvglm(fit_spy_11, new_signal_vec, type = ("response"))

newdata_spy2 <- cbind(new_signal_vec, predictvglm(fit_spy_11,
    new_signal_vec, type = ("response")))
head(newdata_spy2)

newdata_spy2

library(reshape)
library(reshape2)
lnewdat2 <- melt(newdata_spy2, id.vars = c("p_rel_vol_signal",
    "prev_OC_per_change"), variable.name = "close_change_v_prev",
    value.name = "Probability", variable.factor = FALSE)
head(lnewdat2)

# Plot for probablites with all 3 volume signals as facet
appender <- function(string, suffix = "Rel.Vol_Sig - ") paste0(suffix,
    string)
ggplot(lnewdat2, aes(x = prev_OC_per_change, y = Probability,
    colour = close_change_v_prev)) + geom_line() + facet_grid(p_rel_vol_signal ~
    ., labeller = as_labeller(appender, multi_line = TRUE)) +
    labs(y = "Probability", x = "Previous Change Open-Close (not percentized)",
        title = "Predicted cumulative probabilities in the proportional odds model")

###### ##################################### ######

################################## Explored Usefullness of CLASSIFICATION TREES for fun
################################## ############################ Trying out classification
################################## trees since porportional odds model doesn't have grea
################################## Not particularily helpful
```

```
library(rpart)
fit_tree <- rpart(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, method = "class", data = spy_data)
plotcp(fit_tree)
p.fit <- prune(fit_tree, cp = 0.056)
library(rpart.plot)
rpart.plot(p.fit, extra = 1, digits = 4, box.palette = 0)

## https://link.springer.com/article/10.1007/s10260-018-00437-7

# Checking Model Predictions with frequency table
library(PResiduals)
library(sure)
library(VGAM)
library(data.table)
library(Publish)
# Make the OC change a category for each tabulation
# https://publicifsv.sund.ku.dk/~tag/Teaching/share/R-tutorials/Variable-manipulation/
spy_data_DT <- as.data.table(spy_data)
spy_data_DT[, `:=`(prev_OC_cat, cut(prev_OC_per_change, c(-0.1,
    -0.03, -0.015, 0, 0.15, 0.03, 0.1), labels = c("1", "2",
    "3", "4", "5", "6")))]
spy_data_DT$prev_OC_cat

# https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html
# https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf
# Frequency Table
ftable(close_change_v_prev ~ p_rel_vol_signal + prev_OC_cat,
    data = spy_data_DT)
min(spy_data_DT$prev_OC_per_change)
max(spy_data_DT$prev_OC_per_change)

### Can't get fitted values to output the actual levels to make
### a comparison table

##### Cross tab table for occuring values in data set #######

## Calculate compairson table for predicted to actual ####
(pred_vals <- predictvglm(fit_spy_16, spy_data, type = "response"))
levels(close_change_v_prev)[max.col(pred_vals)]
(Predicted_OC_cat <- levels(close_change_v_prev)[max.col(pred_vals)])
f_Predicted_OC_cat <- ordered(Predicted_OC_cat, levels = levels(close_change_v_prev))
class_tab <- xtabs(~close_change_v_prev + f_Predicted_OC_cat,
    data = spy_data)
```

```
class_tab

(CCR <- sum(diag(class_tab))/sum(class_tab))


# Calculate PseudoR2 for model as another signficance test
# https://rstudio-pubs-static.s3.amazonaws.com/747816_ed2c20a650f24802b5646be32807ad64
# psuedoR2 <- 1-fit_spy_11$deviance/fit_spy_11$null.deviance
# Need to put in values manually:
(psuedoR2 <- 1 - 585.84/583.94)

##### Test for Proportial odds assumption ########
prop_odds_fit1 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal,
    family = cumulative(parallel = TRUE), data = spy_data)
prop_odds_fit2 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal,
    family = cumulative(), data = spy_data)

prop_odd_fit3 <- vglm(f_Or_close_change_v_prev ~ prev_OC_per_change,
    family = cumulative(parallel = TRUE), data = spy_data)
prop_odds_fit4 <- vglm(f_Or_close_change_v_prev ~ prev_OC_per_change,
    family = cumulative(parallel = FALSE), data = spy_data)

# A significant p-value associated with this test rejects the
# null hypothesis that the proportional odds assumption
# holds.
# https://conservancy.umn.edu/bitstream/handle/11299/166205/ThomasA_TheProportionalOdd


##### Trying a non proportional odds model since we did not pass
##### the proporitional odds structure test, and the proportional
##### model generally has poor fit ###### Null Model- simplier
##### model
fit_spy_11 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = TRUE),
    data = spy_data)
summary(fit_spy_11)

fit_spy_22 <- vglm(f_Or_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = cumulative(parallel = FALSE ~
    p_rel_vol_signal), data = spy_data)
summary(fit_spy_22)

### Likelikehood ratio test
lrtest_vglm(fit_spy_11, fit_spy_22)  #Signficant difference so partial proportional be
```

```r
extractAIC(fit_spy_11)
extractAIC(fit_spy_22)

# Proportional compared to partial
(psuedoRSquared <- 1 - (585.8445/569.3056))
summary(fit_spy_11)
summary(fit_spy_22)

# Proportional compared to no interactions
summary(fit_spy_0)
summary(fit_spy_11)
(psuedoRSquared <- 1 - (598.4781/585.8445))
lrtest_vglm(fit_spy_0, fit_spy_11)
extractAIC(fit_spy_0)
extractAIC(fit_spy_11)


# Partial proportional compared to no interactions
summary(fit_spy_0)
summary(fit_spy_22)
(psuedoRSquared <- 1 - (598.4781/569.3056))



##### Check the the baseline category logit model for better fit
##### ###### Must reassure that the levels don't change b/c
##### multinomial requres objects to be listed as factors
p_rel_vol_signal <- relevel(p_rel_vol_signal, ref = "down")
levels(p_rel_vol_signal)
f_close_change_v_prev <- factor(spy_data$close_change_v_prev)
levels(f_close_change_v_prev)

fit_spy_23 <- vglm(f_close_change_v_prev ~ p_rel_vol_signal +
    prev_OC_per_change, family = "multinomial", data = spy_data)
summary(fit_spy_23)

### Likelikehood ratio test
lrtest_vglm(fit_spy_23, fit_spy_11)  #Signficant difference so reduced model with propo
extractAIC(fit_spy_11)
extractAIC(fit_spy_23)

# Partial proportional compared to no interactions
summary(fit_spy_11)
summary(fit_spy_23)
```

```
(psuedoRSquared <- 1 - (585.8445/568.1985))

sjPlot::tab_model(fit_spy_23)

### Likelikehood ratio test
lrtest_vglm(fit_spy_23, fit_spy_22)  #Not signficant difference so baseline odds might
extractAIC(fit_spy_22)
extractAIC(fit_spy_23)
# AIC says partial prop is better

lrtest_vglm(fit_spy_0, fit_spy_23)  #Signficant difference so baseline odds might be be
extractAIC(fit_spy_0)
extractAIC(fit_spy_23)
# AIC says baseline is better than null

# Baseline compared to partial
summary(fit_spy_23)
summary(fit_spy_22)
(psuedoRSquared <- 1 - (568.1985/569.3056))

# Baseline compared to no interactions
summary(fit_spy_23)
summary(fit_spy_0)
(psuedoRSquared <- 1 - (598.4781/568.1985))


# Interpreting the effects of the model Predict max open to
# close to min open to close at min rel volume signal - For
# outcome Y=5
pred_minOC_rldn1 <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "down",
    prev_OC_per_change = min(spy_data$prev_OC_per_change)), type = "response")
pred_minOC_rldn1
pred_maxOC_rldn <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "down",
    prev_OC_per_change = max(spy_data$prev_OC_per_change)), type = "response")
pred_maxOC_rldn


pred_minOC_rlup <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "up",
    prev_OC_per_change = min(spy_data$prev_OC_per_change)), type = "response")
pred_minOC_rlup
pred_maxOC_rlup <- predictvglm(fit_spy_11, data.frame(p_rel_vol_signal = "up",
    prev_OC_per_change = max(spy_data$prev_OC_per_change)), type = "response")
pred_maxOC_rlup
```