ALBERT–LUDWIGS–UNIVERSITÄT FREIBURG

M.Sc. Economics

Use Probit Models to Predict the Sign of Financial Returns

Master's Thesis

Examiner: Prof. Dr. Roxana Halbleib

Name: Ha Thu Tran

Matriculation Number: 5363960

Start of the Thesis: 05.12.2024

End of the Thesis: 24.04.2025

universitätfreiburg

Abstract

This thesis adapts the ordered probit model proposed by Hausman et al. (1992) to the recent high frequency financial world with IBM transaction price data on the New York Stock Exchange. The simplified model provides consistent findings with the original paper regarding the impact of order flows and trade size on the following price change. The thesis extends the application of the model and evaluates its forecasting ability. However, the mean percentage of correct prediction only ranges between 62-64%. The forecasting ability even weakens during the crisis periods, i.e., Covid-19 pandemic in 2008 and the financial crisis in 2008.

Contents

List of Figures

List of Tables

List of Algorithms

1. Introduction

Why should one pay attention to secondary financial markets, or namely one of those markets, the stock market? Bond et al. (2012), motivated to verify the impact of the secondary financial market on the real economy, argue that the role of market prices' information is critical to real decision makers (e.g., firm managers, customers, or employees). There are different ways to look at this role. The first instant can be when the decision makers decide based on stock price, e.g., credit rating. Even if one does not make a decision directly, one is still motivated to pay attention to stock price, for example, when an employee's compensation is affected by the company's share price. There are also developed theoretical models that contain the "feedback effect" between the financial market and the real economy.

As stock prices are informative and thus important, many efforts have been made in the literature to investigate their behavior in both theoretical and empirical territories. Looking back through the history of the trading world, from the times of crowded trading floors to the modern days with algorithmic trading, financial data structure transforms from lower frequencies (e.g., daily, weekly, monthly, or annually) to the (ultra-) high frequencies, such as milliseconds or nanoseconds. The high frequency financial data owns some specific characteristics, including the discreteness of transaction price and irregular time between trades (Tsay, 2013). The discreteness of transaction price challenges the traditional assumption of a continuous probability distribution. Hausman et al. (1992) address this problem by introducing an ordered probit model that considers price discreteness. Moreover, their model also captures the effect of market microstructure elements on price, verifying the "information-effect" theory by Easley and O'Hara (1987). One further application of their model, which interests many traders, is to predict the next price movement. Since then, although the so-called "low-latency" trading world continues to evolve and imposes many challenges, research directions (e.g., O'Hara (2015)), as well as newer machine learning models are providing promising results, the ordered probit model's ability to serve both market microstructure and forecasting questions is worth a kind of "test-of-time" study.

Hence, this thesis would revisit the Hausman et al. (1992)'s ordered probit model, attempting to examine the following objectives:

- *Does the information of the sequence of previous price changes affect the next price change? Does the trading volume of the previous trade affect the next price change?*

- and *How effective is our model in forecasting the stock price movement?*

By answering these questions, we hope to learn more about how the financial market microstructure works and the effectiveness of the ordered probit model in our settings.

The remainder of the thesis is organized as follows. Chapter 2 provides an overview of the related literature, specifically in terms of predicting stock price movement. Chapter 3 describes our simplified ordered probit model and its specification. Chapter 4 discusses the empirical results of the main study. Chapter 5 presents the further applications of our model to two special periods: the Covid-19 period in 2020, and the financial crisis in 2008. Chapter 6 reviews the limitations and challenges of this thesis and concludes our findings.

2. Related Works

This Chapter presents related studies on stock price movement prediction. The literature review first gives an overview of methods developed to forecast stock price movement. Afterward, the Chapter discusses publications relevant to ordered probit models, and their close family probit models.

Stock Price Movement Predictability

Determining whether stock returns are predictable is an open matter in the financial econometric literature. Rapach and Zhou (2013) survey time-series regression models with different indicators to forecast US equity premium. These models show weak stock returns predictability. More recently, machine learning approaches have received attention, though their performance still has a long way to go. One recent study by Kelly et al. (2024) shows theoretical support for complex and large machine learning models. A comparative study by Gu et al. (2020) suggests that neural networks and regression trees are the best-performing models to predict asset risk premiums.

While the aforementioned debate is on its way, another line of literature, as well as the focus of this thesis, is, instead, focusing on predicting the sign of returns. Predicting the direction of change is economically meaningful, for instance, in evaluating trading strategies. Earlier result by Leitch and Tanner (1991), considering the case of interest rate forecast, concludes that directional accuracy and profits are closely related. At the same time, that is not the case between traditional summary statistics (average absolute error and root-mean-square error) and profits. Christoffersen and Diebold (2006) find that sign predictability does not necessarily require conditional mean predictability. Christoffersen et al. (2007) further investigate the importance of higher-order conditional moments (skewness and kurtosis) to sign predictability. We can generally model the direction-of-change forecasting problem as a binary classification problem (e.g., going up/down). Different variations of logistic regression models are examined, for instance, by Rydberg and Shephard (2003) or by Anatolyev and Gospodinov (2010). In the first paper, Rydberg and Shephard (2003) decompose trade-by-trade price changes into activity (move or not), direction, and size of movement. Key explanatory variables are lagged variables of the components. In a different context, Anatolyev and Gospodinov (2010)'s decomposition includes the sign component and the absolute value of returns. Their set of predictors includes lagged signed returns/absolute value, along with other macroeconomic variables (e.g., dividend-price ratio, three-month T-bill rate). Beyond the paradigm of classical binary response models, one can also find recently increasing attempts to exploit the power of machine learning (e.g., survey by Bustos and Pomares-Quimbaya (2020)).

Probit Models in Forecasting Stock Price Movement

With an eye on using (ordered) probit models, Table 1 below summarizes relevant studies and their key findings. At first glance, there are two main strands of literature: The first strand mainly stems from the work of Hausman et al. (1992), in which they employ an ordered probit model to study transaction price movement. Some advantages of using ordered probit models are the ability to capture characteristics of high-frequency financial data, such as price discreteness, and to study the behavior of market microstructure. Our paper thus follows the approach of Hausman et al. (1992).

On the other hand, the second strand, including studies by Nyberg (2011), Nyberg and Pönkä (2016), and Pönkä (2017), investigates different variants of binary probit models to predict the direction of monthly excess stock returns. Last but not least, the last paper by Leung et al. (2000) compares the performance of classification models (e.g., probit/logit, probabilistic neural network) versus level models (e.g., vector autoregression with Kalman filter, multilayered feedforward neural network) in forecasting sign of monthly returns. In general, classification models perform better than level models.

Table 1: Probit Models in Forecasting Stock Price Movement.

Study	**Models**	**Data Frequency**	**Key Findings**
This thesis	*Ordered probit model*	*Transaction (tick) level price changes*	*Key Objectives: investigating the impact of sequence of trade and trade size on price, and the forecasting performance of the model.*
Hausman et al. (1992)	Ordered probit model	Transaction (tick) level price changes	The ordered probit model of this study captures discretness of price changes (clustering on eigths of a dollar). Sequence of past price changes and order flows (whether it is buyer-initiated or seller initiated), as well as trade size, do have an impact on transactional prices.
Yang and Parwada (2012)	Ordered probit model with GARCH(2,2) specification for residual series	Transaction (tick) level price changes	The model has an average of 71% accuracy rate in the forecasting exercises both in- and out-of-sample. Dominant buying transactions in the past would increase the probability of price rise, while dominant selling transaction would increase the probability of the price to fall. Volume at the best bid price would have positive impact on price, and in the other way round, the volume at the best ask price have a negative impact. Conditional durations have negative effect on price, suggesting by the joint negative sign of all stocks.

Continued on next page

Table 1: Probit Models in Forecasting Stock Price Movement. (Continued)

Study	**Models**	**Data Frequency**	**Key Findings**
Kim (2014)	Ordered probit model	Transaction (tick) level price changes	The paper revisits Hausman et al. (1992) after the NYSE decimalization: For small firms, the effect of trading-related explanatory variables are more evident to the 1/16th and 1/24th range of dependent variable than 1/8th range of dependent variable.
Nyberg (2011)	Probit models: Static probit, Dynamic probit, Autoregressive probit (Kauppi and Saikkonen, 2008), Dynamic autoregressive probit (Kauppi and Saikkonen, 2008), "Error correction" dynamic autoregressive probit	Monthly excess stock returns	It seems that the direction of the excess stock return is predictable, though with low statistical power. The proposed probit models also outperform the buy-and-hold trading strategy in terms of annualized portfolio returns. For out-of-sample results, the best performing model is the "error correction" dynamic probit model. Noticeably, models employing the recession forecast perform better than the ones incorporating the variables used in recession forecasting.

Continued on next page

Table 1: Probit Models in Forecasting Stock Price Movement. (Continued)

Study	**Models**	**Data Frequency**	**Key Findings**
Nyberg and Pönkä (2016)	Bivariate probit models	Monthly excess stock returns	The bivariate probit models study the interrelationship between US and ten industrialized countries, which allow testing the linkage of US excess return' sign forecast to other markets. In general, both in-sample and out-of-sample results show evidence that bivariate model outperform univariate model (that includes the lagged US return) in most of the markets. This suggests that the predictive power is not just limited to the lagged US return.
Pönkä (2017)	Probit models: Static probit, Dynamic probit, Autoregressive probit, Dynamic autoregressive probit	Monthly and daily excess stock returns	The predictability of the sign of excess market returns seems to be evident in out-of-sample results. Moreover, some industries portfolios do have predictive power for market returns, some of which later also reconfirmed by robustness check with daily data. The metal and construction industry portfolios especially seem to be useful in forecasts for trading strategies.

Continued on next page

Table 1: Probit Models in Forecasting Stock Price Movement. (Continued)

Study	**Models**	**Data Frequency**	**Key Findings**
Leung et al. (2000)	*Classification models*: Discriminant analysis, Logit, Probit, Probabilistic neural network; *Level models*: Adaptive exponential smoothing, Vector autoregression with Kalman filter, Multivariate transfer function, Multilayered feedforward neural network	Monthly excess stock returns	The classification models are employed to predict the sign of returns, while the level models are used to forecast the value. Generally, the classification models outperform level models in their forecasting performance i.e. have better hit rate. Noticeably, trading profits from the classification models are higher than that of the level models.

3. Theoretical Model

In this Chapter, the model specifications are defined and explained. Our ordered probit model is a simplified version of that of Hausman et al. (1992). Last but not least, the Chapter presents the forecasting evaluation metrics.

Convention: Sign of Returns versus Price Movement

In order to improve readability, the notion of *transaction price movement* is used interchangeably with *sign of transaction returns*. If the simple transaction-by-transaction, or **tick returns** is defined as:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}, \tag{1}$$

where P_t is the transaction price at time t, then the positive (negative) sign of returns is equivalent to the upward (downward) price move when $P_t > P_{t-1}$ ($P_t < P_{t-1}$).

3.1. Ordered Probit Model

Our ordered probit model set-up follows Hausman et al. (1992) with some simplification.

In the context of high frequency financial data, "tick" is the smallest unit of price movement. Let $Z_k \equiv P(t_k) - P(t_{k-1})$ be an integer of observed price changes, i.e. multiples of tick. The core idea of the ordered probit model is to map the observable discrete random variable Z_k with an unobserved continuous random variable Z_k^*. In practice, one can partition a finite number of price change categories into m state spaces (e.g. 1-tick change, 2-tick change and so on). Depending on where Z_k^* lies in the state space, Z_k is determined accordingly:

$$Z_k = \begin{cases} s_1 & \text{if } Z_k^* \in A_1, \\ s_2 & \text{if } Z_k^* \in A_2, \\ \vdots & \vdots \\ s_m & \text{if } Z_k^* \in A_m, \end{cases} \tag{2}$$

where s_j are discrete values forming state space $\mathscr{S}$ of Z_k, and A_j are the sets forming the partition of the state space $\mathscr{S}^*$ of Z_k^*.

The unobservable continuous random variable Z_k^* follows a model such that

$$Z_k^* = X_k'\beta + \varepsilon_k, \quad \mathbb{E}[\varepsilon_k \mid X_k] = 0, \quad \varepsilon_k \text{ i.n.i.d. } \mathcal{N}(0, \sigma_k^2), \tag{3}$$

where ε_k is independently but not identically distributed, and X_k is a $q \times 1$ vector of explanatory variables at time t_{k-1}.

Conditional Variance σ_k^2

Noticeably, Hausman et al. (1992)'s ordered probit model also accounts for conditional heteroskedasticity in Z_k^*. Intuitively, their model incorporates the clock-time effect where Z_k^* can be modeled as increments of arithmetic Brownian motion as in the model of Cho and Frees (1988). Let Δt_k be the time between trade k and $k-1$, the variance is linear in Δt_k:

$$X_k'\beta = \mu\Delta t_k, \quad \sigma_k^2 = \gamma^2 \Delta t_k. \tag{4}$$

Hausman et al. (1992) further extends the model and let the variance also depend linearly on other economic variables. However, in the scope of this thesis, we simplify our model by adopting the first specification as in Equation (4).

Explanatory Variables X_k

Our explanatory variables consist of:

- Z_{k-l}: Three lags of price changes variable Z_k (l = 1, 2, 3).
- IBS_{k-l}: Three lags of buyer/seller-initiated trade indicator (l = 1, 2, 3). The classification of trade direction is performed based on the method proposed by Lee and Ready (1991), detailed algorithm is described in Algorithm 1. This different approach than that of Hausman et al. (1992) avoids the indeterminate classification where IBS = 0.

$$\text{IBS}_{k-1} \equiv \begin{cases} 1, & \text{if buyer-initiated trade,} \\ -1, & \text{if seller-initiated trade.} \end{cases} \tag{5}$$

- $lnV_{k-l}IBS_{k-l}$: Three lags of signed transformed dollar volume (l = 1, 2, 3). Volume of a trade is first set to the 99.5 percentile of the traded volume distribution, if it exceeds the 99.5 percentile. Then, the traded volume is multiplied with the transaction price (in dollars), and also divided by \$100. Afterward, we take the natural log of dollar volume. The interaction between transformed dollar volume and the buyer/seller-initiated indicator is taken into account to investigate the impact of trade direction on price. A positive coefficient may indicate that buyer-initiated trade would likely to move price up, and vice versa (Hausman et al., 1992; Easley and O'Hara, 1987).

The complete specification of $X_k'\beta$ is:

$$\begin{aligned} X_k'\beta = \beta_1 Z_{k-1} + \beta_2 Z_{k-2} + \beta_3 Z_{k-3} \\ + \beta_4 IBS_{k-1} + \beta_5 IBS_{k-2} + \beta_6 IBS_{k-3} \\ + \beta_7 \ln V_{k-1} \cdot IBS_{k-1} \\ + \beta_8 \ln V_{k-2} \cdot IBS_{k-2} + \beta_9 \ln V_{k-3} \cdot IBS_{k-3}, \end{aligned} \tag{6}$$

with variance

$$\sigma_k^2 = \gamma^2 \Delta t_k. \tag{7}$$

Conditional Distribution and Log-Likelihood Function

Last but not least, under the assumption of ε_k following normal (or Gaussian) distribution:

$$P(Z_k = s_i \mid X_k, \Delta t_k) = P(\alpha_{m-1} < X_k'\beta + \varepsilon_k \le \alpha_1 \mid X_k, \Delta t_k) \tag{8}$$

$$= \begin{cases} P(X_k'\beta + \varepsilon_k \le \alpha_1 \mid X_k, \Delta t_k) & \text{if } i = 1, \\ P(\alpha_{i-1} < X_k'\beta + \varepsilon_k \le \alpha_i \mid X_k, \Delta t_k) & \text{if } 1 < i < m, \\ P(\alpha_{m-1} < X_k'\beta + \varepsilon_k \mid X_k, \Delta t_k) & \text{if } i = m, \end{cases} \tag{9}$$

$$= \begin{cases} \Phi\left(\frac{\alpha_1 - X_k'\beta}{\sigma_k}\right) & \text{if } i = 1, \\ \Phi\left(\frac{\alpha_i - X_k'\beta}{\sigma_k}\right) - \Phi\left(\frac{\alpha_{i-1} - X_k'\beta}{\sigma_k}\right) & \text{if } 1 < i < m, \\ 1 - \Phi\left(\frac{\alpha_{m-1} - X_k'\beta}{\sigma_k}\right) & \text{if } i = m, \end{cases} \tag{10}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Our model is then estimated with Maximum Likelihood via the following log-likelihood function with `oglmx` package in R:

$$
\begin{aligned}
\mathcal{L}(Z \mid X) = \sum_{k=1}^{n} \Bigg\{ & Y_{1k} \log \Phi\Big(\tfrac{\alpha_1 - X_k'\beta}{\sigma_k}\Big) \\
& + \sum_{i=2}^{m-1} Y_{ik} \log \Big[\Phi\Big(\tfrac{\alpha_i - X_k'\beta}{\sigma_k}\Big) - \Phi\Big(\tfrac{\alpha_{i-1} - X_k'\beta}{\sigma_k}\Big) \Big] \\
& + Y_{mk} \log \Big[1 - \Phi\Big(\tfrac{\alpha_{m-1} - X_k'\beta}{\sigma_k}\Big) \Big] \Bigg\} .
\end{aligned} \tag{11}
$$

3.2. Evaluation Metrics

Last but not least, as this thesis extends the purpose of Hausman et al. (1992)'s model to forecasting, besides the built-in (pseudo) McFadden's R^2 of the `oglmx` package, we adopt two more evaluation metrics to measure the predicting performance of our model:

- **% Accuracy**: For each observation, the prediction is taken with the category that has the highest probability. Then, the accuracy is calculated by comparing the prediction with the actual category of that observation. The final % accuracy is the mean of all % correct predictions.

- **Hand and Till (2001)'s Multiclass AUC (Area Under Curve)**: Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) are common metrics used in binary classification models to measure the ability to rank of the model. The ROC plots the curve by the False Positive Rate (X-axis) ranging from 0 to 1, and the True Positive Rate (Y-axis) also ranging from 0 to 1. Thus, the area under that ROC will have a value ranging from 0 to 1, with 1 being the perfect prediction and < 0.5 performing worse than a random guess. However, for the ordered probit model, it is not straightforward to measure this ranking performance. Hand and Till (2001) proposes a multiclass AUC version, where they evaluate the AUC between classes pairwise and then average the overall AUC. Although the multiclass AUC would be a good start to assess the ability to discriminate between classes of the model, it should be kept in mind that this metric does not consider ordering (or the distance between classes). The multiclass AUC is calculated with the `handtill2001` package.

4. Empirical Results

This Chapter discusses the empirical results at length. The empirical study begins with the essential data preparation steps. In the following part, the Chapter presents detailed descriptive statistics. The results for each research objective, i.e., market microstructure and forecasting, are then reviewed and explained.

4.1. Database

The data used for this study is from the tick-by-tick Trade and Quote database traded at the NYSE. As Hausman et al. (1992) use one-year data for their analysis, we also take the full trading sample of IBM stock on NYSE from January 3 to December 29 of 2023 for the descriptive statistics and market microstructure analysis parts. For the forecasting analysis, the first 10 months of 2023 will be used for in-sample forecasting, and the last 2 months will be left for out-of-sample forecasting.

Before arriving at the final dataset, we first match the trade and quote databases. Since the resolution of the current NYSE database is at nanosecond, we simply backward match the trade to the prevailing quote. We only use the data within the regular trading hours (9:30:00–16:00:00) and do not take into account overnight trading. The transactions happening on the first second of each trading day are removed to reduce contamination from the opening call at the stock exchange. Table 2 summarizes the number of observations used:

Purpose	Timeframe	Number of transactions
Panel A: Total	January 3 - December 29, 2023	1,905,393
Panel B: In-sample	January 3 - October 31, 2023	1,641,694 (86% of total sample)
Panel B: Out-of-sample	November 1 - December 29, 2023	263,699 (14% of total sample)

Table 2.: Summary of Database: IBM traded on NYSE, 2023.

4.2. Descriptive Statistics (Full Sample)

Price Change

Figure 1 depicts the histogram of transaction price changes for IBM on NYSE in 2023 by *tick* (equivalent to \$0.01 or 1 cent). The distribution of price changes seems to be almost symmetric around zero, which is expected for a very liquid stock like IBM. Moreover, the histogram shows that the majority of mass lies in the range of -4 to +4-tick changes. In order to also account for more extreme cases, we could choose the number of the state $m = 14$, including the changes from $<$-6 to $>$+6-tick.

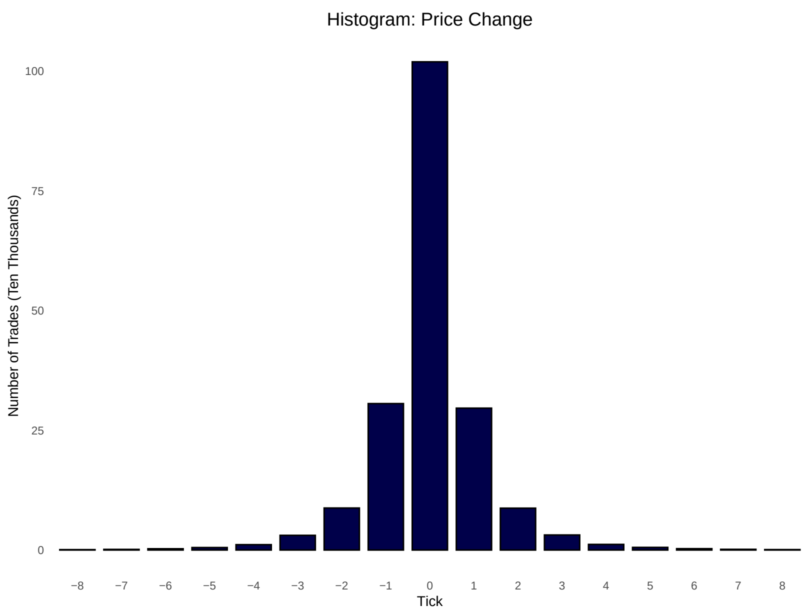

Figure 1.: Histogram of Price Changes: IBM on NYSE, 2023.

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Price Changes	<-6	$[-6,-5)$	$[-5,-4)$	$[-4,-3)$	$[-3,-2)$	$[-2,-1)$	$[-1,0)$	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	≥ 6
Percentage	0.31%	0.16%	0.33%	1.83%	2.68%	8.18%	11.08%	62.00%	7.99%	2.68%	1.89%	0.34%	0.17%	0.34%

Table 3.: Frequencies of Partition (Price Changes in Ticks).

Besides, most of the trade occurs at either higher or lower than mid-quote prices, similar to the results of Hausman et al. (1992) or Kim (2014).

% trades at prices	
$>$ Midquote	40.07
$=$ Midquote	16.65
$<$ Midquote	43.28
Price change, Z_k	
Mean	0.0000
Std. dev.	0.0134

Table 4.: Summary statistics: Price Changes

Trade Direction

As mentioned in the previous chapter, the trade direction is determined via the procedure proposed by Lee and Ready (1991) that incorporates two major steps. Firstly, the trade price is compared with the mid-quote to see if it is greater than mid-quote then it is buyer-initiated trade and vice versa, hereby noted as "mid-quote rule". Secondly, if the price is equal to the mid-quote, then a "tick-test" is applied. The "tick-test" will compare the current trade with the previous trade's price. If there is again no difference between the two consecutive prices, then the sign of the last non-zero change will be selected. Hausman et al. (1992) adopts only the "mid-quote rule" and thus has the third classification of "indeterminate" trade $IBS = 0$. If we used the same procedure, our buyer-initiated classification would simply be the percentage of trade with price > mid-quote (i.e., 40.07%), and seller-initiated classification would be 43.28%. Hence, our final results showed in Table 5 with roughly similar proportions (48.43% buyer-initiated versus 51.57% seller-initiated) are reasonable.

Trade direction, IBS_k	
Buyer-initiated (%)	48.43
Seller-initiated (%)	51.57
Mean	-0.0313
Std. dev.	0.9995

Table 5.: Summary statistics: Trade Direction.

Other Variables

Last but not least, Table 6 gives an overview of some other microstructure variables. IBM is traded even more frequently on the NYSE, with an average of three seconds. This might as well be the effect of high-frequency trading (HFT). The average bid/ask spread is also getting much smaller.

Statistic	This Thesis (Panel A)	Hausman et al. (1992)
Signed transformed volume[a]		
Mean	-0.1183	0.1059
Std. dev.	4.1303	6.1474
Time between trades (seconds)		
Mean	3.0685	27.21
Std. dev.	10.8139	34.13
Bid/ask spread		
Mean	0.0343	1.9470
Std. dev.	0.0311	1.4625

[a] Hausman et al. (1992) performs Box-Cox transformation for dollar volume and estimates the λ parameter via maximum likelihood. Their result for IBM is $\lambda = 0$, equivalent to natural log transformation. In this thesis, we simplify by directly taking natural log of dollar volume.

Table 6.: Summary statistics: Other Variables.

4.3. Panel A: Market Microstructure Analysis

Maximum Likelihood Estimates

Table 7 presents the maximum likelihood estimates for our ordered probit model. Thirteen partition thresholds for fourteen categories, as well as all parameters for explanatory variables and conditional variance, are statistically significant. While Hausman et al. (1992)'s results show a negative sign of all lagged price changes Z_k and argue that it might be a sign of price reversal, the same conclusion could not be found for our results.

Parameter	Value	Thresholds	Value	Statistics	Value
β_1: Z_{-1}	0.0319***	α_1	-1.9050***	AIC	4884659.97
	(0.0003)		(0.0036)		
β_2: Z_{-2}	0.0163***	α_2	-1.7981***	Log Likelihood	-2442306.98
	(0.0003)		(0.0031)		
β_3: Z_{-3}	0.0061***	α_3	-1.6489***	Num. obs.	1904393
	(0.0003)		(0.0026)		
β_4: IBS_{-1}	-0.0433***	α_4	-1.2736***	Iterations	8
	(0.0011)		(0.0017)		
β_5: IBS_{-2}	0.0149***	α_5	-1.0251***	McFadden's R^2	0.0743
	(0.0011)		(0.0014)		
β_6: IBS_{-3}	0.0064***	α_6	-0.6611***		
	(0.0011)		(0.0009)		
β_7: $lnV_{-1} \cdot IBS_{-1}$	0.0114***	α_7	-0.3961***		
	(0.0003)		(0.0007)		
β_8: $lnV_{-2} \cdot IBS_{-2}$	0.0016***	α_8	0.6121***		
	(0.0003)		(0.0009)		
β_9: $lnV_{-3} \cdot IBS_{-3}$	0.0009***	α_9	0.9822***		
	(0.0003)		(0.0013)		
γ: $\Delta t/100$ [a]	0.1424***	α_{10}	1.2328***		
	(0.0002)		(0.0017)		
		α_{11}	1.6137***		
			(0.0025)		
		α_{12}	1.7622***		
			(0.0030)		
		α_{13}	1.8751***		
			(0.0035)		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

[a] Note: Since the `oglmx` package (Carroll, 2018) only takes the standard deviation argument in the form of $\sigma = exp(\cdot)$, we take natural log of time between trade variable before estimate the model i.e., $\sigma = exp(ln(x)) = x$.

Table 7.: Ordered Probit Model Estimation (IBM on NYSE, 2023).

Impact of Sequence of Trade on the Next Price Change

As part of a study to explore the impact of trading volume on security price, Easley and O'Hara (1987) propose an "information-effects" theory. In this model, the price process will not follow a Markov process, i.e., information on the entire trade sequence is essential to calculate the distribution of the next price, not only the current price.

Putting this theory in the context of our ordered probit model, Hausman et al. (1992) discuss that if the entire sequence of trade would have an impact on the conditional mean of the next price, then one simple case we could test is whether the coefficients (β_1, β_2, β_3) of the lagged price changes different to each other or not. For example, in case the sequence of trade matters, the effect of the price change sequence of -1/1/-1 (total change -1 tick) on the conditional mean would be different than 1/-1/-1 (total change also -1 tick). The formulation of the test statistic used by Hausman et al. (1992) is similar to a joint Wald test for a linear hypothesis (Wooldridge, 2010):

$$H_0 : \beta_1 = \beta_2 = \beta_3,$$
$$H_1 : \text{At least one of } \beta_1, \beta_2, \beta_3 \text{ differs.}$$

Restriction:
$Z_{-1} - Z_{-2} = 0,$
$Z_{-2} - Z_{-3} = 0.$
Model 1: restricted model
Model 2: full model, as in Equation (6)

Df	χ^2	Pr($>\chi^2$)
2	2823.1	$< 2.2\text{e-}16$ ***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 8.: Sequence of Trade: Linear Hypothesis Test.

With this test result, we could reject the null hypothesis at 0.1% significance level. This finding is consistent with Hausman et al. (1992), and therefore also supports the "information-effects" theory of Easley and O'Hara (1987).

Impact of Trade Size on the Next Price Change

In our model, the impact of trade size or volume is measured with variable $lnV \cdot IBS$. The maximum likelihood estimates in Table 7 show consistent results with Hausman et al. (1992) that the coefficients are significantly positive, and the most recent transaction has the most impact.

However, the estimated coefficients relate to the impact on the conditional mean of the latent variable Z_k^*. Hausman et al. (1992) instead attempt to make inference on the impact of trade size on the observed price change Z_k. In general, they proceed to calculate the conditional mean for different scenarios following these steps:

1. **Determining specific values for explanatory variables** X_k: since our subject-of-interest is the impact of the last trade size, all other variables should be set to identical values while the first-lag of volume varies. The values of these variables are fixed at:
   - $V_{-2} = V_{-3} =$ median of dollar volume $\cdot \frac{1}{100}$. (Note that for lagged volumes, although in the original model we have the interaction term between volume and buyer/seller-initiated trade indicator IBS_k, here we only include lags of volumes.);
   - $IBS_{-1} = IBS_{-2} = IBS_{-3} = 1$ (Assuming the last three transactions are buy-initiated trades);
   - $\Delta t_k =$ sample mean;
2. **Determining sequence-of-trade scenarios**: as the sequence of price change does matter to the conditional mean, we would need to make assumption on the sequences we would like to investigate. Following Hausman et al. (1992) there are two cases:
   - Increasing price, where $Z_{-1} = Z_{-2} = Z_{-3} = +1$
   - Constant price, where $Z_{-1} = Z_{-2} = Z_{-3} = 0$;
3. **Calculating the probabilities**: by substituting the estimated coefficients, with the aforementioned fixed values of X_k, based on the conditional distribution Equation (10);
4. **Finally, calculating the expected price**;

Table 9 report the computed conditional means for each specific scenario. The first upper part of the table shows the price impact of trade size in ticks. Since our data in 2023 has much smaller tick size (1 tick = \$0.01) than in 1988 data of Hausman et al. (1992) (1 tick = \$0.125), the base price in this thesis is \$500 instead of \$5,000. Following the conditional mean of \$500 are the changes in conditional mean for the next additional dollar volume (e.g., $\Delta E[Z_k]$ of \$1,000 depicts the change of conditional mean when buying an additional \$500). Moreover, the conditional mean of \$500 is negative, which might be affected by the bid/ask bounce. Hausman et al. (1992)

argue that while we assume all three previous trades are buys ($IBS_k = +1$), the next trade could be a sell leading to the negative sign. This effect could be dismissed by calculating the difference in conditional mean for the larger dollar volume than the base $500, and hence, we could still capture the price impact effects of trade size. Our results for both increasing- and constant-price-sequence show the consistent message with their study: a larger trade size would have a larger price impact.

Along with the price impact measured in ticks, the price impact in percent of Table 9 shows the changes as the percentages of the average of the max ($166.34) and min ($120.55) trade price. As IBM is a very liquid stock, the trade size is expected to have minimal impact on price. Moreover, Figure 2 plots the price impact in percent as a function of the dollar volume, i.e., a "price response function" for our IBM stock, conditioning on the increasing-price-sequence as well as the last three trades are buys. A very flat curve again illustrates the high liquid characteristic of IBM (Hausman et al., 1992).

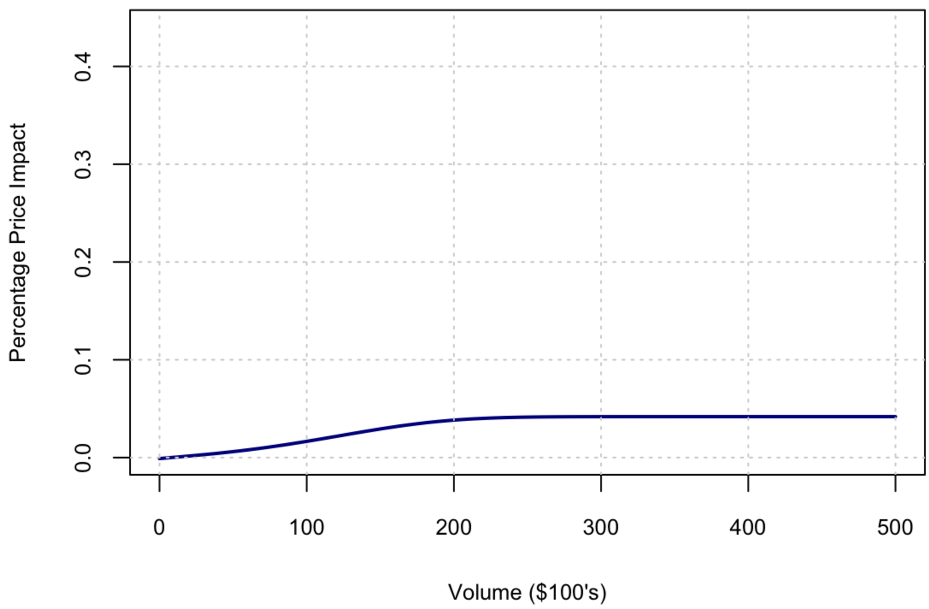

Figure 2.: Price Response Function for IBM on NYSE, 2023.

	$Volume	Impact
Increasing price sequence (1/1/1)		
Price impact in ticks		
$\mathrm{E}[Z_k]$	500	-0.0545
$\Delta E[Z_k]$	1,000	0.0943
$\Delta E[Z_k]$	1,500	0.1887
$\Delta E[Z_k]$	2,000	0.2843
$\Delta E[Z_k]$	2,500	0.3818
$\Delta E[Z_k]$	5,500	1.0496
Price impact in percent		
$\mathrm{E}[Z_k]$	500	-0.0004
$\Delta E[Z_k]$	1,000	0.0007
$\Delta E[Z_k]$	1,500	0.0013
$\Delta E[Z_k]$	2,000	0.0020
$\Delta E[Z_k]$	2,500	0.0027
$\Delta E[Z_k]$	5,500	0.0073
Constant price sequence (0/0/0)		
Price impact in ticks		
$\mathrm{E}[Z_k]$	500	-0.1450
$\Delta E[Z_k]$	1,000	0.0949
$\Delta E[Z_k]$	1,500	0.1891
$\Delta E[Z_k]$	2,000	0.2836
$\Delta E[Z_k]$	2,500	0.3793
$\Delta E[Z_k]$	5,500	1.0217
Price impact in percent		
$\mathrm{E}[Z_k]$	500	-0.0010
$\Delta E[Z_k]$	1,000	0.0007
$\Delta E[Z_k]$	1,500	0.0013
$\Delta E[Z_k]$	2,000	0.0020
$\Delta E[Z_k]$	2,500	0.0026
$\Delta E[Z_k]$	5,500	0.0071

Table 9.: Impact of Trade Size.

4.4. Panel B: Forecasting Analysis

In-sample

Our in-sample period from January 2023 to October 2023 is first estimated using the ordered probit model as in the previous section. Table 10 presents the maximum likelihood estimation results. Similar to the full sample estimation, our in-sample coefficients are all statistically significant. The pseudo R^2, i.e., McFadden's R^2, is also low as expected. Nevertheless, when we would like to extend the model (e.g.,, add more explanatory variables), we could use this as an evaluation metric.

Parameter	Value	Thresholds	Value	Statistics	Value
β_1: Z_{-1}	0.0303***	α_1	-1.8903***	AIC	4226828.7780
	(0.0004)		(0.0038)		
β_2: Z_{-2}	0.0157***	α_2	-1.7849***	Log Likelihood	-2113391.3890
	(0.0004)		(0.0033)		
β_3: Z_{-3}	0.0059***	α_3	-1.6331***	Num. obs.	1640858
	(0.0004)		(0.0027)		
β_4: IBS_{-1}	-0.0403***	α_4	-1.2617***	Iterations	8
	(0.0012)		(0.0018)		
β_5: IBS_{-2}	0.0136***	α_5	-1.0195***	McFadden's R^2	0.0748
	(0.0012)		(0.0014)		
β_6: IBS_{-3}	0.0056***	α_6	-0.6513***		
	(0.0012)		(0.0010)		
β_7: $lnV_{-1} \cdot IBS_{-1}$	0.0110***	α_7	-0.3930***		
	(0.0003)		(0.0007)		
β_8: $lnV_{-2} \cdot IBS_{-2}$	0.0017***	α_8	0.6026***		
	(0.0003)		(0.0010)		
β_9: $lnV_{-3} \cdot IBS_{-3}$	0.0010***	α_9	0.9764***		
	(0.0003)		(0.0014)		
γ: $\Delta t/100$	0.1432***	α_{10}	1.2210***		
	(0.0002)		(0.0018)		
		α_{11}	1.5978***		
			(0.0027)		
		α_{12}	1.7484***		
			(0.0032)		
		α_{13}	1.8597***		
			(0.0037)		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 10.: In-sample Estimation (January 3 - October 31, 2023).

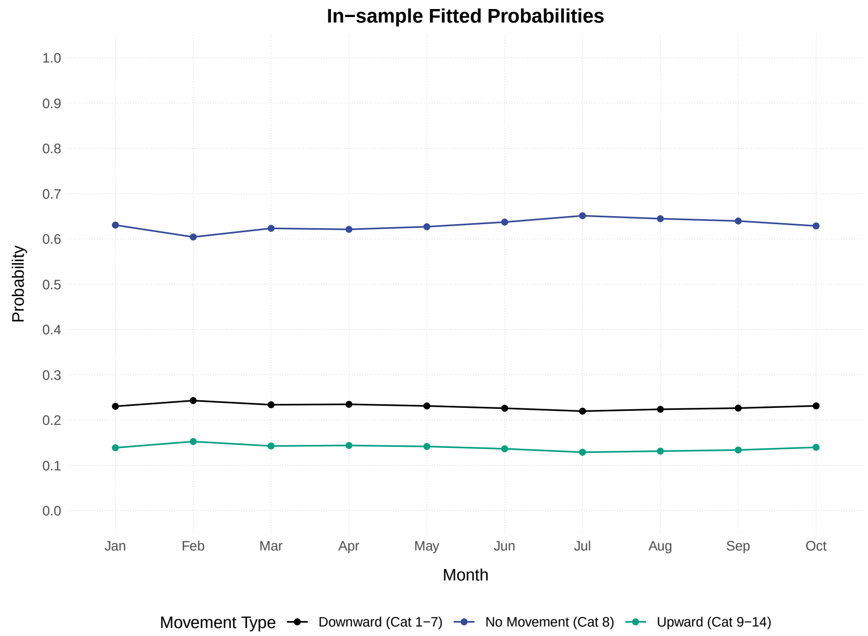

Figure 3.: Aggregated Monthly Fitted Probability for In-sample Period (January - October 2023).

To have a simpler illustration regarding the transaction price movement, we could cluster our 14 categories into three major movements: downward movement (category 1 to 7), no change (category 8), and upward movement (category 9 to 14). (The definition of each category could be reviewed in Table 3.) Figure 3 illustrates the average monthly fitted probability for the in-sample observations. The sum of the fitted probability of falling into each category for each observation equals one (1). Accordingly, the *no movement* category has the highest fitted probability. Nevertheless, the primary evaluation metrics for in-sample prediction will be discussed with the following out-of-sample results.

Out-of-sample

Table 11 presents the forecasting performance of both in- and out-of-sample cases. The prediction accuracy is relatively low for both cases. Noticeably, the multiclass AUC values are only slightly above 0.5, which means the model barely performs better than a random guess.

Metric	**% Accuracy**	**Hand-Till Multiclass AUC**
In-sample	61.77 %	0.5732
Out-of-sample	63.54 %	0.5697

Table 11.: Forecasts Evaluation for In- and Out-of-sample.

One possible culprit for this poor performance might be over-predicting the *no change* category. In Figure 4, we could observe that more than 60% of fitted probabilities are falling into *no change*. Previous studies employing similar ordered probit models also capture this tendency, with more than 50% prediction as *no change* (Yang and Parwada, 2012; Kim, 2014).

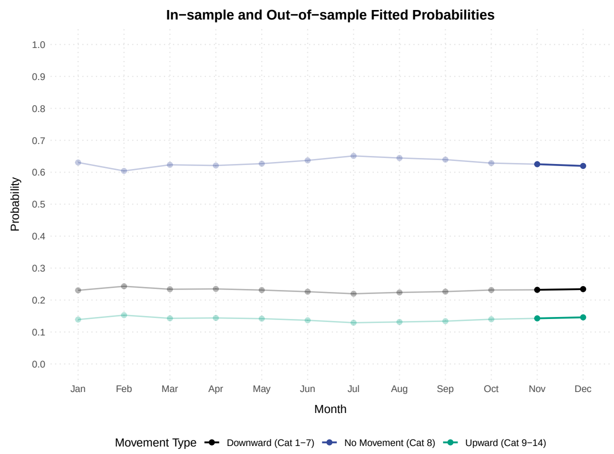

Figure 4.: Aggregated Monthly Fitted Probability for In-sample (January - October 2023) & Out-of-sample (November - December 2023).

Figure 5 again shows the major miss-classification of another category into category 8, i.e., *no change in price*. In general, our simplified ordered probit model seems not to be able to handle the class imbalance, as 62% of our data is in category 8 (see Table 3).

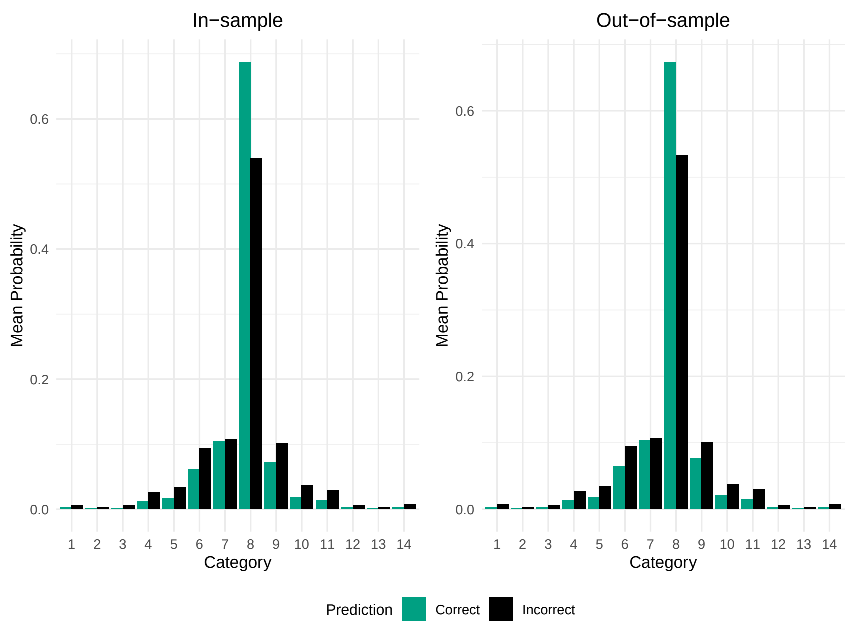

Figure 5.: Mean Probability by Prediction Class of In- and Out-of-sample.

5. Further Applications

This Chapter further applies the ordered probit model to two special periods. The first section provides and discusses the main findings of the Covid-19 period. Afterward, the second section presents the results of the financial crisis in 2008. Both sections follow the same structure as the main empirical result chapter.

5.1. Covid-19 Period (February-April 2020)

The Covid-19 pandemic has brought significant disturbance to (not only) the financial and economic world. Since the stock data analyzed throughout this thesis is IBM traded on the NYSE, the chosen focus is on the US stock market. Cox et al. (2020) observe the turbulence of the S&P 500 stock market index in the period between February to April 2020: from February 19 to March 23, the index lost 33.7%, but rose back 29% from March 24 to April 17. Hence, the data used in this section are taken from February to April 2020 to test the reaction of our model towards these disrupting times.

Similar to the main study, this section for the Covid-19 period also has two panels: one for the full-sample market microstructure study and the other for the forecasting study. Table 12 presents the database structure. For the forecasting panel, the first ten weeks are taken for in-sample analysis, and the last three are for out-of-sample analysis.

Purpose	Timeframe	Number of transactions
Panel C: Total	February 3 - April 30, 2020	786,598
Panel D: In-sample	February 3 - April 11, 2020	660,487 (84% of total sample)
Panel D: Out-of-sample	April 12 - April 30, 2020	126,111 (16% of total sample)

Table 12.: Summary of Database: IBM traded on NYSE (February-April 2020).

Table 13 shows the frequency distribution of the Covid-19 period by each category. Noticeably, although the majority of mass still clusters around zero (category 8), there are more percentages at the extreme cases (<-6 and

≥6). Besides, the summary statistics of the explanatory variables (lagged price changes Z_k, trade direction indicator IBS_k, signed transformed volume $lnV_k \cdot IBS_k$) could be found from Table 22 to Table 24 in Appendix A.1. The proportion between price traded > Miquote and < Midquote is still roughly balanced (48.32% vs. 45.44%), and a similar picture could also be detected for trade direction (51.45% buyer-initiated vs. 48.55% seller initiated). The average time between trades is 1.8 seconds.

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Price Changes	<–6	[–6,–5)	[–5,–4)	[–4,–3)	[–3,–2)	[–2,–1)	[–1,0)	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	≥ 6
Percentage	2.98%	0.57%	2.04%	4.14%	2.39%	9.58%	5.71%	51.28%	9.28%	2.37%	4.04%	1.96%	0.55%	3.08%

Table 13.: Frequencies of Partition (February-April 2020).

The ordered probit model is estimated for the Covid-19 period data, detailed estimation results are reported in Table 25 in Appendix A.1. The McFadden's R^2 is low with 0.0914. After estimating the model with maximum likelihood, as before, the impact of sequence-of-trade and trading volume on price change are tested. The significant χ^2 at 0.1% level from Table 14 again rejects the null hypothesis that there is no difference between order of trade i.e., order flow matters.

Df	χ^2	$\Pr(>\chi^2)$
2	2835.2	$< 2.2\text{e-}16$ ***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 14.: Sequence of Trade: Linear Hypothesis Test (February-April 2020).

In Table 15, the conditional mean for each dollar trading volume case is presented. The finding is consistent with the main study that a larger trade size would move the subsequent price changes upward. Noticeably, the absolute magnitude of the changes in conditional mean for both impact in ticks and impact in percent is larger than the main study period (2023). For instance, conditioning for the increasing price sequence scenario and the base price of \$500, an increase of \$1000 trading volume in Covid-19 would increase the conditional mean by 0.33 tick, while in the main study period (2023), it only increases by 0.19 tick (see Table 9). Although more tests are needed to verify whether these gaps are significant, one possible explanation is the decrease of market liquidity on NYSE during the Covid-19 period (Chung and Chuwonganant, 2023).

	$Volume	Impact
Increasing price sequence (1/1/1)		
Price impact in ticks		
$\mathrm{E}[Z_k]$	500	0.0802
$\Delta E[Z_k]$	1,000	0.1664
$\Delta E[Z_k]$	1,500	0.3337
$\Delta E[Z_k]$	2,000	0.5027
$\Delta E[Z_k]$	2,500	0.6742
$\Delta E[Z_k]$	5,500	1.7823
Price impact in percent		
$\mathrm{E}[Z_k]$	500	0.0006
$\Delta E[Z_k]$	1,000	0.0013
$\Delta E[Z_k]$	1,500	0.0027
$\Delta E[Z_k]$	2,000	0.0040
$\Delta E[Z_k]$	2,500	0.0054
$\Delta E[Z_k]$	5,500	0.0143
Constant price sequence (0/0/0)		
Price impact in ticks		
$\mathrm{E}[Z_k]$	500	−0.0144
$\Delta E[Z_k]$	1,000	0.1663
$\Delta E[Z_k]$	1,500	0.3330
$\Delta E[Z_k]$	2,000	0.5009
$\Delta E[Z_k]$	2,500	0.6709
$\Delta E[Z_k]$	5,500	1.7655
Price impact in percent		
$\mathrm{E}[Z_k]$	500	−0.0001
$\Delta E[Z_k]$	1,000	0.0013
$\Delta E[Z_k]$	1,500	0.0027
$\Delta E[Z_k]$	2,000	0.0040
$\Delta E[Z_k]$	2,500	0.0054
$\Delta E[Z_k]$	5,500	0.0142

Table 15.: Impact of Trade Size (February-April 2020).

Finally, the in-sample panel is estimated before computing the in- and out-of-sample fitted probabilities. For the Covid-19 period, the ordered probit model performs worse than the main study in 2023. As shown in

Table 16, % accuracy is only 51.8% for in-sample and 48.6% for out-of-sample. Besides, the tendency of predicting *no change* movement (or category 8) is again prominent as depicted in Figure 6.

Metric	**% Accuracy**	**Hand-Till Multiclass AUC**
In-sample	51.8 %	0.6133
Out-of-sample	48.6 %	0.6088

Table 16.: Forecasts Evaluation for In- and Out-of-sample (February-April 2020).

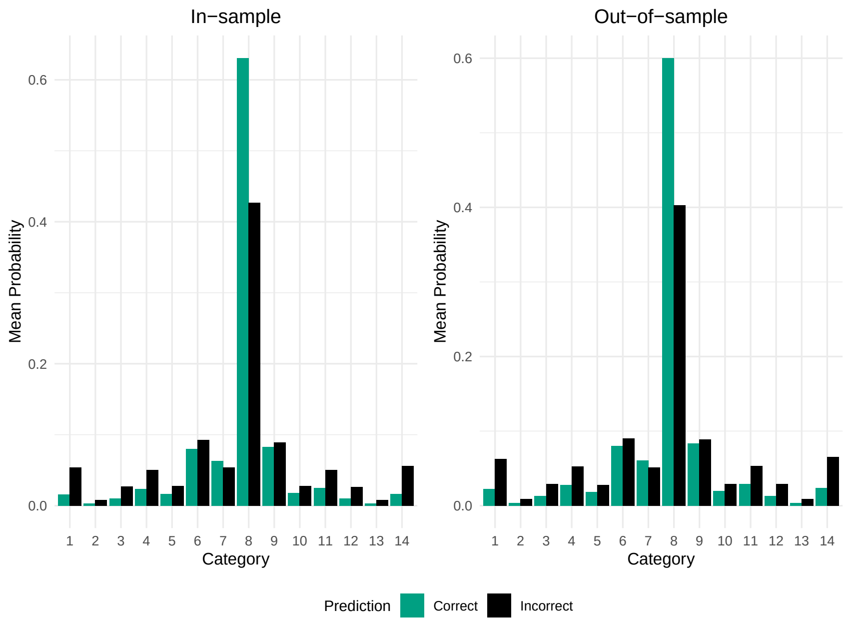

Figure 6.: Mean Probability by Prediction Class of In- and Out-of-sample (February-April 2020).

5.2. 2008 Financial Crisis (September-October 2008)

On September 15, 2008, Lehman Brothers, back then one of the largest investment banks in the US, filed for bankruptcy. Following its collapse, the Dow Jones Industrial Average (DJIA) went down by 500 points on that trading day (Johnson and Mamun, 2012). This is one of the key events that caused the global financial crisis to last roughly until 2009, or one could go with the "Great Recession" (Islam and Verick, 2011). For this thesis, the investigated period is from September 2 to October 31, with seven weeks for the in-sample and two weeks for the out-of-sample forecast. A complete event study would need to look at a more extended period. Nevertheless, Table 17 describes the sample size for this chosen period.

Table 18 presents the distribution of price changes by tick as well as by our category's definition. Comparing to the previous case study of the recent Covid-19 pandemic, there are even more mass in the two tails. Moreover, further summary statistics of other X_k variables are shown from Table 26 to Table 28 in Appendix A.2. Overall, the trade direction proportion is roughly even between buyer-initiated (49.9%) and seller-initiated (50.1%). The bid/ask spread (mean of 0.08) is quite similar to the previous Covid-19 study (mean of 0.09), and both of these periods have wider spread than the main study's period (mean of 0.03).

Moreover, Table 29 in Appendix A.2 reports the ordered probit model's estimation results. Accordingly, when roughly comparing with our previous studies in 2023 and Covid-19 2020, this model performs worst with McFadden's R^2 of only 0.0252. Only the first lag of price changes estimate is statistically significant at 0.1% level as in the main study, while the second is only significant at 5% level, and the third lag coefficient is not statistically significant.

Purpose	**Timeframe**	**Number of transactions**
Panel E: Total	September 2 - October 31, 2008	508,714
Panel F: In-sample	September 2 - October 19, 2008	390,540 (77% of total sample)
Panel F: Out-of-sample	October 20 - October 31, 2008	118,174 (23% of total sample)

Table 17.: Summary of Database: IBM traded on NYSE (September-October 2008).

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Price Changes	<-6	[–6,–5)	[–5,–4)	[–4,–3)	[–3,–2)	[–2,–1)	[–1,0)	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	≥ 6
Percentage	6.53%	0.79%	3.48%	4.95%	1.99%	9.98%	4.10%	40.81%	9.70%	1.98%	4.95%	3.50%	0.80%	6.44%

Table 18.: Frequencies of Partition (September-October 2008).

The test statistic for the effect of order flow is performed and presented in Table 19 with statistical significant result that is able to reject the null hypothesis. In summary, throughout all of our three case studies, our model' results show consistent support for Easley and O'Hara (1987) "information-effect" theory.

Df	χ^2	$\Pr(>\chi^2)$
2	135.71	$< 2.2\text{e-}16$ ***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 19.: Sequence of Trade: Linear Hypothesis Test (September-October 2008).

Moving on to the impact of trade size on the next price change, Table 20 presents the expected price impact for our dollar volume scenarios. Besides the similar trend of the larger volume the larger the impact, as noted for the Covid-19 period, the absolute impact in ticks for increasing amount of dollar volume for this "Great Recession" period is even greater. Again, conditioning for the increasing price sequence scenario and the base price of \$500, an increase of \$1000 trading volume in our financial crisis 2008 period would increase the conditional mean by 0.73 tick, while the impact of the previous studies are 0.33 tick (February-April 2020) and 0.19 tick (2023). A deeper analysis would needed to be able to investigate whether the seemingly decreasing in trade size's impact due to the increasing liquidity of IBM stock through time, or it due to the effects of the crisis periods that cause the turbulence.

	$Volume	Impact
Increasing price sequence (1/1/1)		
Price impact in ticks		
$\mathrm{E}[Z_k]$	500	−0.1707
$\Delta E[Z_k]$	1,000	0.3654
$\Delta E[Z_k]$	1,500	0.7286
$\Delta E[Z_k]$	2,000	1.0898
$\Delta E[Z_k]$	2,500	1.4490
$\Delta E[Z_k]$	5,500	3.5094
Price impact in percent		
$\mathrm{E}[Z_k]$	500	−0.0017
$\Delta E[Z_k]$	1,000	0.0036
$\Delta E[Z_k]$	1,500	0.0072
$\Delta E[Z_k]$	2,000	0.0107
$\Delta E[Z_k]$	2,500	0.0143
$\Delta E[Z_k]$	5,500	0.0346
Constant price sequence (0/0/0)		
Price impact in ticks		
$\mathrm{E}[Z_k]$	500	−0.1550
$\Delta E[Z_k]$	1,000	0.3653
$\Delta E[Z_k]$	1,500	0.7284
$\Delta E[Z_k]$	2,000	1.0895
$\Delta E[Z_k]$	2,500	1.4487
$\Delta E[Z_k]$	5,500	3.5072
Price impact in percent		
$\mathrm{E}[Z_k]$	500	−0.0015
$\Delta E[Z_k]$	1,000	0.0036
$\Delta E[Z_k]$	1,500	0.0072
$\Delta E[Z_k]$	2,000	0.0107
$\Delta E[Z_k]$	2,500	0.0143
$\Delta E[Z_k]$	5,500	0.0346

Table 20.: Impact of Trade Size (September-October 2008).

Metric	% Accuracy	Hand-Till Multiclass AUC
In-sample	41.47 %	0.5661
Out-of-sample	38.65 %	0.5796

Table 21.: Forecasts Evaluation for In- and Out-of-sample (September-October 2008).

Last but not least, the forecasting performance of our ordered probit model is also the worst among studies' periods, with both in-sample and out-of-sample % accuracy below 50% reporting in Table 21. The model undoubtedly needs further refinement to be robust for different periods.

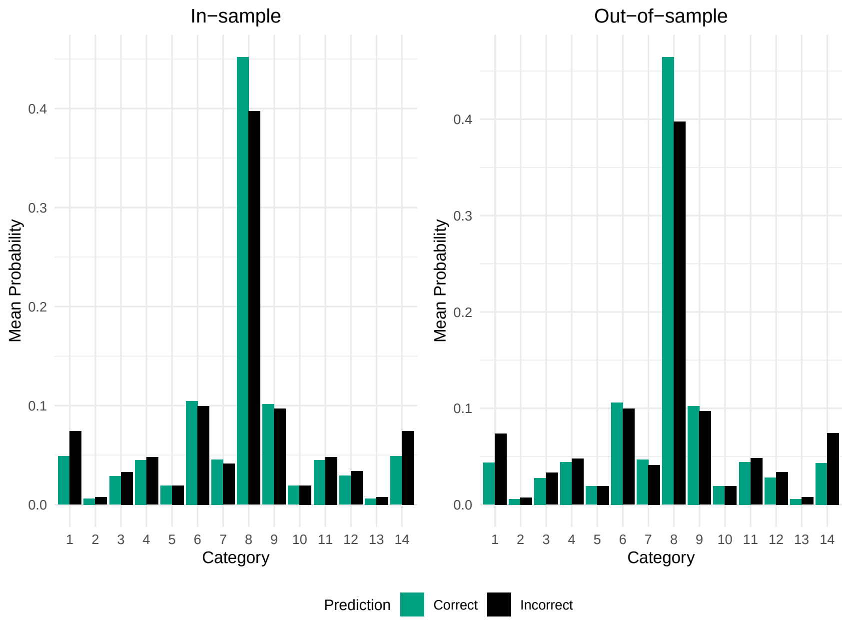

Figure 7.: Mean Probability by Prediction Class of In- and Out-of-sample (September-October 2008).

6. Conclusion

This Chapter discusses the limitations and challenges imposed on this thesis, explores possible future improvements, and concludes with the main findings and remarks.

Limitation, Challenges, & Outlook

This thesis faces several limiting factors. At first, the simplified ordered probit model does not incorporate the full list of explanatory variables proposed by Hausman et al. (1992) that could impact the model's performance substantially, such as the S&P 500 index that could be informative about the overall US financial market. The signed transformed volume is separately pre-transformed with natural log, while in the original study, the Box-Cox transformation is estimated jointly with maximum likelihood estimation. Furthermore, the conditional variance that accounts for conditional heteroskedasticity is simplified to only time between trade. The restrictive usage of the `oglmx` package constrains how the variance could be modeled. A customized, self-developed function would be needed to replicate the ordered probit model completely. However, due to time and resource constraints, so far, this thesis only investigates IBM, a large and very liquid stock. To arrive at a robust and generalized conclusion, one should need a more representative number and size of stocks.

Secondly, the changing nature of the high frequency financial data characteristic, compared to the 1992 period of the original paper, leaves this thesis further challenges in cleaning and pre-processing the data. One instant would be to verify whether our choice of trade direction classification's algorithm is already the most suitable, though the % trade at a price equal to mid-quote is minor. The issue of split transaction (Hautsch, 2012) is also not addressed in this thesis, which could be a significant issue when not treated right, leading to the over-representation of the *no change* price movement in the dataset.

Lastly, updating the results with the more recent market microstructure theory regarding information-based trading would be more insightful. For the time being, the relationships between order flow and price change (or between trade size and price changes) have not been discussed in depth. The thesis only revisits the test statistics and experiments from Hausman et al. (1992) regarding the behavior of the market microstructure.

Concluding Remarks

The ordered probit model introduced by Hausman et al. (1992) provides a means to model the transaction price movement that accounts for the discreteness characteristic of price change and conditional heteroskedasticity by inducing conditional variance. This thesis attempts to reconstruct a simplified version of the ordered probit model to explore some aspects of the financial market microstructure. In particular, our model shows consistent results with the original study regarding the effect of trading sequence on the next price change. Indeed, the order flows (e.g., buy/sell/buy or sell/buy/buy) do have an impact on the next price change. Moreover, the model allows us to estimate the trade size's impact on the conditional mean of price change in different scenarios, i.e., constant price sequence of increasing price sequence. The result shows a trend of larger trade size leading to a larger price impact.

In addition, this thesis evaluates the forecasting performance of our simplified ordered probit model. The prediction ability is relatively weak. For the main study of 2023's IBM stock traded on NYSE, % accuracy ranges only from 62-64%. The forecasting performance is even poorer when applying for the later special periods, the Covid-19 pandemic and the financial crisis of 2008. Noticeably, it is evident that the model suffers from class imbalance, and tends to predict *no change* when actually there is a price change. This issue might require us to develop a more sophisticated model for forecasting purposes. Despite the limitations and challenges, the ordered probit model provides informative findings on how stock price moves and its market microstructure behave.

Bibliography

Anatolyev, S. and Gospodinov, N. (2010). Modeling Financial Return Dynamics via Decomposition. *Journal of Business & Economic Statistics*, 28(2):232–245.

Bond, P., Edmans, A., and Goldstein, I. (2012). The Real Effects of Financial Markets. *Annual Review of Financial Economics*, 4(1):339–360.

Bustos, O. and Pomares-Quimbaya, A. (2020). Stock market movement forecast: A Systematic review. *Expert Systems with Applications*, 156:113464.

Carroll, N. (2018). oglmx: A Package for Estimation of Ordered Generalized Linear Models.

Cho, D. C. and Frees, E. W. (1988). Estimating the Volatility of Discrete Stock Prices. *The Journal of Finance*, 43(2):451–466.

Christoffersen, P., Diebold, F. X., Mariano, R. S., Tay, A. S., and Tse, Y. K. (2007). Direction-of-Change Forecasts Based on Conditional Variance, Skewness and Kurtosis Dynamics: International Evidence. *SSRN Electronic Journal*.

Christoffersen, P. F. and Diebold, F. X. (2006). Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics. *Management Science*, 52(8):1273–1287.

Chung, K. H. and Chuwonganant, C. (2023). COVID-19 pandemic and the stock market: Liquidity, price efficiency, and trading. *Journal of Financial Markets*, 64:100803.

Cox, J., Greenwald, D., and Ludvigson, S. (2020). What Explains the COVID-19 Stock Market? Technical Report w27784, National Bureau of Economic Research, Cambridge, MA.

Easley, D. and O'Hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19(1):69–90.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hand, D. J. and Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45:171–186.

Hausman, J. A., Lo, A. W., and MacKinlay, A. C. (1992). An ordered probit analysis of transaction stock prices. *Journal of Financial Economics*, 31:319–379.

Hautsch, N. (2012). *Econometrics of Financial High-Frequency Data*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Islam, I. and Verick, S. (2011). *The Great Recession of 2008–09: Causes, Consequences and Policy Responses*, pages 19–52. Palgrave Macmillan UK, London.

Johnson, M. A. and Mamun, A. (2012). The failure of Lehman Brothers and its impact on other financial institutions. *Applied Financial Economics*, 22(5):375–385.

Kauppi, H. and Saikkonen, P. (2008). Predicting U.S. Recessions with Dynamic Binary Response Models. *Review of Economics and Statistics*, 90(4):777–791.

Kelly, B., Malamud, S., and Zhou, K. (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance*.

Kim, H. (2014). Stock price discreteness and clustering: decimals and ordered probit model. *Journal of Financial Regulation and Compliance*, 22(1):49–60.

Lee, C. M. C. and Ready, M. J. (1991). Inferring Trade Direction from Intraday Data. *The Journal of Finance*, 46(2):733–746.

Leitch, G. and Tanner, J. E. (1991). Economic Forecast Evaluation: Profits Versus The Conventional Error Measures. *The American Economic Review*.

Leung, M. T., Daouk, H., and Chen, A.-S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, 16:173–190.

Nyberg, H. (2011). Forecasting the direction of the US stock market with dynamic binary probit models. *International Journal of Forecasting*, 27(2):561–578.

Nyberg, H. and Pönkä, H. (2016). International sign predictability of stock returns: The role of the United States. *Economic Modelling*, 58:323–338.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116(2):257–270.

Pönkä, H. (2017). Predicting the direction of US stock markets using industry returns. *Empirical Economics*, 52(4):1451–1480.

Rapach, D. and Zhou, G. (2013). Forecasting Stock Returns. In *Handbook of Economic Forecasting*, volume 2, pages 328–383. Elsevier.

Rydberg, T. H. and Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics*.

Tsay, R. S. (2013). *An Introduction to Analysis of Financial Data with R*. John Wiley Sons.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, Mass, 2nd ed edition.

Yang, J. W. and Parwada, J. (2012). Predicting stock price movements: an ordered probit analysis on the Australian Securities Exchange. *Quantitative Finance*, 12(5):791–804.

A. Additional Tables

A.1. Covid-19 Period (February-April 2020)

% trades at prices	
> Midquote	48.32
= Midquote	6.24
< Midquote	45.44
Price change, Z_k	
Mean	0.0000
Std. dev.	0.0294

Table 22.: Summary statistics: Price Changes (February-April 2020).

Trade direction, IBS_k	
Buyer-initiated (%)	51.45
Seller-initiated (%)	48.55
Mean	0.0290
Std. dev.	0.9996

Table 23.: Summary statistics: Trade Direction (February-April 2020).

Statistic	This Thesis (Panel C)	Hausman et al. (1992)
Signed transformed volume		
Mean	0.1188	0.1059
Std. dev.	4.1146	6.1474
Time between trades (seconds)		
Mean	1.8410	27.21
Std. dev.	4.7175	34.13
Bid/ask spread		
Mean	0.0877	1.9470
Std. dev.	0.0795	1.4625

Table 24.: Summary statistics: Other Variables (February-April 2020).

Parameter	Value	Thresholds	Value	Statistics	Value
β_1: Z_{-1}	0.0142***	α_1	-1.0118***	AIC	2597221.9548
	(0.0002)		(0.0023)		
β_2: Z_{-2}	0.0066***	α_2	-0.9589***	Log Likelihood	-1298587.9774
	(0.0002)		(0.0021)		
β_3: Z_{-3}	0.0024***	α_3	-0.8152***	Num. obs.	786350
	(0.0002)		(0.0018)		
β_4: IBS_{-1}	-0.0087***	α_4	-0.6269***	Iterations	9
	(0.0013)		(0.0015)		
β_5: IBS_{-2}	0.0197***	α_5	-0.5492***	McFadden's R^2	0.0914
	(0.0013)		(0.0013)		
β_6: IBS_{-3}	0.0108***	α_6	-0.3395***		
	(0.0012)		(0.0009)		
β_7: $lnV_{-1} \cdot IBS_{-1}$	0.0082***	α_7	-0.2522***		
	(0.0003)		(0.0008)		
β_8: $lnV_{-2} \cdot IBS_{-2}$	-0.0013***	α_8	0.3323***		
	(0.0003)		(0.0009)		
β_9: $lnV_{-3} \cdot IBS_{-3}$	-0.0007*	α_9	0.5441***		
	(0.0003)		(0.0013)		
γ: $\Delta t/100$	0.2013***	α_{10}	0.6240***		
	(0.0004)		(0.0015)		
		α_{11}	0.8135***		
			(0.0018)		
		α_{12}	0.9546***		
			(0.0021)		
		α_{13}	1.0064***		
			(0.0023)		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 25.: Ordered Probit Model Estimation (February-April 2020).

A.2. 2008 Financial Crisis (September-October 2008)

% trades at prices	
> Midquote	48.82
= Midquote	2.20
< Midquote	48.98
Price change, Z_k	
Mean	-0.0001
Std. dev.	0.0475

Table 26.: Summary statistics: Price Changes (September-October 2008).

Trade direction, IBS_k	
Buyer-initiated (%)	49.90
Seller-initiated (%)	50.10
Mean	-0.0020
Std. dev.	1.0000

Table 27.: Summary statistics: Trade Direction (September-October 2008).

Statistic	This Thesis (Panel E)	Hausman et al. (1992)
Signed transformed volume		
Mean	-0.0183	0.1059
Std. dev.	5.2592	6.1474
Time between trades (seconds)		
Mean	2.0216	27.21
Std. dev.	3.7012	34.13
Bid/ask spread		
Mean	0.0800	1.9470
Std. dev.	0.0693	1.4625

Table 28.: Summary statistics: Other Variables (September-October 2008).

Parameter	Value	Thresholds	Value	Statistics	Value
β_1: Z_{-1}	-0.0022***	α_1	-0.7307***	AIC	2054380.9622
	(0.0001)		(0.0030)		
β_2: Z_{-2}	-0.0004*	α_2	-0.7000***	Log Likelihood	-1027167.4811
	(0.0001)		(0.0029)		
β_3: Z_{-3}	0.0000	α_3	-0.5907***	Num. obs.	508538
	(0.0001)		(0.0025)		
β_4: IBS_{-1}	-0.0946***	α_4	-0.4745***	Iterations	9
	(0.0045)		(0.0021)		
β_5: IBS_{-2}	0.0430***	α_5	-0.4353***	McFadden's R^2	0.0252
	(0.0045)		(0.0019)		
β_6: IBS_{-3}	0.0275***	α_6	-0.2742***		
	(0.0045)		(0.0014)		
β_7: $lnV_{-1} \cdot IBS_{-1}$	0.0119***	α_7	-0.2184***		
	(0.0008)		(0.0012)		
β_8: $lnV_{-2} \cdot IBS_{-2}$	-0.0033***	α_8	0.2740***		
	(0.0009)		(0.0014)		
β_9: $lnV_{-3} \cdot IBS_{-3}$	-0.0029***	α_9	0.4320***		
	(0.0009)		(0.0019)		
γ: $\Delta t/100$	0.3073***	α_{10}	0.4713***		
	(0.0014)		(0.0021)		
		α_{11}	0.5880***		
			(0.0025)		
		α_{12}	0.6984***		
			(0.0029)		
		α_{13}	0.7293***		
			(0.0030)		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 29.: Ordered Probit Model Estimation (September-October 2008).

B. Additional Algorithm

Algorithm 1 Trade Direction Algorithm (Lee and Ready (1991))

Data : Trade price (P_t), Midquote (M_t), Previous trade price (P_{t-1}), Previous IBS (IBS_{t-1})

Result : IBS (Initiated Buy-Sell indicator)

if $P_t > M_t$ **then**
    $IBS_t \leftarrow 1$; `// Buyer initiated`
end
else if $P_t < M_t$ **then**
    $IBS_t \leftarrow -1$; `// Seller initiated`
end
else
    ; `//` $P_t = M_t$ `- apply tick test`
    if $M_t > P_{t-1}$ **then**
        $IBS_t \leftarrow 1$; `// Buyer initiated`
    end
    else if $M_t < P_{t-1}$ **then**
        $IBS_t \leftarrow -1$; `// Seller initiated`
    end
    else
        ; `// No price change between consecutive trades`
        $IBS_t \leftarrow IBS_{t-1}$; `// Use previous classification`
    end
end
return IBS_t

Declaration of Authorship

I certify that the thesis at hand was made without unauthorized help and that I only used the tools denoted. All statements literally or logically taken from publications are marked as quotes.

Place, Date

Ha Thu Tran