Measuring and Modelling Patterns of Behaviour in Datasets of Individual Investor Trading Records Using Flexible Methods

by

Matthew Burgess

A thesis submitted to the University of Warwick for the degree of

Doctor of Philosophy

Department of Statistics

September 2017

Contents

Acronyms **xiii**

Abstract **xv**

Acknowledgements **xvi**

Declaration **xvii**

Overview **xviii**

1 Defining, measuring and decomposing the disposition effect **1**

1.1 Introduction 1

1.2 Data 3

1.3 Past approaches to measuring the disposition effect 5

1.3.1 Aggregate measures 5

1.3.2 Regression models 7

1.4 Modified Odean-ratio 11

1.5 Decomposing the disposition effect 13

1.5.1 The disposition effect as a time series 15

1.5.2 Explaining the fall in 1995/6 17

1.5.3 The disposition effect and market capitalization . . . 22

1.5.4 The influence of individual stocks on the aggregate disposition effect 27

1.5.5 Summary 35
1.6 The Cox model in detail 36
1.7 Measuring the disposition effect in the LDB dataset using a Cox model 39
1.7.1 Sample formation 40
1.7.2 Model results 41
1.7.3 Testing the proportional hazards assumption 42
1.8 Connection with count-based measures of the disposition effect 46

2 Using a frailty model to measure the effect of covariates on the disposition effect 49
2.1 Introduction 49
2.2 The problem of correlation 51
2.2.1 Marginal model 52
2.2.2 Frailty model 53
2.2.3 Comparison 55
2.3 Covariates 57
2.4 Summary statistics 62
2.5 Comparison of marginal and frailty models 63
2.5.1 Frailty model results 70
2.6 Robustness checks and supplementary models 79
2.6.1 Model without demographic information 79
2.6.2 Model without demographic variables and larger sample 80
2.6.3 Recurrent positions 82
2.7 Summary and discussion 85
2.7.1 Model results 86

3 Using a mixed-effects logistic model to analyse the determinants of lottery stock purchases 92
3.1 Introduction 92
3.2 Stocks as lotteries 94

3.2.1 Empirical definition 96
3.2.2 Past findings 97
3.3 Modelling approach 99
3.3.1 Investor-level correlation 100
3.3.2 Mixed-effects logistic regression 101
3.4 Covariates 102
3.4.1 Time-varying covariates 103
3.4.2 Static covariates 106
3.5 Model estimation results 110
3.5.1 Testing significance of the random effect component . 110
3.5.2 Residuals 110
3.5.3 Interpretation of full model 116
3.6 Supplementary models 123
3.6.1 Is the recent performance of lottery stocks less important for some groups of investors? 123
3.6.2 Equivalent model for non-lottery stocks 124
3.6.3 The importance of repurchases 128
3.6.4 Without demographic covariates and a larger sample 130
3.7 Summary 132

Appendix **134**

List of Figures

1.1 A monthly time series of the aggregate disposition effect, across all investors and stocks. The component parts are split according to the month they occurred in, and a disposition effect score calculated according to the formula described in Section 1.4. The data run from January 1991 to November 1996. The first few months are very volatile due to a low number of stocks being under observation, hence the figure is started in April 1991. 16
1.2 A monthly time series of the proportion of gains realized (PGR) and the proportion of losses realized (PLR). Or alternatively, the proportion of opportunities to sell for a gain or loss that investors take in aggregate. Again the plot is started in April 1991 due to the low amount of data available to calculate the ratios with at the start of the period. 18
1.3 The number of sales for a gain and the number of sales for a loss each month during the data period, across all investors and stocks. 19
1.4 The number of days each month, summed across all investors and stocks, where a stock position was trading at a gain and was not sold, and the number where a stock position was trading at a loss and was not sold. 20

1.5 Monthly time series of the disposition effect for stocks in the top 20% of the cap-size distribution (updated annually) and separately for all other stocks. 24
1.6 The number of sales for a gain and a loss each month across all stocks in the bottom 80% of the cap-size distribution. . . 25
1.7 The number of gain and loss days each month for stocks in the bottom 80% of the cap-size distribution. 26
1.8 Plots of the bottom (left) and top (right) 10% of DE influence scores across all stocks. 29
1.9 A plot of a smoothed curve fitted to the scaled Schoenfeld residuals for the paper gain indicator, with a dashed line at the level of the coefficient estimate β. 45

2.1 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the 'Limited' group for self-assessed experience, in the frailty and marginal models. Each curve is plotted with an approximate 95% confidence interval. A horizontal dashed line is plotted at the coefficient estimate for the variable in the respective model. 67
2.2 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and age, in the final frailty model and an equivalent marginal model. Each curve is plotted with an approximate 95% confidence interval. A horizontal dashed line is plotted at the coefficient estimate for the variable in the respective model. 67

2.3 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the gender indicator (equalling one if the investor is male) in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate. 71
2.4 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the 'limited' experience category in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate. . 73
2.5 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the third diversification group (holding between 7 and 11 stocks) in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate. 74
2.6 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the December indicator in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate. 76
2.7 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the first capitalization size quintile, containing stocks with the smallest cap-sizes. 77
2.8 Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the second capitalization size quintile. 78

3.1 Monthly time series of aggregate preference for each of the three stock categories: lottery, non-lottery and other. The aggregate preference for a category is the percentage of the total value of all purchases made during the month that is due purchases of stocks in that category. 106
3.2 Cubic smoothing spline fits for the response variable (1 if the buy was of a lottery stock and 0 otherwise) plotted on the logit scale against covariate values. The regression model assumes a linear relationship on this scale, hence the plots for the transformed variables on the right demonstrate closer adherence to this assumption. 109
3.3 QQ plot comparing a realization of the RQRs for the mixed-effects logistic model with the standard normal distribution. The RQRs have been converted into quantiles of the standard normal distribution by applying the inverse CDF of this distribution to them. 112
3.4 Four realizations of the RQRs for the mixed-effects logistic model, plotted against the fitted values of the model. A smoothing spline has been added to each plot to help identify deviations from the expected value of 0.5. 114
3.5 Four realizations of the RQRs for the logistic model without random effects, plotted against the fitted values of the model. A smoothing spline has been added to each plot to help identify deviations from the expected value of 0.5. 115

List of Tables

1.1 Disposition effect score for positions in stocks which are in each capitalization size quintile. Quintile 1 contains stocks with the lowest capitalization sizes and quintile 5 those with the highest. 23

1.2 Percentage of stock-year observations belonging to each DE influence group where the stock was in a certain cap-size quintile, for each of the five quintiles. Note that the columns sum to 100 (with some allowance for rounding), but the rows do not. This is because the quintiles are defined based on all stocks in the CRSP database and hence the distribution across quintiles of stocks traded in this dataset does not need to be even. 31

1.3 Percentage of stock-year observations belonging to each DE influence group where the stock was in a certain volatility quintile, for each of the five quintiles. Note that the columns sum to 100 (with some allowance for rounding), but the rows do not. This is because the quintiles are defined based on all stocks in the CRSP database and hence the distribution across quintiles of stocks traded in this dataset does not need to be even. 32

1.4 The percentage of pairs of consecutive years for all stocks (49,060 in total), where the stock was in two particular DE influence groups across the two years, for all combinations of influence groups. 33
1.5 The mean change in cap-size and volatility for stocks in two particular DE influence groups across two years, for pairs of consecutive years across all stocks. 34
1.6 Parameter estimates resulting from fitting a Cox model to positions in the LDB dataset, reported as hazard ratios (HR), and p-values for Wald tests of the estimates being different from zero (or equivalently for HRs, being different from 1). . 42

2.1 Summary statistics for the continuous covariates. Calculated across investors, rather than positions, and only including investors with at least one position in the sample that will be used for model estimation. Diversification is the number of stocks the investor had in their portfolio in the first month for which there is a record of their positions. 62
2.2 Distribution of categorical variables across investors and positions. The ratio is the percentage of positions divided by the percentage of investors. Income has been split into quartiles, with quartile 1 containing those with the lowest incomes. Experience is self-reported by the investor. 63
2.3 Hazard ratios for the interaction terms that were significant at the 5% level in both the marginal and frailty models. . . . 64
2.4 Hazard ratios for the interaction terms that were significant at the 5% level in the frailty model but not the marginal model. 65
2.5 Statistics comparing the overall adequacy of the marginal and frailty models. 65
2.6 Hazard ratios, standard errors and p-values for the interaction terms in the frailty model with all interaction terms included. 69

2.7 Hazard ratios for interaction terms in a frailty model that does not include any demographic information and estimated using a much larger sample (364,000 positions compared to 85,000), denoted by (L). Compared with hazard ratios for the same interactions as estimated in the full frailty model from section 2.5.1, which used the smaller sample, denoted by (S). 81
2.8 Hazard ratios and p-values for interaction terms in the model estimated with: all data (1), first positions only (2), and repurchased positions only (3). First positions occur the first time an investor purchases a particular stock during the data period and all other positions repurchased positions. 84

3.1 Percentage of stock purchases where the investor making the purchase had sold at least one lottery stock in the year prior to the purchase, and the percentage where the investor held at least one lottery stock at the time of the purchase. 104
3.2 Summary statistics for the time-varying retuns covariates, calculated across purchases. This includes the value-weighted mean return of the investor's lottery sales in the past year, conditional on at least one such sale having been made; the value-weighted mean return of the investor's currently held lottery stock positions, conditional on them having at least one such position; and the value-weighted mean return of all lottery stocks in the market in the previous calendar month 104

3.3 Summary statistics for the time-varying covariates containing information about the investor's behaviour, calculated across purchases. Includes the proportion of the value of purchases made so far that the investor spent on lottery stocks, the number of positions the investor currently holds and the number of actions they have taken so far (buy or sell) divided by the number of days since the start of the data period. Also describes the transformation applied to each variable before being entered into the model. 105
3.4 Distribution of static categorical covariates across investors and purchases 107
3.5 Summary statistics for static continuous covariates. Calculated across investors, rather than stock purchases. Also describes the transformation, if any, applied to the variable before being entered into the model. Initial diversification is the number of stocks the investor held at the start of the data period. 108
3.6 Results from estimating a logistic regression model with investor-level random effects for lottery stock purchases. 117
3.7 Change in odds for some example increases in the proportion of the total value of their purchases to date that an investor has spent on lottery stocks. 118
3.8 Change in odds for some example differences in age 122
3.9 Results for interactions with the variable recording the return of lottery stocks in the previous calendar month (RLPM) when added to the logistic regression model with investor-level random effects. 124
3.10 Examples of the effect on the odds of a lottery stock purchase for increases in the return of lottery stocks in the market in the previous calendar month (denoted in percentage points) for investors who currently hold 1, 5, and 10 positions. . . . 125

3.11 Results from estimating a logistic regression model with investor-level random effects for non-lottery stock purchases. 126
3.12 Results from estimating a logistic regression model with investor-level random effects for lottery stock purchases. Repurchases were removed, so the sample only contained the first purchase an investor made of any particular stock. 129
3.13 Results from estimating a logistic regression model with investor-level random effects for lottery stock purchases. No demographic variables were included, and as a result a much larger sample was used. 131
3.14 Correlation coefficient between the scaled Schoenfeld residuals and survival times for each interaction term in the marginal and frailty models. Reported with the χ^2 test statistics for whether this correlation is non-zero and associated p-values. 135
3.15 Hazard ratios, standard errors and p-values for the main effects in the frailty model with all interaction terms included. 136
3.16 Hazard ratios, robust standard errors and Wald test p-values for the main effects in the marginal model. Standard errors are robust to correlation at the investor level. 137
3.17 Hazard ratios, robust standard errors and Wald test p-values for the interactions in the marginal model. Standard errors are robust to correlation at the investor level. 138

Acronyms

AIC: Akaike information criterion

CDF: Cumulative distribution function

CPT: Cumulative prospect theory

CRSP: The Centre for Research in Security Prices

DE: Disposition effect

D&Z: Dhar and Zhu (2006)

F&S: Feng and Seasholes (2005)

GLM: Generalized linear model

GLMM: Generalized linear mixed model

HR: Hazard ratio

IQR: Interquartile range

LDB: Large discount brokerage

LRT: Likelihood-ratio test

PGR: Proportion of gains realized

PH: Proportional hazards

PLR: Proportion of losses realized

QQ: Quantile-quantile

RQR: Randomized quantile residuals

SIC: Standard industrial classification

SSR: Scaled Schoenfeld residuals

Abstract

Datasets of individual investor trading records have been an important source of empirical evidence in the field of behavioural finance. This thesis contributes to two topics within this empirical literature using a dataset of trading records from a discount brokerage. The first topic is the disposition effect (DE), the tendency for investors to sell winning positions at a faster rate than losing positions. A version of the aggregate DE score introduced by Odean (1998) is analysed, as a time series and at the level of individual stocks. The influence of each stock on the aggregate DE score is calculated, and the characteristics of high and low influence stocks compared. A formal relationship is derived between this DE score and the hazard ratio estimated in a proportional hazards (PH) model.

PH models have been used in the literature to measure the effect of covariates on the DE at the investor level. Past approaches have used a marginal model to address the problem of correlation between positions at the investor-level, which involves computing robust standard errors after estimation of the model. A shared frailty model is tested as a more flexible alternative, where unobserved heterogeneity is modelled through the use of latent variables. It provides a significantly improved fit relative to the corresponding marginal model, and adheres more closely to the PH assumption.

The second topic is the preference of investors for lottery stocks. These are stocks that are low in price and high in volatility and skewness, a scheme of stock categorisation suggested by Kumar (2009a). The theme of using more flexible models to accommodate investor-level correlation is continued, with a mixed-effects logistic regression being used to study the factors affecting the decision to purchase a lottery stock. This allows the comparison of both time-varying and static factors.

Acknowledgements

I would like to thank my supervisor, Dr. Julia Brettschneider, for her support and guidance during the completion of this thesis and the work it describes. In particular I am grateful for the generosity with which she has shared her time and wisdom, and her endless enthusiasm for the topic. Thanks are also due to the members of my review panel, Drs. Ben Graham, Vicky Henderson and Dario Spano for their helpful feedback throughout the process. I am grateful to the EPSRC for funding, and to the department of statistics for providing me with an academic home during my studies.

I would also like to thank my family, and partner, Sophie, for the love and encouragement they have provided during my time as a student.

Declaration

This thesis contains my original work and has not been submitted for examination at any institution other than the University of Warwick.

Overview

It is now widely accepted in the fields of economics and finance that the decisions people make are often not rational in a way that is consistent with expected utility theory. This has given rise to a rich literature of new theories of decision making, with prospect theory (Kahneman and Tversky, 1979) and mental accounting (Thaler, 1985) being two prominent examples. An important inspiration for these theories has been the empirical evidence collected on decision making in a variety of settings. One such source has been datasets of financial trading records. The stock market represents a high-stakes environment that is highly structured and relatively self-contained compared to other domains in which people make important decisions. For example, the value and riskiness of an asset can be clearly defined using market data, and the timing of events is recorded with high precision.

This makes trading record datasets ideal for detecting patterns of behaviour that can provide evidence for or against theoretical models. Financial trading is also of inherent interest due to the welfare implications of systematically poor decision making. Datasets of trading records are not widely available due to the sensitive nature of the information they contain. But a number of large datasets of this kind have been studied in the literature, including the LDB dataset used in this thesis.[1] This empirical literature has provided some of the best evidence on decision making in general. These

[1] See section 1.2 for details.

empirical studies and the rigorous theoretical models of financial decision making together comprise the behavioural finance literature. An overview of this literature as it relates in particular to the decision making of individual investors can be found in Barber and Odean (2011).

This thesis contributes to two topics on the empirical side of this literature. The first is the disposition effect (DE), which is the tendency for investors to sell stocks that have increased in price since purchase at a faster rate than stocks that have decreased in price. One approach to measuring the DE has been the use of an aggregate score first suggested by Odean (1998), which compares the number of sales for a gain or loss to the number of total opportunities to make such a sale that occurred. A modified version of the Odean score is analysed in chapter 1. This includes decomposing it into a time series and to the level of individual stocks. The influence of each stock on the aggregate DE score can then be quantified, and the characteristics of high and low influence stocks compared. This analysis aids the understanding of how an aggregate DE, as measured by the Odean score, arises in practice.

More recent approaches to measuring the DE have used proportional hazards (PH) regression models, a class of models from survival analysis that estimate the rate at which positions are sold directly. The Cox PH model (Cox et al., 1972) is introduced in chapter 1 and demonstrated using the LDB dataset. Particular attention is paid to checking the PH assumption, a step that is often neglected in the existing literature. A formal relationship is derived between the hazard ratio for gains relative to losses, as estimated by the Cox model, and the Odean DE score. This has been missing from the literature thus far, and helps connect the DE with survival analysis methods more closely.

A key aim of the DE literature has been to determine which factors affect the severity of the DE at the investor level. PH models have been a popular choice for this task. An important feature of trading record datasets is the natural grouping of positions based on which investor it was that held them. Correlation between positions held by the same investor can be problematic in a model where one investor can contribute multiple positions. For the LDB dataset, this problem is made worse by the fact that the distribution of positions across investors is extremely imbalanced: many investors hold only a few positions during the data period whereas some hold many hundreds. Past approaches have dealt with this correlation by using a marginal model, which entails computing robust standard errors after the model has been estimated. Chapter 2 explores the use of a shared frailty model as a more flexible alternative, where unobserved heterogeneity between investors is modelled explicitly through the use of latent variables, called frailties. Compared to a corresponding marginal model, the frailty model provides a significantly improved fit to the LDB data. It is also shown to adhere much more closely to the proportional hazards assumption.

Chapter 3 continues with the theme of using flexible models to account for investor level correlation. It uses this approach to study purchases of lottery stocks. These are stocks that are low in price and high in volatility and skewness, a scheme of stock categorisation introduced by Kumar (2009a). This chapter uses logistic regression at the level of individual purchases to model the odds of a particular purchase being of a lottery stock. Since investors can make many purchases during the data period, random effects are included at the investor level to control for this source of correlation. This approach allows the inclusion of both time-varying covariates capturing information about an investor's portfolio and recent behaviour, and static information about their demographic background. The results of this analysis provide new evidence on the relative importance of these different factors, particularly an investor's recent experience with lottery stocks and the recent performance of lottery stocks as a group in the market.

Chapter 1

Defining, measuring and decomposing the disposition effect

1.1 Introduction

The disposition effect (DE) is the tendency for investors to sell gains (assets that have increased in price since purchase) and hold losses (assets that have decreased in price since purchase). More quantitative definitions refer to the rate at which positions are sold, with the DE corresponding to the situation where gains are sold at a greater rate than losses. The DE is the most consistently observed behavioural pattern in datasets of investor trading records, having been documented in a number of different financial markets and countries[1]. It has received a great deal of attention since, as an investment strategy, it contradicts the standard advice to cut losses and let winners run, and is sub-optimal in terms of minimizing capital gains

[1] See Barber and Odean (2011) for a comprehensive list

tax.[2]

There is also evidence that investors earn lower returns as a result of the DE due to the momentum present in stock prices. This momentum is documented in the U.S. by Jegadeesh and Titman (1993), who find that a momentum trading strategy generates significant positive returns for holding periods up to one year. In a dataset of investor trading records, Odean (1998) finds that gains that are sold outperform losses that remain unsold by an average of 3.4 percentage points over the year following the sale. Together with the consideration of tax, this suggests that investors would improve their performance if they did not exhibit the DE.

A variety of different methods have been used to measure the DE in the empirical literature. This chapter will describe the different approaches and examine two in detail. The first is a modification of the score introduced by Odean (1998), based on finding the ratio of sales for a gain or a loss to the number of opportunities investors had to make such a sale. This score is analysed using the LDB dataset of individual investor trading records, which is used throughout this thesis and introduced in the next section. For exploratory purposes, and to understand how an aggregate DE arises in practice, the score is decomposed both into its component parts and calculated for subsets of the data. This includes calculating it monthly to produce a time series, which leads to an investigation into why the aggregate DE score falls in 1995/96. The score is also calculated for individual stocks, which allows the contribution of each stock to the aggregate DE score to be quantified. This leads to a notion of influence analogous to the DFBETA quantity used to measure influence in regression analysis. The characteristics of stocks that have high influence on the aggregate DE score are studied, along with the changes in influence over time.

The second method to be examined is the Cox proportional hazard regression

[2] This is discussed in Shefrin and Statman (1985), with reference to the tax-optimal behaviour derived in Constantinides (1984).

model, a method from survival analysis. This method has become popular in the study of the DE, and in particular for measuring the effect of covariates on the DE at the investor level. The method is introduced and demonstrated using the LDB dataset. Particular focus is given to testing the important proportional hazards assumption, a step that is typically neglected in the existing literature. Finally, the relationship between the modified Odean measure of the DE and the hazard ratio produced by a Cox model is formally derived. It is shown that the Odean score is proportional to a non-parametric estimator for the hazard ratio, and a simple condition is given for when the latter will exceed the former in magnitude.

The remainder of the chapter is organised as follows. Section 1.2 introduces the datasets that will be used throughout this thesis. Section 1.3 discusses past approaches to measuring the DE. Section 1.4 defines the modified Odean score that will be used subsequently. Section 1.5 presents analysis resulting from decomposing the aggregate DE score. Section 1.6 introduces the Cox model and section 1.7 demonstrates its use with the LDB dataset. Section 1.8 derives a relationship between the modified Odean score and a non-parametric estimator for the hazard ratio that is produced by the Cox model.

1.2 Data

The main dataset used in this thesis consists of the trading records and a variety of demographic information for 78,000 investors at a large discount brokerage firm in the U.S. from the beginning of January 1991 until the end of November 1996. The data were obtained by Odean from a large discount brokerage and are commonly referred to in the literature as the LDB dataset.[3] A detailed description of the dataset can be found in Barber

[3] We are grateful to Terrance Odean for sharing this dataset with us.

and Odean (2000). The LDB dataset is notable for the large number of investors it contains, the relatively long period of time it covers and the range of demographic information about the investors that was collected alongside the trading records. Several studies that are important references for this thesis also used the LDB dataset in their respective analyses, a fact that will be pointed out when these references are introduced.

As reported in Goetzmann and Kumar (2008), 62,387 investors in the dataset trade common stocks, which is the only asset type considered here. The median investor holds a portfolio consisting of three stocks, with a total value of $13,869. Barber and Odean (2000) find that the mean investor in the dataset turns over 75% of their stock portfolio each year and underperforms the market by 1.5% annually. A defining feature of the LDB dataset is the large degree of heterogeneity that is present in essentially all aspects of investor behaviour. This includes total portfolio value, level of diversification, trading frequency and stock-type preference. In models where one investor can contribute multiple observations to the sample, controlling for this heterogeneity and the correlation between groups of observations that it causes is an important part of the modelling process. Investor-level correlation was a key motivation behind the choice of the models used in chapters 2 and 3, and the topic will be discussed in more detail there.

Data on stock prices, SIC industry and capitalization were obtained from the CRSP, and the analysis is limited to stocks that have price information in the CRSP database. These prices, as well as those in the LDB files were corrected for splits and dividends using the information provided by the CRSP. Multiple buys and sells of the same stock on the same day by an investor were aggregated, a standard processing step.

1.3 Past approaches to measuring the disposition effect

1.3.1 Aggregate measures

Holding-period approaches

The DE was first theorised by Shefrin and Statman (1985), and the authors provided some empirical support for its existence. They analyse a dataset containing the trading records for 2,506 individual investors in the U.S. during the period 1964-70, originally studied in Schlarbaum et al. (1978). The authors find that 60% of round-trip trades, where both the purchase and sale are observed during the data period, end in a sale for a gain. This percentage remains the same for different lengths of holding period. The authors cite this as evidence for the DE, as sales for a loss should be more common at short holding periods if investors are minimising their capital gains tax burden.

The holding period of round-trip trades is also used by Shapira and Venezia (2001) in their study of the trading records of 4,330 individual investors in Israel during 1994. They find that the average holding period of stocks sold for a loss is 63 days, compared to 20 days for stocks sold for a gain. For investors receiving professional advice the gap was somewhat smaller, with an average holding period of 55 days for losses and 25 days for gains. Importantly there was no capital gains tax in Israel at the time, so their analysis is not complicated by the issue of tax.

Odean count-based approach

As pointed out in Odean (1998), the problem with measures based on holding periods is that they do not consider the number of opportunities an investor had to sell for a gain or a loss before they eventually did so. This can lead to a DE not being detected in a situation where we would expect it to be based on the qualitative definition of the DE. As a simple example, suppose an investor buys stocks A and B on the same day. On the next day, both stocks are trading at a loss and the investor sells stock A, and on the next day stock B is trading at a gain and the investor sells it. So 1/1 opportunities were taken to sell for a gain, compared to 1/2 for losses. The holding period for losses was shorter than for gains which would indicate no DE, yet the investor exhibited a greater eagerness to realize gains.

Instead, Odean (1998) proposes calculating the ratio of the number of sales for a gain or loss relative to the number of opportunities an investor had to do so. An opportunity to sell for a gain (loss) is defined as a day where the daily low and high prices for the stock being held are above (below) the price the stock was purchased at. Days when an investor had the opportunity to sell for a gain/loss are labelled as 'realised' gains/losses, and days where they had the opportunity but did not take it are labelled as 'paper' gains/losses. Whether a stock position is currently trading at a gain or a loss will be referred to as the paper status of the position. Importantly, under the specification in Odean (1998), paper gains and losses are only counted on days when the investor sells at least one stock in their portfolio. This choice will be discussed in section 1.4.

Using this information two ratios are computed, the proportion of gains realized (PGR) and the proportion of losses realized (PLR) with formulas given by

$$\text{PGR} = \frac{\text{Realized Gains}}{\text{Realized Gains} + \text{Paper Gains}} \tag{1.1}$$

$$\text{PLR} = \frac{\text{Realized Losses}}{\text{Realized Losses} + \text{Paper Losses}} \tag{1.2}$$

with PGR > PLR indicating that there is a DE. The ratio PGR/PLR is commonly used to summarise the DE in a single value, with PGR/PLR > 1 indicating a DE is present. These ratios are computed in Odean (1998) using the trading records of 10,000 individual investors in the U.S. during the period 1987-93[4]. He reports a PGR of 0.148 and a PLR of 0.098, which gives a ratio of PGR/PLR of 1.51. Importantly, the DE is reversed, i.e. PGR/PLR is < 1 if only days in December are counted. This supports the hypothesis that investors engage in tax-loss selling in December, the last month in which it is possible to do so before the end of the tax year, in order to reduce their tax burden.

1.3.2 Regression models

The work of Odean provided the first robust evidence for the existence of the disposition effect in a large dataset of individual investor trading records. The next step pursued by several authors was to test theories about factors that may effect the extent of the DE, such as the characteristics of individual investors. The main approaches will be detailed below.

Dhar and Zhu (2006) calculate the DE individually for 14,872 investors using their trading records from the period 1991-96. This is the same dataset as used for the analysis in this thesis.[5] The authors then regress these DE scores on a range of variables including the investor's age, the number of trades they made during the data period, an indicator for whether they work in a professional occupation and categories for high and low income. They find that older, wealthier investors who work in a professional occupation and trade more frequently have a significantly reduced DE. Their results will be compared to those of the analysis conducted in chapter 2.

[4]This dataset comes from the same discount brokerage as the LDB dataset.

[5]The dataset contains trading records for 78,000 investors, but only 14,872 make at least six trades (a condition for inclusion made by the authors) and have a sufficient number of observed round-trip trades for a DE to be calculated.

As the authors note, DE scores at the individual level are highly dependent on the number of trades an investor makes, and extreme values for the score are not uncommon. These usually occur as a result of either the PGR or PLR being close to or exactly zero. In addition to this, many investors have to be excluded from the analysis because they do not have a sufficient number of trades for a DE score to be computed. An alternative approach is to model position lifetimes, and test whether the paper status of a position affects its chances of being sold. This avoids having to calculate the DE at the level of individual investors, and instead the effect of the paper status on an investor's decision to sell can be estimated using information from all investors. Two types of model have been used for this purpose: logistic regression and proportional hazards (PH) regression. Examples of both will be described before the key differences between them are highlighted.

Grinblatt and Keloharju (2001) use logistic regression to examine how past returns and price patterns of stocks affect the probability of a position being sold in a dataset from Finland. The dataset contains the trades made by all Finnish investors, both individual and institutional, during the period from the end of December 1994 to the beginning of January 1997. Similarly to Odean (1998), for each investor, the authors only record observations on days when the investor made at least one sale. On each such day, for all positions in the investor's portfolio, the value of a sell indicator variable is recorded plus the values of the accompanying covariates at that point in time.

The sell indicator equals one if at least some of the investor's position in the stock is sold on that day, and zero otherwise. The covariates may reflect characteristics of the particular stock or the investor, and can be fixed over the whole period (e.g. the category of the investor) or change over time (e.g. the past return of the stock over a particular horizon). The value of the sell indicator and the accompanying covariate values constitute one observation in the subsequent regression. They find that recent negative returns are associated with a decreased probability of selling and recent positive returns are associated with an increased probability of selling, consistent with the presence of a DE.

Feng and Seasholes (2005) use a parametric proportional hazards model to investigate whether investor sophistication and experience reduce the disposition effect (see section 1.6 for some detail on proportional hazards models). They study the trading records of 1,511 accounts held at a large brokerage in China from January 1999 to December 2000. Unlike in Odean (1998) and Grinblatt and Keloharju (2001), the status of the stocks in an investors portfolio is recorded on every day that they are held. Specifically, for each day a stock is held, the value of a sale indicator is recorded (equal to one if the stock is sold on that day) and the value of the accompanying covariates is recorded. This is done for every position an investor holds during the data period.[6]

The key covariates are indicators for whether a stock is trading at a gain or a loss on a certain day. The trading gain indicator (TGI) is equal to one if the daily low price is above the purchase price, and the trading loss indicator (TLI) is equal to one if the daily high price is below the purchase price. The authors estimate separate models for each of these two covariates, and find a reduced rate of selling when a stock is trading at a loss and a greatly increased rate when trading at a gain. The authors also propose measures of investor sophistication and experience, and test whether they have an effect on the strength of the DE in their data. These results will again be compared to the similar analysis which is conducted in chapter 2.

The main difference between logistic and PH regression models is that in the logistic model there is no distinction between observations that occur at different times during the holding period of a position. All intervals (single

[6]The authors define the holding period of a position as starting when the investor first purchases a stock (which they do not already hold) and ending when they have sold the position in its entirety.

days in this case) are effectively treated as having happened simultaneously. In a PH model, positions that have been sold are compared only to other positions that were at risk of being sold at that time.[7] This is important because in a PH model there is a component, called the baseline hazard function, that captures the risk of being sold that is common across all positions, analogous to an intercept term in a linear regression model. However the baseline hazard function can be dependent on time, allowing the risk of a position being sold to be different depending on how long the investor has already held it. In contrast, the logistic model implicitly assumes there is no such difference. If it does change over time then a PH model will better reflect the situation in reality.

PH models separate out this baseline hazard from the effect of the covariates, and assume only that the covariates have a multiplicative effect on the baseline hazard that is constant over time. The PH model that is used in this thesis will be described in detail in section 1.6. Whilst a PH model will be used for formal testing of the effect covariates have on the DE in chapter 2, a count-based measure of the DE, similar to that of Odean, will be used in the present chapter to complement the regression analysis. The count-based measure will be useful for detecting aggregate patterns that can then be tested formally using the regression model. Decomposing the measure into its component parts, and into a time series will also add to the understanding of how aggregate measures of this kind work in datasets of trading records. The next section will provide a definition of the count-based measure of the DE that will be used in subsequent sections.

[7] Note that 'time' refers here to the time since a position was purchased. Hence sold stocks are compared to those that had been held for the same number of days, but not necessarily at the same calendar time. This distinction is discussed in section 1.6.

1.4 Modified Odean-ratio

In their examination of the aggregate DE in the LDB dataset, the following sections will use the PGR and PLR ratios introduced in Odean (1998), with two important differences. Odean's method only counted paper gains and losses on days when the investor sold at least one other position in their portfolio. Here, all days during the holding period of a position on which it is not sold will be counted.[8] The criteria for a position representing either a paper gain or a loss are the same: it is a paper gain if the daily low price is above the purchase price, and a paper loss if the daily high price is below the purchase price. Hence there will be days during the holding period where the position is neither a paper gain nor loss, and it will not contribute to any of the DE components i.e. PGR or PLR.

There are arguments on both sides for which method best reflects the conditions under which investors were making their decisions at the time. On a day when an investor sells a position, it is reasonable to assume they also checked the current price of their other positions. Hence the presence of a sale on that day provides the strongest signal that they considered selling a position but decided not to. However, there are also likely to be days when an investor checks the price of stocks they hold but takes no action. These days would not be counted when using Odean's method, but would by the method used here.

It is not clear which total will best reflect the number of days when an investor was aware of the price of a stock they held and chose not to sell. During the 1990s, the time in which the LDB data was collected, stock prices were available in print, by telephone, on TV channels and later, via the Internet[9]. So investors who wanted to know the current status of their

[8] To be precise, 'all days' means U.S. business days i.e. days on which it was possible to trade stocks.

[9] The majority of actual trading in this dataset was conducted via telephone, but with a significant shift towards trading online towards the end of the data period. See Barber

positions would easily have been able to do so. Counting all days during the holding period also follows the general principle in statistics of using as much information as possible in any analysis that is being done.

The second difference is that, unlike in Odean's method, sales of a position that occur when an investor holds no other positions at the time will be included in the counts of realized gains and losses. An investor holding only one position is a common occurrence in this dataset, and making this change doubles the number of sales that are recorded in total (from ∼315,000 to ∼630,000). These two changes bring the method more in line with more recent survival analysis approaches using PH regression models, such as in Feng and Seasholes (2005) and Barber and Odean (2011). The connection with survival analysis methods will be discussed in section 1.8. In these models the units of analysis are the holding periods of stock positions. The holding period of a position is the number of days from the date of purchase until the position is sold, hence recording paper gains and losses on all days during the holding period of a position makes a count based method more comparable to results from a PH regression model.

Together these changes greatly increase the number of investors and stocks that contribute to the aggregate DE components. An investor or stock contributes to the aggregate DE when positions held by the investor, or in the stock, add to at least one of the four DE components i.e. realized gains/losses or paper gains/losses. Under Odean's original scheme, 28,450 investors make a contribution to the aggregate DE, compared to 62,473 investors in the modified scheme. Amongst stocks, 2,299 make a contribution in the original scheme and 9,812 do after the modification.

For a disposition effect score, the ratio of PGR and PLR will be used

$$\mathrm{DE} = \frac{\mathrm{PGR}}{\mathrm{PLR}} \tag{1.3}$$

and Odean (2001b, 2002) for some detail on this.

If this ratio is greater than 1 then there is a disposition effect i.e. investors take a greater proportion of their opportunities to sell for a gain than for a loss.

To compare the original Odean scheme with the modified version, this score can be computed for all stocks traded in the dataset. In both cases, for it to be possible to compute a score a stock must have at least one realised loss, otherwise PLR, the denominator of the score, would be zero.[10] This restriction means it is possible to compute both scores for only 2,031 stocks. For this set of stocks, the Pearson correlation between the two scores is 0.62 and there is agreement in 75% of cases as to whether there is a DE (score > 1) or not.

1.5 Decomposing the disposition effect

Scores of the kind discussed in the previous section have mainly been used in the literature to establish the presence of the DE, averaged across large groups of investors, stocks and also across time. Further insight can be gained however by decomposing the aggregate score. The following sections will do this by separating it into its component parts, and by calculating it for subsets of the full dataset. Doing so will reveal broad trends in the data, and thus serve as a useful exploratory step prior to the regression modelling that will be done in chapter 2. Results of the analysis here will inform the choice of covariates in the models that will subsequently be estimated. But this kind of decomposition will also help explain how an aggregate DE, as a statistical phenomenon, arises in practice.

Section 1.5.1 will consider the DE as a time series by calculating it for

[10] This problem would be avoided by using DE = PGR − PLR, as some authors do. However, using their ratio makes the result more comparable with results from other studies that have been computed under a different specification, for example the original results in Odean (1998).

each month during the data period. Whilst a tax-motivated fall in the DE in December has been found by Odean (1998) and others, changes in the aggregate DE over time have not been examined in detail. This is partly because most datasets used in the literature do not cover more than a few years, unlike the LDB dataset which covers a six year period. By decomposing the DE into its component parts and seeing how they change over time, section 1.5.2 is able to establish which part is the primary driver behind the fall in aggregate DE in '95 and '96. Rather than large changes in the number of stocks investors sell for a gain or a loss, it is the dramatic fall in the number of opportunities to sell for a loss that causes the fall in DE. This can in turn be connected with the market conditions which were in effect at the time: strong stock market performance at the start of what became the dotcom boom meant investors had far fewer opportunities to sell their positions at a loss, but likely still wanted to do so in order to off-set capital gains and reduce their tax burden.

Some work has been done on variations in magnitude of the DE between stocks. Kumar (2009b) finds that difficult to value stocks exhibit a greater disposition effect. This includes stocks with higher volatility, lower market capitalization and weaker price momentum. That positions in stocks with smaller market capitalisations exhibit a stronger DE is confirmed in section 1.5.3. Section 1.5.4 extends the work of Kumar by decomposing the DE into the parts contributed by each stock. The influence of each stock on the aggregate DE can then be calculated. Influence is highly concentrated amongst a small group of stocks, the majority of which are in the top quintile in terms of market capitalization. However, this is true for both stocks that make the aggregate DE stronger and those that make it weaker. So whilst large cap stocks do have a lower DE as a group, it is a small number of large cap stocks that make the DE stronger to the greatest degree.

By defining high and low influence groups using percentiles in the influence distribution, section 1.5.4 also shows how the impact stocks have on the aggregate DE can change over time. Examining the characteristics of stocks as they move between influence groups shows that stocks which become more DE strengthening also increased in volatility on average, whilst stocks moving in the other direction decreased.

1.5.1 The disposition effect as a time series

When the DE component parts are counted for each stock position, the date of the count observation is also recorded. Hence for any time period lasting at least one day, the component counts that occurred during that period can be summed and a DE score calculated. Due to the relevance of calendar months in financial markets, an aggregate DE score for each month during the data period will be calculated in order to construct a time series for the DE.

A monthly scale is also important due to the presence of a strong 'December effect' in data of this kind. As a result of the time of the deadline for recording capital gains and losses in the U.S. tax system, the number of realized losses increases dramatically in December. In a similar dataset to the one studied here, Odean (1998) finds that the DE is actually reversed in December, with the proportion of losses realized being greater than the proportion of gains realized. Studying a monthly time series of DE scores will highlight the change in investor behaviour during December, and also allow comparison between Decembers in different years.

Since the data start in January '91 and end in November '96 there are 71 months in total for which a DE score can be calculated. At the start of the data period the number of stock positions under observation increases from zero as investors make their first recorded trades.[11] The magnitude of DE component counts is therefore much lower in the first few months of

[11]There are significantly more buys than sells in this dataset, so the number of stock positions under observation steadily increases throughout the period.

'91 relative to the rest of the data period, and as a result the corresponding ratios are sensitive to small changes in the data. To keep them informative, the time series plots in Figures 1.1 and 1.2 start in April '91 rather than January.

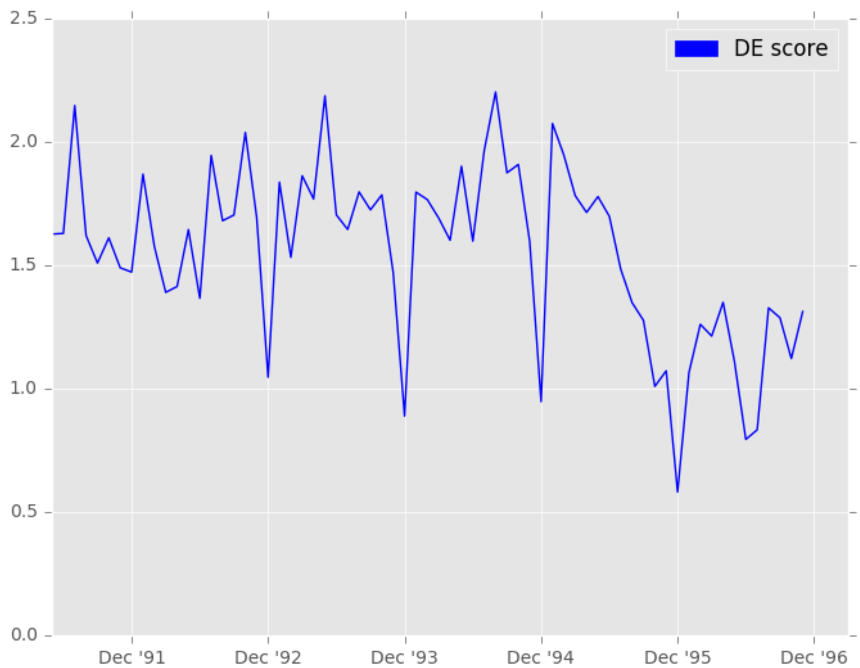

Figure 1.1: A monthly time series of the aggregate disposition effect, across all investors and stocks. The component parts are split according to the month they occurred in, and a disposition effect score calculated according to the formula described in Section 1.4. The data run from January 1991 to November 1996. The first few months are very volatile due to a low number of stocks being under observation, hence the figure is started in April 1991.

The monthly series of DE scores is shown in Figure 1.1. The most notable features of this series are the clearly evident December effect in years '92-'95, and the dramatic fall in DE during '95, which is maintained during '96. This latter feature will be discussed in depth in section 1.5.2. Whilst the December effect is visible in '92-'95 (the data period only goes as far as November in '96), the DE score does not fall in December '91. One of the results in chapter 2 is that the December effect is weaker for positions that have not been held very long. DE components are only counted for positions where the purchase was observed during the data period, hence all of the positions under observation in December '91 were relatively new at the time. This suggests the apparent lack of a December effect in that year is really an artefact of the dataset.

To understand why the DE changes over time, it can be split into its component parts with separate time series plotted for each. Figure 1.2 shows the monthly series for the proportion of gains realized (PGR) and proportion of losses realized (PLR), starting in April '91. The large values at the start of both series are mostly a result of the low number of positions under observation at the time. As this number increases, both PGR and PLR steadily fall up until part way through '94. Yet the ratio between them does not change much except in December months, hence the consistent structure in the DE score series during this period.

1.5.2 Explaining the fall in 1995/6

Figures 1.3 and 1.4 show the number of gain and loss sales per month, and number of paper gain and loss days per month respectively. Paper gain and loss days occur when a position is trading at a gain or loss relative to the purchase price, but the investor does not take the opportunity to sell the stock and realize the gain or loss. As mentioned previously, there are no positions under observation at the start of the data period, hence all the series start at zero. Since there are more buys than sells overall, the number under observation steadily grows over the course of the data period. This explains the general upwards trend visible in these figures.

Figure 1.2 showed that the fall in aggregate DE in '95 and '96 was primarily

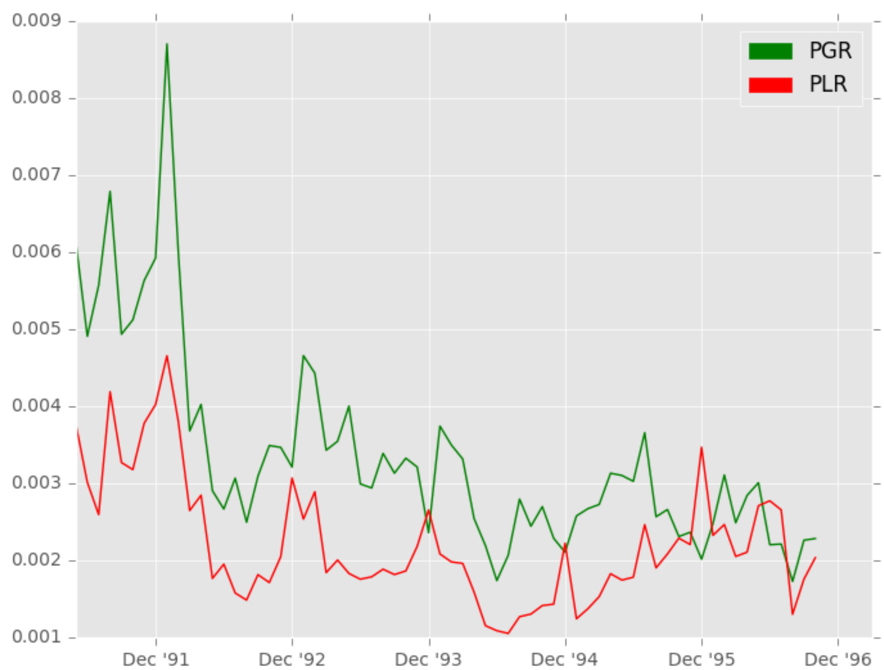

Figure 1.2: A monthly time series of the proportion of gains realized (PGR) and the proportion of losses realized (PLR). Or alternatively, the proportion of opportunities to sell for a gain or loss that investors take in aggregate. Again the plot is started in April 1991 due to the low amount of data available to calculate the ratios with at the start of the period.

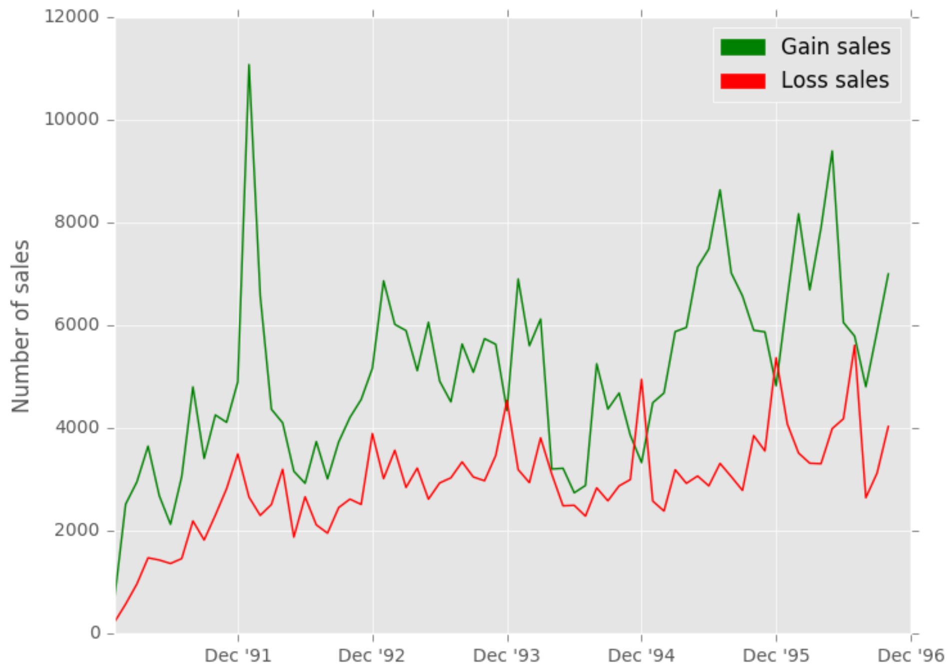

Figure 1.3: The number of sales for a gain and the number of sales for a loss each month during the data period, across all investors and stocks.

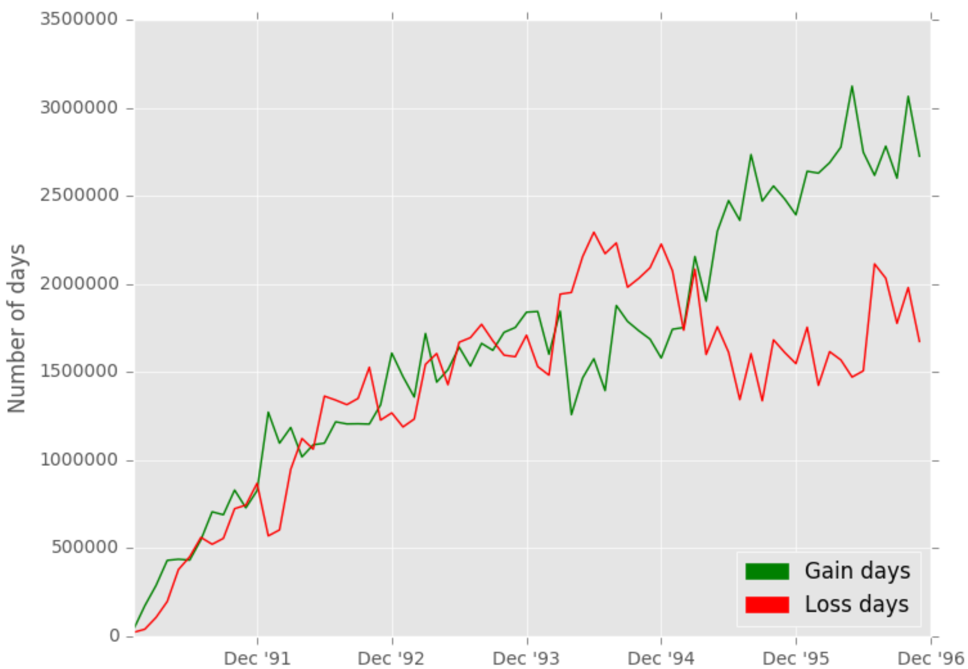

Figure 1.4: The number of days each month, summed across all investors and stocks, where a stock position was trading at a gain and was not sold, and the number where a stock position was trading at a loss and was not sold.

due to the increase in the proportion of losses realized; investors took more of their opportunities to sell for a loss during these two years. Figure 1.3 shows that the number of sales for a loss does not change significantly when PLR rises, it continues on the steady upwards trend that it had been following for most of the first four years of the data period. The increase in PLR is instead explained by the significant fall in the number of loss days, in absolute terms but particularly when compared to its upwards path over the first four years. This means that investors had far fewer opportunities to sell stocks for a loss, but still chose to sell a similar number of losses in absolute terms. Hence the proportion of losses realized increases and the DE falls.

The fall in the number of days on which investors' stock positions were trading at a loss can be explained by the strong performance of the U.S. stock market in general starting in '95. In the period 1991-93, the S&P 500 index returned 7.6%, 10.0% and 1.32% respectively each year. In '95, the index returned 37.8%, signalling the start of a stock market boom that continued until the dotcom crash in 2000. In market conditions where prices were rising rapidly across the full spectrum of stocks, investors simply did not have as many opportunities to sell positions at a loss compared to the previous few years.

As can be seen in figure 1.3, investors did increase the number of stocks they sold for a gain in response to the strong performance of their positions. Since the number of opportunities to sell for a gain had increased the proportion of gains realized ends up not changing much compared to the previous few years. But as mentioned previously, investors maintained, and in fact steadily increased, the number of stocks they sold for a loss despite the large fall in the number of opportunities they had to do so. One explanation for this is that investors still wanted to realize capital losses in order to offset the now increased amount of capital gains they were realizing, and hence minimize the capital gains tax they needed to pay.

In a different dataset, Seru et al. (2010) also find a fall in the aggregate DE over time and show that it is partly due to low ability investors deciding to cease trading. This phenomenon could also explain some of the observed decline in aggregate DE in the LDB dataset. However, in the regression model estimated in chapter 2, there is an indicator variable for whether a position is being held during the years 1995/6 or not. This indicator is highly significant in explaining variation in the DE even whilst variables capturing the experience and sophistication of the investor holding the position are also included. It is unlikely this would be the case if all of the decline in aggregate DE during these two years could be explained by changes in the composition of investors who were actively trading.

1.5.3 The disposition effect and market capitalization

An important question which will be addressed in this and subsequent sections is whether the DE affects different types of stocks to different extents. Market capitalization, or cap-size, is an important stock-level characteristic and is the primary way that stocks will be categorised in this section. Investors in the LDB dataset tilt their portfolios towards stocks with small cap-size relative to a value-weighted market portfolio, as found in Barber and Odean (2000). But a great majority of their positions are still in large cap stocks, by both count and dollar value.

Using cap-size information from the CRSP stock file, all stocks traded in the LDB dataset are assigned to a cap-size quintile. Note that the quintiles are defined using all stocks in the CRSP database and not just those traded in the LDB dataset. Cap-size information is updated annually in the CRSP database, hence the quintile a stock is in can change during the course of the data period. At any one time, roughly 70% of the total value of LDB investors' common stock positions is held in stocks that are in the largest cap-size quintile i.e. the largest 20% of stocks. This decreases to 10% for

Capitalization quintile	DE score
1	3.3
2	2.4
3	2.1
4	1.76
5	1.40

Table 1.1: Disposition effect score for positions in stocks which are in each capitalization size quintile. Quintile 1 contains stocks with the lowest capitalization sizes and quintile 5 those with the highest.

the next quintile and roughly halves thereafter for each remaining quintile. Any aggregate pattern such as the DE will therefore mostly be due to what investors are doing with their positions of large cap stocks.

A DE score can be calculated for each cap-size quintile by summing the DE components recorded for positions in stocks that are in that quintile. The results of these calculations are shown in table 1.1. There is a clear decrease in DE as cap-size increases, with the smallest stocks in terms of cap-size having a DE that is twice that of the largest cap-size stocks. This result is in agreement with those in Kumar (2009b), which show that more volatile stocks tend to have a stronger DE as a group, and volatility decreases with cap-size. These results and those of Kumar both contradict an earlier paper on the topic by Ranguelova (2001), which also uses the LDB dataset and finds the exact opposite: that the DE is stronger for stocks with larger cap-size. An Odean-type score was used there too, but with some different choices made about what exactly is counted. However, an attempt to follow the methodology set out by Ranguelova as closely as possible still produced the same result, that the DE is weaker for larger cap-size stocks. The reason for the discrepancy remains unknown.

These scores can also be broken down into monthly time series as was done in the previous section. Figure 1.5 shows a monthly series for the largest

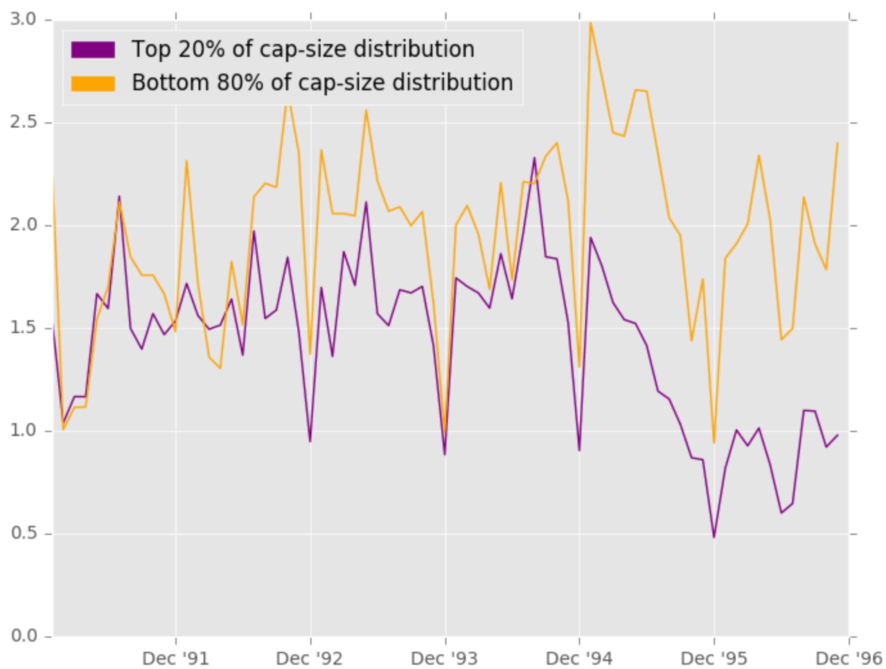

Figure 1.5: Monthly time series of the disposition effect for stocks in the top 20% of the cap-size distribution (updated annually) and separately for all other stocks.

cap-size quintile and one for the other 80% of stocks grouped together. The series for the largest cap-size quintile is essentially the same as the series for all stocks in figure 1.1. But the DE score for smaller cap stocks is consistently higher, most notably in the last two years. Although there is a fall starting in mid '95 that mirrors that of the largest cap-size quintile, it was at an unusually high level at the start of the year and rebounded to a level comparable with the rest of the series in '96, whereas the large cap series stayed low.

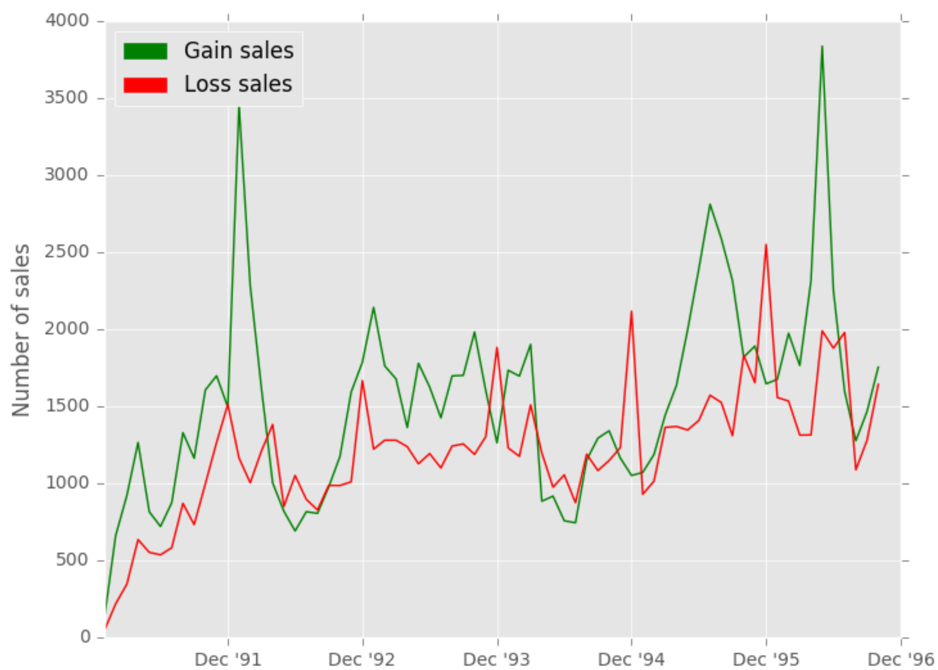

Figure 1.6: The number of sales for a gain and a loss each month across all stocks in the bottom 80% of the cap-size distribution.

As with the DE score, the series for the number of sales made for a gain or loss, and the number of gain or loss days where a stock was not sold are again very similar to the aggregates (in Figures 1.3 and 1.4) for the large cap

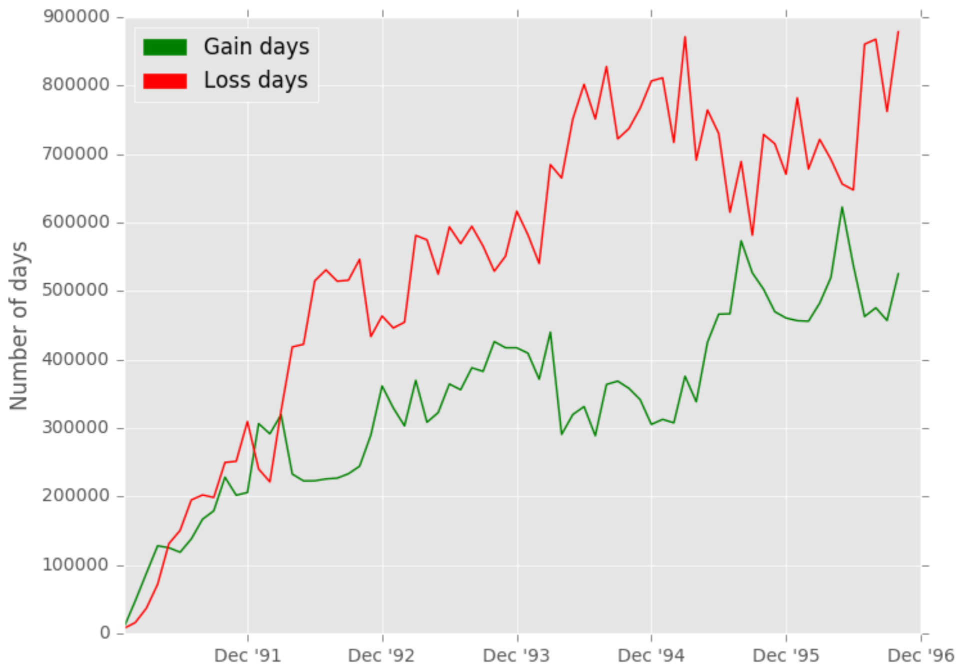

Figure 1.7: The number of gain and loss days each month for stocks in the bottom 80% of the cap-size distribution.

group. Figures 1.6 and 1.7 show the corresponding series for the smaller cap group. Breaking it down into these components reveals that the DE score for small cap stocks is generated in a different way to the large cap group. The number of sales for a gain and a loss are much closer together throughout the period, hence for there to be a DE > 1 the number of opportunities to sell for a loss has to be much larger than the number of opportunities to sell for a gain. The second figure shows that this is indeed the case; there is a fall in the number of loss days during '95, as there is in the aggregate series, but even then it remains higher than the number of gain days. This is a good example of something that would not be apparent looking only at the DE score by itself.

1.5.4 The influence of individual stocks on the aggregate disposition effect

Given that there is a DE in aggregate across all stocks, a next step is to investigate the relative importance of individual stocks in producing the aggregate effect. The component counts due to an individual stock, which will be referred to as the stock-specific components, can be summed across investors and used to quantify the effect that trades of the stock have on the aggregate DE.

Measuring influence

The statistical concept of influence can be used to do this, where the parameter estimate is calculated with and without the observation in question and the two resulting estimates are compared.[12] In this case, the aggregate DE score is recalculated after subtracting the stock-specific components from

[12]The formula used here is the same as the DFBETA measure of influence commonly used in regression analysis. It is also related to the jackknife resampling method.

the aggregate totals e.g. the number of times the stock was sold for a gain is subtracted from the total number of realized gains across all stocks. The new DE score is then subtracted from the original, giving a measure of the stock's total effect on the aggregate DE. This will be referred to as the stock's influence on the aggregate DE, or DE influence. Recalling the formula for the DE score given in 1.3, the DE influence for a stock indexed by j is defined as

$$IF_j = DE - DE_{-j}$$

where DE is the aggregate score across all stocks and DE_{-j} is the same score calculated without the component counts contributed by stock j. A positive value of DE influence means the stock makes the aggregate score larger; a negative value means it makes it smaller. Such stocks will be labeled DE strengthening and DE weakening respectively.

The distribution of influence across stocks is extremely uneven, and is essentially zero for the majority of the 9,812 stocks traded in the dataset. The most common reason for a stock to have near-zero influence is a low number of positions of that stock. As noted in section 1.5.3, investment value is highly concentrated in a small number of large cap-size stocks, with a long tail of smaller stocks that relatively few investors hold. For stocks with few positions, subtracting their component counts from the aggregates will not change the DE score much, regardless of the ratios of the stock-specific components. There will also be some stocks with a large number of positions whose component ratios closely match the aggregate ratios, and hence will have an influence close to zero.

Figure 1.8 plots curves through the ordered influence scores for stocks in the bottom and top (left and right in the figure) 10% of the empirical distribution. Note that one percentile contains approximately 98 stocks. Decay in influence is very rapid in both cases as a stock moves away from the extremes of the distribution, essentially reaching zero by the 10th and 90th percentiles. Much of the decay happens within the first and last percentiles. If stocks with influence scores in the top 1% are removed, the aggregate DE score falls by 15%, and if the bottom 1% are removed, it increases by 17%.

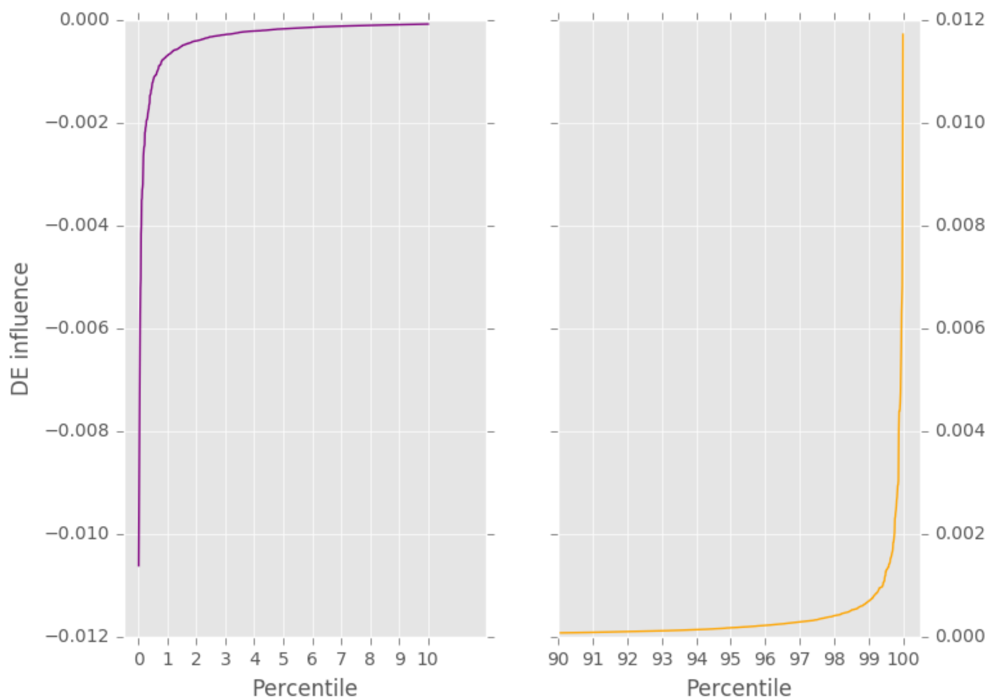

Figure 1.8: Plots of the bottom (left) and top (right) 10% of DE influence scores across all stocks.

Stocks can be naturally separated into three groups as a result of this structure. A small group at the top of the influence distribution which make the aggregate DE stronger, a large group in the middle who have near-zero influence and a small group at the bottom which make the aggregate DE weaker. To make this categorisation formally, the stocks will be split at the 5th and 95th percentile values of the distribution of influence. The resulting three groups will be referred to as DE strengthening, DE weakening and DE neutral. The decay in influence is smooth at both the top and bottom of the distribution, hence changing the cut-off percentiles does not significantly alter the rest of the analysis.

Capitalization and volatility

First the composition of each influence group in terms of market capitalization and volatility will be examined. Annual cap-size and volatility data were obtained from the CRSP stock files, and quintile groups defined for each of the two variables in the same way as before. As with cap-size, the volatily quintile a stock is assigned to is updated annually using the returns data for the previous year.

For each year, the DE influence for each stock is calculated using the stock-specific and aggregate components for that year. Stocks are then assigned to one of the three influence groups by splitting them at the 5th and 95th percentiles of the year's influence score distribution. This generates an annual sequence of DE influence groups that a stock is part of during the data period. Given that there are 9,812 stocks for which DE components can be counted, every year the DE strengthening and DE weakening groups will both contain 982 stocks, and the DE neutral group will contain 7,848.

Along with the influence group, the stock's cap-size and volatility quintiles are also recorded for each year. Across all stock-year observations which are in a particular influence group, the percentage which were in each quintile at the time can be calculated. Table 1.2 gives these percentages for each combination of influence group and cap-size quintile. The stocks with the smallest cap-sizes are in quintile 1 and the largest are in quintile 5.

In the DE weakening group, stock-year observations are heavily concentrated in the largest cap-size quintiles; over 80% are in the two largest quintiles.

Cap quintile	DE Weakening %	DE Neutral %	DE Strengthening %
1	1.68	19.98	3.95
2	3.57	21.69	9.33
3	9.87	20.96	17.23
4	20.83	20.28	25.33
5	64.04	17.09	44.16

Table 1.2: Percentage of stock-year observations belonging to each DE influence group where the stock was in a certain cap-size quintile, for each of the five quintiles. Note that the columns sum to 100 (with some allowance for rounding), but the rows do not. This is because the quintiles are defined based on all stocks in the CRSP database and hence the distribution across quintiles of stocks traded in this dataset does not need to be even.

Stock-year observations in the DE strengthening group are also concentrated in the larger cap-size quintiles, but to a lesser extent. Whilst large cap stocks have a lower DE as a group, there is clearly an important subset that have high influence and make the aggregate DE stronger. This would not be evident when looking only at individual stock DE scores, or aggregate scores for cap-size groups. In contrast to the high influence groups, the DE neutral observations are distributed evenly across all cap-size quintiles. This confirms that as well as low cap-size stocks with few positions, there are also large cap stocks with low influence due to the ratios of their component counts being close to the aggregate values.

Table 1.3 gives the corresponding percentages for the volatility quintiles. The stocks with the lowest volatilities are in quintile 1 and those with the highest are in quintile 5. Since volatility tends to decrease with cap-size, stocks in quintile 5 for cap-size will often be in quintile 1 for volatility. DE weakening stock-year observations are concentrated more in the lower volatility quintiles, and DE strengthening observations more in the higher volatility quintiles. However the distribution in both cases is much more even than for cap-size, showing that there are still a range of volatilities

Volatility quintile	DE Weakening %	DE Neutral %	DE Strengthening %
1	29.76	16.83	7.95
2	27.22	19.36	13.78
3	21.15	21.29	18.47
4	14.83	21.56	31.18
5	7.03	20.97	28.62

Table 1.3: Percentage of stock-year observations belonging to each DE influence group where the stock was in a certain volatility quintile, for each of the five quintiles. Note that the columns sum to 100 (with some allowance for rounding), but the rows do not. This is because the quintiles are defined based on all stocks in the CRSP database and hence the distribution across quintiles of stocks traded in this dataset does not need to be even.

present in the largest cap-size quintile. The distribution across quintiles is again very even for the DE neutral group, reflecting the much wider cross-section of stocks which are in this group at any time. As in Section 1.5.3, these results again provide support for the finding in Kumar (2009b) that stocks with greater volatility have a stronger DE.

Changes from year to year

Assigning stocks to an influence group each year generates a time series of group membership for each stock. In order to examine the movement, or otherwise, of stocks between different influence groups, the data period is now split into pairs of consecutive years. For each stock and pair of years, the influence group of the stock in both years is recorded, along with information about its cap-size and volatility. As a result of the way the influence groups are defined, 76% of the 9,812 stocks remain in the DE neutral group for the whole data period. But there is still movement in and out of the DE strengthening and weakening groups each year. The pairs of years across all stocks can be grouped based on the combination of influence groups the stock was in during the two years.

1st group $\rightarrow$ 2nd group	% of observations
S $\rightarrow$ W	0.59
S $\rightarrow$ N	2.04
S $\rightarrow$ S	2.38
N $\rightarrow$ W	2.05
N $\rightarrow$ N	85.9
N $\rightarrow$ S	2.04
W $\rightarrow$ W	2.36
W $\rightarrow$ N	2.06
W $\rightarrow$ S	0.58

Table 1.4: The percentage of pairs of consecutive years for all stocks (49,060 in total), where the stock was in two particular DE influence groups across the two years, for all combinations of influence groups.

Table 1.4 shows the percentage of all pairs of years per stock, 49,060 in total, with each influence group combination. Each influence group is abbreviated to its first letter, so for example S $\rightarrow$ N indicates the stock was in the DE strengthening group in the first of the two years, and in the DE neutral group the second. W $\rightarrow$ W indicates that the stock was in the DE weakening group in both years. These results confirm that there is significant movement between the neutral group and the two high influence groups, and even some very large jumps between the strengthening and weakening groups.

Table 1.5 gives the mean percentage change in volatility and cap-size from the first year to the second for each combination of influence groups. The rows are ordered according to their mean percentage change in volatility in order to highlight the pattern in this variable. At the top of the table are the three group combinations where a stock switches to a group closer to the DE weakening end of the distribution i.e. switching from the strengthening group to the weakening or neutral groups, or from the neutral group to the weakening group. These group combinations all have a decrease in mean volatility to go along with their move towards more DE weakening influence groups. The stocks that make the largest movement in this direction, from

1st group $\rightarrow$ 2nd group	Mean volatility change %	Mean cap-size change %
S $\rightarrow$ W	-7.44	29.87
S $\rightarrow$ N	-4.92	28.4
N $\rightarrow$ W	-1.11	32.5
S $\rightarrow$ S	0.01	28.34
W $\rightarrow$ N	0.12	11.29
W $\rightarrow$ W	0.66	16.0
N $\rightarrow$ N	1.06	37.09
N $\rightarrow$ S	4.1	36.06
W $\rightarrow$ S	8.52	10.52

Table 1.5: The mean change in cap-size and volatility for stocks in two particular DE influence groups across two years, for pairs of consecutive years across all stocks.

the strengthening straight to the weakening group also have the largest fall in mean volatility.

At the other end of the table, the two group combinations with the greatest increase in mean volatility are those where stocks move closer to the DE strengthening end of the distribution, with the stocks making the largest move again having the largest increase in volatility. Stocks moving from the weakening to the neutral group also have a positive mean change in volatility, but it is smaller and comparable with stocks that stay in the same group, who all have small but positive changes in volatility.

All group combinations have a positive change in mean cap-size. This is because stock prices were generally rising during the data period and particularly in the last two years. The fact that stocks moving from the weakening to the strengthening group have the smallest increase in mean cap-size suggests that poor performance and a large increase in volatility can make a usually DE weakening large cap stock become DE strengthening instead. Similarly, the row above suggests that an increase in volatility coupled with unexpectedly strong performance can cause an otherwise low influence stock to switch to the DE strengthening group.

1.5.5 Summary

Existing work on individual investor trading behaviour tends to look only at the disposition effect and the factors affecting it in cross-section, and aggregated over a large number of investors or stocks. By decomposing the aggregate DE into its component parts, this section has been able to establish how the different parts interact to create changes in the DE over time. The fall in aggregate DE during '95 and '96 is driven by a dramatic decrease in the number of opportunities to sell stocks for a loss at the time. This in turn is due to the strong performance of the stock market in '95 and '96, the start of a boom that culminated in the dotcom crash. Despite the fall in the number of opportunities to sell for a loss, investors still chose to sell a similar number of stocks for a loss as in previous years, likely due to the desire to offset capital gains and reduce their tax burden. Hence the proportion of losses realized increased and the disposition effect fell.

Looking at time series of the different component parts also revealed a difference between the largest 20% of stocks in terms of market capitalization, and those in the rest of the distribution. This large cap group had a lower DE than the rest of the stocks for most of the data period, but with consistently more sales for a gain than for a loss. In contrast the smaller cap group had roughly equal numbers of sales for a gain and a loss each month. The lower DE for the large cap group is produced by the much greater number of opportunities to sell these stocks for a gain than for a loss; the opposite was true for the smaller cap stocks with many more opportunities to sell for a loss in most months.

Decomposing the aggregate DE into the parts contributed by individual stocks showed that a small number of stocks have extremely large influence in terms of determining the aggregate DE, with the majority having essentially zero influence. Most of these high influence stocks are in the largest 20% of stocks in terms of cap-size due to these stocks being by far the most commonly held in the dataset. This shows that whilst smaller cap stocks have a higher DE as a group, it is in fact a small number of stocks in the large cap group that increase the aggregate DE by the greatest amount.

Looking at these high and low influence groups over time has shown that stocks do move between them from year to year. Stocks which move closer to the DE strengthening end of the distribution exhibit an increase in volatility, whilst stocks which move in the opposite direction exhibit a decrease.

1.6 The Cox model in detail

The count-based measure of the DE used in the preceding sections was useful for exploratory purposes and detecting aggregate patterns in the data. But to formally test the effect of a covariate on the DE, whilst controlling for the effect of others, a regression model is the best option. Proportional hazards (PH) regression models are preferred to other models such as linear or logistic regression, due to the reasons described in section 1.3.2. Hypothesising the presence of a DE is a claim about the rate at which positions are sold, specifically that gains are sold at a greater rate than losses. Hence modelling this rate directly, as is done in a PH model, will provide a better framework for testing claims about the DE. This section will introduce the Cox model, a particular type of PH model, and discuss its use for measuring the DE.

Let $\lambda(t, X(t))$, for $t > 0$, be the rate at which stock positions are sold, which shall be referred to as the hazard rate. It is dependent on the time t and the p-dimensional vector $X(t)$, which consists of observations on p covariates at time t. The hazard rate, or function, is defined as

$$\lambda(t, X(t)) = \lim_{h \downarrow 0} \frac{\mathbb{P}(t \leqslant T < t + h | T \geqslant t,\, X(t))}{h}$$

where T is the time at which the stock is sold. The hazard function is therefore approximately the probability of the stock being sold in the infinitesimal period of time immediately after t, conditional on the stock not having been sold prior to t and on the 'history' of the covariates during the period $[0, t]$.

Proportional hazards (PH) models assume that the hazard rate can be split into two components. The first, called the baseline hazard function, is common to all positions and depends only on time. The second depends on the current values of the covariates, which may themselves depend on time. The key assumption of PH models is that changes in the covariates have a multiplicative effect on the baseline hazard, and that this effect is constant over time. Checking this assumption will be an important part of the analysis presented below. In such models, the hazard function can be written as

$$\lambda(t, X(t), \beta) = \lambda_0(t) \exp(\beta^\top X(t)) \tag{1.4}$$

where $\lambda_0(t)$ is the baseline hazard function, and β is a p-dimensional vector of coefficients. Fully parametric PH models specify a functional form for the baseline hazard function and estimate its parameters along with β. However, there may not be a good basis for choosing one particular form over another. Cox et al. (1972) proposed a method for estimating β whilst leaving the baseline hazard function unspecified. Since the effect of covariates is of primary interest and there is no strong reason for choosing a particular form for the baseline hazard, the Cox model will be used throughout this chapter. Estimation of the Cox model involves maximizing the 'partial likelihood' that depends only on β, given by

$$l_p(\beta) = \prod_{i=1}^{n} \left(\frac{exp[\beta^\top X_i(T_i)]}{\sum_{j \in R(T_i)} exp[\beta^\top X_j(T_i)]} \right)^{c_i} \tag{1.5}$$

where n is the number of positions across all investors, and c_i and $R(T_i)$ are the censoring indicator and risk set - concepts from survival analysis that will be explained in the following. Positions are under observation for a set period of time after being purchased, called the follow-up time. If a position has not been sold by the end of its follow-up time, then it is said to be censored. A position is also censored if it has not been sold by the end of the data period. In (1.5), T_i is the time that position i is sold or censored. The censoring indicator c_i equals one if the position is sold at time T_i and zero if it is censored. $R(T_i)$ is the risk set, containing all positions that have not been sold or censored prior to time T_i, including position i itself. Hence censored positions only contribute to the likelihood by their presence in risk sets at times when other positions are sold.[13]

The follow up time used in the analysis below and in chapter 2 is 500 days. This decision was made because for longer holding periods it becomes less likely that the investor is actively considering selling the position, and the position is therefore not informative about the effect of covariates on the decision to sell. Related to this is the time scale that is being used to record a position's holding period and how this impacts the risk set that is constructed when the position is sold or censored. The holding period is defined as the time in days since the stock was first purchased. As a consequence of this, the risk set of positions contributing to the denominator in (1.5) are those that had been held for the same amount of time as the position which was sold. Hence positions in the risk set did not necessarily exist at the same calendar time as the sold stock, but did exist at the same

[13]This formula ignores the possibility of tied event times, which occur in this dataset when two or more positions are sold on the same day. The method proposed in Efron (1977) is used to approximate the likelihood contribution for a time point when there are multiple events.

'follow-up time'. This means that calendar effects, such as whether the hazard of selling is greater in December months, can be tested through the inclusion of dummy variables in the model.

The log of the likelihood in (1.5) is twice-differentiable and can therefore be maximised using the Newton-Raphson algorithm. Standard errors for the parameter estimates are obtained as the inverse of the information matrix, which is minus the second derivative of the likelihood for β.

In a Cox model for the hazard rate of a position being sold, the DE can be defined as follows. let $X_1(t)$ be an indicator function which equals one if the market price of the stock at t is greater than or equal to the price the stock was purchased at, and zero otherwise. Let β_1 be the regression coefficient associated with $X_1(t)$. Then there is a DE if $\beta_1 > 0$, as this implies that the position trading at a gain increases the hazard of it being sold, relative to if it were trading at a loss. Often the hazard ratio, defined as $\exp(\beta_1)$, will be used to summarise the effect of a covariate, as it gives the proportional change in hazard for a one unit change in the covariate. Once a Cox model has been estimated, the presence of a DE can be tested formally using the Wald test statistic

$$z = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$$

which asymptotically has the standard normal distribution.

1.7 Measuring the disposition effect in the LDB dataset using a Cox model

This section will estimate a simple Cox model using the LDB data in order to demonstrate this method for measuring the DE. The formation of the dataset used for model estimation will be discussed in detail, along with the choices made regarding the definition of the holding period of a position. After the model has been estimated, a method for checking the proportional hazards assumption will be described.

1.7.1 Sample formation

Positions are defined and recorded as follows. A position starts when an investor purchases a stock they do not already hold. There is information on the stocks an investor holds at the start of the data period, but not the purchase date or price, hence these pre-existing positions cannot be included, and are ignored for the purpose of determining if a stock purchase marks the start of a new position or not. If an investor purchases more of a stock they already hold, the quantity held and share-weighted average purchase price are updated accordingly. A position ends when the stock is sold. If only part of the quantity held is sold, the position is still considered to have ended on that date and will be entered into the dataset as such. This is because the holding period and reference price (the price used to determine if a position is trading at a gain or not, and if a sale was for a gain or not) after the first sale are ambiguous.

If a partial sale has occurred, the quantity held is still tracked and a new position in this stock cannot start until after the entire quantity held has been sold. There are some cases where a sale is made for a greater quantity of stock than was observed being purchased, due to the investor already holding some of the stock before the start of the data period. In these cases the sale is not considered valid for the purposes of this sample and the quantity held is set to zero. A subsequent purchase of the stock will start a new position as usual.

The other way a position can end is by being censored if it is not sold before the end of the 500 day follow up period. In the LDB dataset, 83%

of positions are sold before the end of this 500 day follow up period. A position can also be censored if a sale has not been observed before the end of the data period, which is November 29th, 1996. In the LDB dataset there are approximately 618,500 valid positions according to the criteria set out above. They are held by 55,300 unique investors and 70% are sold before being censored, for a total of 436,100 sales.

On each day during the holding period of a position, including the day it is sold or censored but not the day it is purchased, the paper status of the position is recorded. Following Odean (1998) The position is trading at a gain if the daily low price is above the average purchase price, at a loss if the daily high price is below the average purchase price, and is considered 'neutral' otherwise.[14]

1.7.2 Model results

Table 1.6 shows the results from fitting a Cox model to the set of positions extracted as described in the previous section. Note that all of the Cox models in this chapter and the next were estimated using the 'survival' package in R.[15] The sole covariate is a categorical variable indicating the current paper status of the position, with levels 'gain', 'loss', and 'neutral'. The 'loss' category is the reference level. The hazard ratio for the gain indicator is 1.80, meaning that positions trading at a gain have an 80% increased hazard of being sold relative to positions trading at a loss. Paper neutral stocks have a 6% increased hazard of being sold relative to losses. Both parameter estimates have highly significant Wald statistics, as indicated by the reported p-values. These results therefore confirm the basic finding that there is a strong DE in this dataset.

[14] On the date of a sale the sell price is used for this comparison, so a stock sold for a gain will always be recorded as having been trading at a gain on that date, and likewise with losses.

[15] Therneau (2015)

Variable	HR	P-value
Paper status: Gain	1.80	< 0.001
Paper status: Neutral	1.06	< 0.001

Table 1.6: Parameter estimates resulting from fitting a Cox model to positions in the LDB dataset, reported as hazard ratios (HR), and p-values for Wald tests of the estimates being different from zero (or equivalently for HRs, being different from 1).

1.7.3 Testing the proportional hazards assumption

The most important implication of the proportional hazards (PH) assumption is that the effect of a covariate is constant throughout the follow-up period. With the above model results, this means that the hazard of a position being sold is 80% greater if it is trading at a gain rather than a loss, regardless of how long the position has already been held. Whilst violation of the assumption does not invalidate the model, it does significantly alter the interpretation, particularly when only hazard ratios are reported. If an effect does change over time then the hazard ratio is only an average of this process, and if it changes a lot then this average can be misleading. Despite its importance, authors using PH models in the literature generally do not report having checked the assumption. In an effort to correct this, checking the PH assumption will be a particular focus when models that make it are used in this and the following chapter.

This will be done with a method that makes use of the residual process for Cox models suggested by Schoenfeld (1982).[16] The derivative of the log partial likelihood w.r.t β_k, the coefficient corresponding to the k-th covariate, is

[16] The derivation followed here is from Hosmer Jr et al. (2011), modified to accommodate the inclusion of time-dependent covariates.

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} c_i \left[z_{ik}(T_i) - \bar{z}_k(T_i) \right] \tag{1.6}$$

where c_i is the censoring indicator, z_{ik} denotes the value of the k-th parameter for the i-th unit and

$$\bar{z}_k(T_i) = \frac{\sum_{j \in R(T_i)} z_{jk}(T_i) \exp[\beta^\top Z_j(T_i)]}{\sum_{j \in R(T_i)} \exp[\beta^\top Z_j(T_i)]}$$

Substituting in the partial likelihood estimator $\hat{\beta}$ in place of β produces the estimator of the Schoenfeld residual for the i-th unit on the k-th covariate

$$\hat{s}_{ik} = c_i[z_{ik}(T_i) - \hat{\bar{z}}_k(T_i)] \tag{1.7}$$

The residual $\hat{s}_{ik}$ can be interpreted as the difference between the value of covariate k for the i-th unit at the time that it fails T_i, and the weighted average of covariate k values for all units still in the risk set at time T_i (which includes unit i). The weight for unit j is given by the 'risk score' $\exp[\hat{\beta}^\top Z_j(T_i)]$ i.e. a measure of how likely the unit was to fail at time T_i, relative to other units in the risk set, as estimated by the model.

A violation of the PH assumption implies that the true parameter β_k is not constant and is in fact a function of time, i.e. $\beta_k(t)$. Suppose that this is the case, but a standard Cox model is fitted to the data, producing the parameter estimate $\hat{\beta}_k$. Grambsch and Therneau (1994) show that

$$\mathbb{E}(s^\star_{ik}) \approx \beta_k(T_i) - \hat{\beta}_k$$

Where $s^\star_{ik}$ is the scaled Schoenfeld residual, defined as the k-th component of the scaled vector[17]

$$s_i^\star = [\widehat{\mathrm{Var}}(\hat{s}_i)]^{-1}\hat{s}_i$$

Hence the PH assumption can be assessed for covariate k by plotting $s_{ik}^\star + \hat{\beta}_k$ against each residual's respective failure time and adding a smoothed line through the points. A non-zero slope in this line indicates non-proportional hazards, with a positive slope implying the hazard ratio for this covariate increases over time and a negative slope implying that it decreases. It may be the case that PH is only violated during a certain period of time, and holds for the remainder.

Figure 1.9 shows this plot for the paper gain indicator in the model estimated in section 1.7.2. Since there are a very large number of events, and therefore residuals, only the smoothed line and a 95% confidence interval for it are plotted. Note that the event times plotted on the x-axis have undergone a transform recommended by the R 'survival' package author in order to assure an even spread of points across the axis i.e. intervals with fewer events, such as at the end of the data period, appear shortened. The downward slope starting after 100 days indicates that the DE becomes weaker after this point, meaning there is less of a difference between gains and losses in terms of the hazard of them being sold. This is an important caveat to the conclusion that there is a DE in this dataset, and one that would not have been detected had the PH assumption not been checked.

[17] In fact, $[\widehat{\mathrm{Var}}(\hat{s}_i)]^{-1}$ is usually approximated by $m\widehat{\mathrm{Var}}(\hat{\beta})$, where m is the total number of events, since it is easier to compute. All mentions of scaled Schoenfeld residuals in the remainder of this thesis refer to approximate residuals calculated in this way.

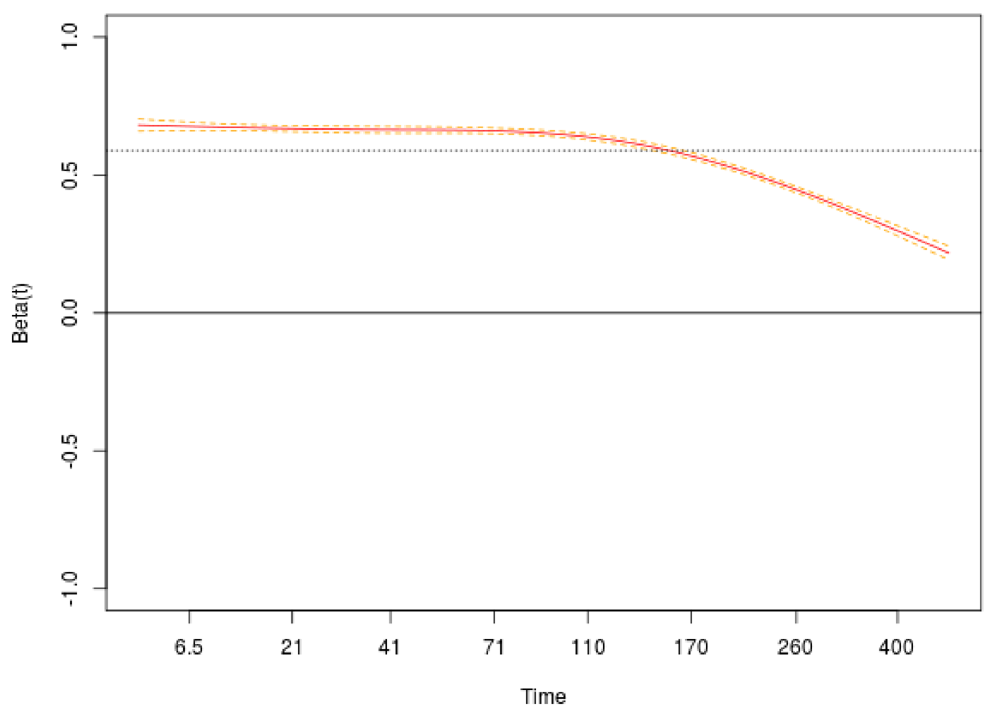

Figure 1.9: A plot of a smoothed curve fitted to the scaled Schoenfeld residuals for the paper gain indicator, with a dashed line at the level of the coefficient estimate β.

1.8 Connection with count-based measures of the disposition effect

The previous section used a Cox model to estimate the difference in the rate at which gains and losses are sold i.e. the hazard ratio for the paper gain indicator. The ratio PGR/PLR, introduced in section 1.4 and used throughout section 1.5, has a qualitatively similar interpretation. This section will formalise the relationship between the two quantities.

The hazard ratio can also be estimated non-parametrically using a formula proposed in Peto and Peto (1972)

$$\widehat{HR} = \left(\frac{O_G}{E_G} \Big/ \frac{O_L}{E_L} \right) \tag{1.8}$$

where O_G and O_L are the total number of sales observed for a gain and loss respectively. E_G is the number of sales for a gain that would be expected if there was no difference between the rate at which gains and losses are sold, and is defined by

$$E_G = \sum_{i=1}^{k} \frac{m_i}{R_i} N_{Gi}$$

where $i = 1, ..., k$ are the unique event times, m_i is the number of sales that occurred at time i, N_{Gi} is the number of positions trading for a gain at time i and R_i is the total number of positions that were at risk of being sold at time i.[18] E_L is defined similarly. The 'observed' and 'expected' quantities, O and E, are the same as those from the log-rank statistic for testing the equality of two survival curves. Note that for simplicity it is being assumed

[18]For a large dataset such as this, there is at least one event on each day of the follow up period, so sums over the unique event times are effectively sums over the days of the follow up period.

that a position must be either trading for a gain or a loss at all times during its holding period.

From the definition of the terms in formulas (1.1) and (1.2), PGR and PLR can be re-expressed using the variables from (1.8)

$$PGR = \frac{O_G}{N_G}, \qquad PLR = \frac{O_L}{N_L}$$

where

$$N_G = \sum_{i=1}^{k} N_{Gi}$$

and likewise for N_L. Now the Odean ratio from (1.3) can be written as

$$DE = \left(\frac{O_G}{N_G} \bigg/ \frac{O_L}{N_L} \right) \tag{1.9}$$

Dividing the hazard ratio estimator in (1.8) by this quantity produces a scaling factor that does not depend on the observed counts O_G and O_L

$$\frac{\widehat{HR}}{DE} = \frac{O_G E_L O_L N_G}{O_L E_G O_G N_L} = \frac{E_L N_G}{E_G N_L} \tag{1.10}$$

This implies that $\widehat{HR} > DE$ if

$$\frac{E_L N_G}{E_G N_L} > 1$$

or equivalently

$$\frac{E_L}{N_L} > \frac{E_G}{N_G}$$

Meaning the hazard ratio will be greater in magnitude than the modified Odean score if the ratio of the expected number of sales to the total number of positions in the risk set (summed across the unique event times) is greater for losses than it is for gains.

To demonstrate this relationship, the scaling factor derived in (1.10) can be calculated for the LDB dataset and applied to the DE score produced using the sample of positions collected as described in section 1.7.1. Due to the derivation above using the restriction that a position must be in either the paper gain or paper loss category at all times, intervals where a position is in the paper neutral category are recoded as paper losses. Estimating a Cox model produces a hazard ratio for the gain indicator of 1.78. The DE score for this sample is 1.57, and the scaling factor is 1.12. This produces an estimate of the hazard ratio of 1.76. The remaining discrepancy is likely due to the downward bias of $\widehat{HR}$ in large samples with tied event times, as discussed in Bernstein et al. (1981).[19]

[19]Note that $\widehat{HR}$ is equivalent to their O/E estimator since for this dataset all tables are informative i.e. there are gain and loss positions at risk at all time points.

Chapter 2

Using a frailty model to measure the effect of covariates on the disposition effect

2.1 Introduction

An important aim in the DE literature has been to determine which factors either strengthen or weaken the effect, particularly the characteristics of the investors themselves. Starting with Feng and Seasholes (2005), the use of a proportional hazards regression model, introduced in section 1.6, has been a common approach to this problem. In the context of survival analysis, an important property of trading record datasets is the natural grouping of trades at the investor level. Positions held by the same investor are likely to be dependent due the particular investment style of the investor, and this unobserved heterogeneity needs to be addressed when estimating a regression model. The problem of investor-level dependency is made more important by the extreme imbalance that can be present in datasets of this kind. A small number of investors account for a large proportion of overall trading activity, hence effect estimates will disproportionately reflect the idiosyncrasies of these investors if no effort is made to control for them.

In the existing literature the marginal method is used, which adjusts standard errors after the model has been estimated to account for the investor-level correlation between positions. This chapter will explore the use of frailty models as an alternative. The analysis will again make use of the LDB dataset, which was introduced in section 1.2. In a frailty model, the propensity to sell a position that is unique to each investor is modelled directly in a way that is analogous to the use of random effects in linear regression. For the dataset used here, the addition of a frailty component significantly improves upon the equivalent marginal model and allows a greater number of effects to be significantly estimated. In addition, the frailty model adheres much more closely to the proportional hazards assumption, which makes the parameter estimates more useful as summaries of the effect over time.

Results from the frailty model provide some new evidence on experience and learning; the number of trades an investor has made does not have a significant effect on the DE when the investor's self-assessed experience level is included in the model, and the length of time an investor has held an account does not appear to be a reliable measure of experience in this dataset, as those who opened an account most recently exhibit the weakest DE. Graphical checking of the proportional hazards assumption adds nuance to the interpretation of some variables. For example, the weakening of the DE in December is much larger for positions that have already been held for a long period of time, and differences in the DE between positions in small and large cap-size stocks only start to materialize after they have been held for roughly 100 days.

The remainder of the chapter is organised as follows. Section 2.2 discusses the problem of investor-level correlation in detail, along with both the marginal and frailty approaches to dealing with with. Section 2.3 describes the covariates that will be used for model estimation. Section 2.5 presents results from estimating both a frailty and marginal model, and highlights how they differ. Section 2.6 presents some additional models that check the robustness of the main results. Section 2.7 summarises the advantages that using a frailty model provided for this analysis and compares the results to those in the literature.

2.2 The problem of correlation

Datasets of brokerage trading records are naturally grouped at the investor level. In models where the observational units are stock positions, the possibility of dependence between positions held by the same investor needs to be addressed. Survival models for the lifetimes of stock positions are a popular method for measuring the disposition effect (DE) and factors that may affect its severity, hence investor-level correlation is an important issue in empirical work on the DE. In the existing literature, the marginal model approach has been used, where standard errors that are robust to correlation between positions are calculated after the model has been estimated.[1] [2]

This chapter will explore the use of frailty models as an alternative, where investor-level correlation is modelled explicitly through the use of latent variables. This method is analogous to the use of random effects in linear regression. Frailty models make assumptions about the structure of depen-

[1]Marginal models encompass a wider set of strategies for dealing with complex survival data, but for models of stock position lifetimes, the marginal approach only requires the calculation of robust standard errors.

[2]Ivkovic et al. (2005) deal with the issue of correlation caused by unobserved heterogeneity by stratifying the model based on the investor that held the position. This means each investor has a different baseline hazard function, which can absorb any heterogeneity not captured by the model covariates. However, this approach precludes the inclusion of covariates that are fixed at the investor-level, such as gender, since the variable only takes on a single value within each strata and a coefficient cannot be estimated. Hence, whilst effective, this approach cannot be adopted here due to the importance of demographic information in the analysis.

dence within and between groups of positions, whereas no such assumptions are made in the marginal approach. However if these assumptions are reasonable then using a frailty model will provide a number of advantages over the corresponding marginal model. This section will describe the different modifications that are made to the basic Cox model, as introduced in section 1.6, to produce the marginal and frailty models. The differences between the two methods in practice will then be discussed.[3]

2.2.1 Marginal model

In the marginal approach, all positions are assumed to be independent for the purpose of estimating β, the vector of model coefficients. Estimation proceeds in the same way as the standard Cox model, with robust standard errors being calculated afterwards. Lin and Wei (1989) derive an estimator for the variance of $\hat{\beta}$ in the familiar 'sandwich' form that is consistent in the presence of correlation between groups of units, in this case correlation between positions held by the same investor. Lipsitz et al. (1996) provide an estimator that is asymptotically equivalent to that of Lin and Wei, but easier to compute in practice using an infinitesimal jackknife method.

After a Cox model has been estimated, positions are split into q groups, one for each investor. The jackknife residual for group j is

$$J_j = \hat{\beta} - \hat{\beta}_{-j}$$

where $\hat{\beta}_{-j}$ is the vector of parameter estimates resulting from fitting the same model but after deleting the observations in group j. Fitting a new model for every group would be very costly, so instead $\tilde{\beta}_{-j}$, an approximation of $\hat{\beta}_{-j}$, is computed in the following way. Estimate $\hat{\beta}$ by letting the

[3]For a more detailed discussion of the theoretical and practical differences between marginal and frailty models, see Wienke (2010).

Newton-Raphson algorithm run until convergence as usual, then run it for another step after deleting the observations in group j, producing $\tilde{\beta}_{-j}$. The approximate jackknife variance estimate is then

$$\left(\frac{q-p}{q}\right)\sum_{j=1}^{q}(\tilde{\beta}_{-j}-\hat{\beta})(\tilde{\beta}_{-j}-\hat{\beta})'$$

where p is the number of parameters being estimated. In the model estimated in section 2.5 the robust standard errors are significantly larger than their naive counterparts, but in general they can be larger or smaller.

2.2.2 Frailty model

In contrast to a marginal model, frailty models assume a specific structure for the dependence between positions held by the same investor and incorporate it in the model directly. Specifically, a shared frailty model can be used where each investor is assumed to have a certain propensity for selling positions they hold. Conditional on this shared frailty and the model covariates, the survival times of positions held by the same investor are assumed to be independent. Similarly, the effect of covariates on the survival of positions held by different investors is assumed to be constant, conditional on the investor-specific frailties.

These investor-level frailties are included in the Cox model as fixed quantities that act multiplicatively on the baseline hazard function. In a shared frailty model for position lifetimes with $j = 1, ..., q$ investors and $i = 1, ..., n_j$ positions associated with each investor, the hazard function for the i-th position of the j-th investor is

$$\lambda_{ij}(t) = Z_j \lambda_0(t) \exp(\beta^{\top} X_{ij}(t))$$

Where X_{ij} is the covariate vector for the position and Z_j the unobserved frailty associated with investor j. The vector of frailties Z are assumed to be random variables with a common distribution, the parameters of which can be estimated. The gamma distribution is a common choice of frailty distribution, and is the one that will be used here. As discussed in Wienke (2010), the gamma distribution is used as a frailty distribution primarily because of its mathematical properties: it produces non-negative values, it can flexibly model a variety of distributional 'shapes' as its variance changes and the simplicity of its Laplace transform means the frailty terms can be easily integrated out of the log partial likelihood in the Cox model.

Since a scaling factor common to the frailties of all subjects can be absorbed into the baseline hazard of the Cox model, the gamma distribution for the frailties is taken to have expectation equal to one. Hence only the variance of the frailty distribution must be estimated as an additional parameter. Testing whether this variance is significantly different from zero is the primary way of establishing if investors do indeed have differing levels of frailty when it comes to selling positions they hold.

Derivations of the partial log-likelihood for a Cox model with gamma frailties are provided by Klein (1992) and Nielsen et al. (1992). As with many latent variable problems the EM algorithm can be used in this situation, as described by these references. In the R 'survival' package[4], which is used here, the problem is placed in the framework of penalised regression. When gamma frailties are used, this approach produces the same estimates as the EM algorithm, as shown in Therneau and Grambsch (2000). The package authors report from experience that this implementation typically converges significantly faster than an equivalent EM implementation. Details of the estimation procedure can also be found in this reference, and in the documentation accompanying the 'survival' package.

[4]See Therneau (2015) for some detail.

Shared frailty is a natural fit for investor trading data since there are a large number of investors and variation at the investor level is more of a nuisance factor that could obscure the effects of covariates in the model, which are the main focus of the analysis. Grouping variables can also be included as fixed effects, for example gender divides the data into two groups and is included in the models estimated later in the chapter. But this would not be feasible for investors since there are thousands of them and many have as little as one position in the dataset. Additionally, there is no need to test hypotheses about differences between specific investors, it is sufficient to have an estimate of the variation between investors in general, as is provided in the frailty model.

Possible reasons for there being such a difference between investors include their preference for risk, their beliefs about the market (e.g. whether there is price momentum or not), their investment objectives and the particular strategy they are following. For example, some more serious investors may trade very frequently and follow a strategy based on short term changes in stock prices. The holding periods of these investors will therefore be shorter than other investors in the sample. Shared frailty can separate out this kind of difference from the effects of covariates included in the model, that are in theory common across all investors.

2.2.3 Comparison

Because of the assumptions each model makes, the interpretation of parameter estimates is different in marginal and frailty models (note that any reference to a frailty model will from now on mean specifically a shared frailty model). Marginal models estimate average effects at the population level, with the average being across all investors. Hence the coefficient of a covariate describes the expected difference in survival for positions that differ in this covariate, regardless of which investors hold the positions. In a frailty model, a coefficient describes the difference in hazard between two positions that differ in the covariate, conditional on the frailties of the investors.

Since the marginal model is estimating effects averaged across all investors, the estimates can be biased if the sample either contains positions from only a small number of investors, or the distribution of positions across investors in the sample is very imbalanced. The former problem is not relevant in this dataset since over 5,000 investors are represented in the sample used for estimating the models in section 2.5.[5] The latter is more of a concern though since the sample is extremely imbalanced in terms of positions per investor. The median number of positions per investor is 7, whilst the maximum is 451 and investors in the top decile of this distribution account for 46% of the total number of positions. If the survival of positions is correlated at the investor level, then coefficient estimates that ignore this, as in a marginal model, will disproportionately reflect the behaviour of these most active investors. By separating the coefficient estimates out from each investor's static propensity for selling positions, the frailty model can provide a more accurate summary of the effect a covariate has on the hazard of selling.

Another advantage of frailty models is their ability to explain apparent violations of the PH assumption when effects appear to weaken over the course of the follow-up time. Positions that remain unsold for a long time are more likely to be held by investors with a lower propensity for selling i.e. frailty, hence the effect of all covariates seems weaker for positions that have been held for a long time when differing levels of frailty are not accounted for. This is more apparent when the follow-up time is long since there is more scope for observing longer survival times. The follow up time here is 500 days, compared to the median holding period of 86 days for positions where a sale is observed, so this will be a relevant issue. Effects that weaken over time can be detected using the Schoenfeld residuals, as discussed in Section 1.7.3.

[5] A small number of groups would suggest that the grouping variable should be included using fixed effects, rather than as a frailty term.

The results of section 2.5 show that for this dataset, adding frailty terms greatly improves the model's adherence to the PH assumption.

If shared frailty is a good description of the dependence structure in the data, then adding a frailty component will produce a more useful model for the reasons described above. In particular it will help isolate the effects of covariates in the model and their interactions, which is the main aim in analyses of the disposition effect. Assessing whether the addition of frailties does improve the model will be the focus of section 2.5.

2.3 Covariates

This section contains a description of the covariates that will be included in the model. Their effect on the DE will also be tested through the use of interaction terms. Following previous work on the topic, primarily Feng and Seasholes (2005) (F&S) and Dhar and Zhu (2006) (D&Z), many of them are related to the hypothesis that more sophisticated and experienced investors suffer less from the DE. Some stock-level variables will also be tested. Since the dataset is large, it should be possible to significantly estimate even small effects. This analysis again makes use of the LDB dataset, which was introduced in section 1.2. The LDB dataset was also used by D&Z in their analysis.

- Paper status: this variable records whether the position is currently trading at a gain (daily low and high prices above the purchase price), loss (daily low and high prices below the average purchase price) or neither. This third state shall be referred to as paper neutral. If positions in the paper gain group have a greater hazard of being sold than those in the paper loss category, then there is a DE. The paper loss category is used as the reference level, hence the model will contain an indicator for whether the position is paper neutral rather than a paper loss, and one for whether it is a paper gain rather than a paper loss. This latter indicator measures the strength of the DE, as discussed in section 1.6.

- Gender: coded as an indicator which equals one if the investor is male. 92% of investors are male, but significant differences between genders have been found in the LDB dataset. Barber and Odean (2001a) find that men trade 45% more than women and that their net return is roughly one percentage point lower per year as a result. However F&S note that, in a different dataset, gender only appears important in explaining the propensity to sell when few other control variables are included. Since a number of controls are being included here, further evidence on the importance of gender can be found.

- Age: the investor's age, as recorded on June 8 1997, 7 months after the end of the data period.

- Professional occupation indicator: whether the investor works in a professional occupation or not. The category is labelled as 'professional/technical', and contains 42% of investors. The remaining investors are in categories such as 'administrative/managerial', 'sales/service', 'clerical/white collar' or are retired. D&Z find that investors who work in professional occupations exhibit a disposition effect that is 20% weaker than those who work in non-professional occupations.

- Income: provided by the brokerage house is a proxy for income that is not continuous yet contains a large number of categories. This variable was therefore converted to a categorical variable by splitting it into quartiles. The four groups shall be referred to as the 'low', 'medium', 'high' and 'very high' income groups. D&Z find that high-income investors exhibit a DE 10% smaller than low income investors.

- Self-assessed experience: When opening an account, investors were asked to describe their investing experience on a four point scale. The levels are 'None', 'Limited', 'Good', and 'Extensive'.

- Tenure: the time in (possibly fractional) years since the investor opened their first account at the brokerage, recorded at the start of the data period. Negative values occur when an investor opened their first account after the start of the data period, as is the case for 12% of investors. Tenure is included as a measure of 'passive' experience, with the theory being that more experience investors will suffer less from the DE. The variable is split into four groups based on the bins defined by (-1, 1], (1, 4], (4, 8] and (8, 16]. Note that the maximum recorded value is 15.8 years. These groups will be referred to as Tenure 1-4, in the same order.

- Initial Diversification: The number of stocks the investor holds, across all their accounts, in the first month for which they have a record in the dataset. This variable is split into bins defined by (0,3], (3,7], (7,12] and (12, ∞), and the groups will be referred to as Diversification 1-4. F&S find that investors with greater initial portfolio diversification suffer less from the DE.

- Number of sales made to date: a time-varying covariate that only includes sales where the purchase occurred after the start of the data period. This variable is highly skewed, so its natural logarithm will be used when estimating models, as has been done in past research. F&S and D&Z find that having completed more sales reduces an investor's DE, implying that investors learn from their mistakes.

- Capitalization quintile of the stock: this data is taken from CRSP and is a time-varying covariate since it is updated annually. Quintile 1 contains stocks with the smallest cap-size, and quintile 5 those with the largest. The analysis in section 1.5.1 found that positions in stocks from higher cap-size quintiles exhibited a lower DE. Larger cap-size stocks have lower volatility, hence this finding is consistent with Kumar

(2009b), who finds that lower volatility stocks exhibit less of a DE. For each cap-size quintile, table 2.3 gives the percentage of holding days across all positions for which the stock held was in that quintile. Quintile 5 was used as the reference level for this variable in the models below, to increase the chance of finding significant differences between groups.

Cap-size quintile	1	2	3	4	5
% of holding days	2.29	4.77	8.14	14.45	70.35

- SIC major group: also from CRSP, this variable separates stocks into one of 10 industry groups. These are described in table 2.3, which also gives the percentage of positions due to stocks in each of the groups. This variable is included mainly as a control for possible correlation between positions in similar stocks. The manufacturing group (MAN) was used as the reference level for this variable.

SIC major group	Group label	% of positions
Agriculture, forestry & fishing	AFF	0.03
Construction	CON	0.57
Finance, insurance & real-estate	FIRE	8.52
Manufacturing	MAN	48.64
Mining	MIN	3.81
Public administration	PUBA	0.13
Retail trade	RETL	7.82
Services	SERV	12.29
Transport & public utility	TRAN	9.75
Wholesale trade	WHOL	2.46

- December indicator: equalling one for intervals of the position's holding period which are in December, and is hence time-varying. Due to the ending of the tax year, trading behaviour is significantly different in December. Odean (1998) finds that the DE is actually reversed for December months, which is known as the December effect. This effect was found in section 1.5.1, although with some variation between years. Including it in the model here may provide some more evidence on the topic.

- 1995/96 indicator: equalling one for intervals of the position's holding period which are in years 1995 or 1996. Analysis in section 1.5.1 showed that the aggregate DE decreases significantly in 1995/96, the last two years of the data period. Controlling for this effect will therefore be important when trying to isolate the effect of other variables.

For fitting the models described in the following sections, the sample was restricted to only those investors and positions for which full information was available on all covariates described in the previous section. This leaves a sample of 5,577 investors who hold 84,975 positions. Missing demographic information accounts for the majority of excluded positions. In fact, 70% of all positions observed in the dataset are held by an investor without occupation information, 50% each by investors without gender information or self-assessed experience, and 40% by investors without age information. The effect these variables have on the DE is of primary interest, so it will still be worthwhile to fit models using this heavily reduced sample. However a model without these variables will be estimated using a much larger sample in section 3.6.4. Parameter estimates for variables that appear in both models can then be compared to see if using the reduced sample makes a significant difference. The remainder of this section and those that follow shall deal exclusively with the reduced sample unless otherwise specified.

Of the 84,975 positions in the reduced sample, 60.4% are sold within 500 days of first being purchased, with the remainder being censored at that time. As discussed in section 1.7.1, only the first sale of a position is recorded, since the holding period and reference price of subsequent sales is ambiguous. The median holding period is 155 days, and 86 days if censored observations are not included.

2.4 Summary statistics

	Min.	Median	Mean	Max.
Age	24.00	48.00	50.53	92.00
Tenure (years)	-0.75	3.78	3.99	15.81
Diversification (# stocks)	1.00	3.00	4.98	269.00
Total sales	0.00	3.00	10.09	470.00

Table 2.1: Summary statistics for the continuous covariates. Calculated across investors, rather than positions, and only including investors with at least one position in the sample that will be used for model estimation. Diversification is the number of stocks the investor had in their portfolio in the first month for which there is a record of their positions.

Table 2.1 presents some summary statistics, at the investor level, for the continuous variables that will be used in the modelling to follow. Note that negative values of the tenure variable indicate that the investor opened their first account at the brokerage after the start of the data period, at the beginning of 1991. Most of note is the large range and skewness present in the measure of diversification (number of stocks present in the investor's accounts in the first month for which there is a record) and the total number of sales each investor makes. The latter will enter into any model as a time-varying covariate, taking the value of the current total. As mentioned in section 2.3, diversification is split into a categorical variable with four groups, and the log of number of sales made is used.

For the categorical variables, table 2.2 gives the percentage of investors and positions in each category, along with the latter divided by the former. This highlights groups of investors which hold a disproportionate number of positions. Investors who describe their experience as extensive and investors in the highest income quartile hold 40% and 60% more positions respectively than would be expected if all investors held the mean number of positions i.e. if there were no differences between groups. This shows the imbalance in the distribution of positions across investors, and again highlights the im-

	% of investors	% of positions	Ratio (Pos/Inv)
Female	8.01	6.73	0.84
Male	91.99	93.27	1.01
Non-professional	57.48	57.11	0.99
Professional	42.52	42.89	1.01
Income 1	31.53	27.46	0.87
Income 2	30.95	27.48	0.89
Income 3	25.75	28.78	1.12
Income 4	11.77	16.29	1.38
Experience: None	3.31	2.48	0.75
Experience: Limited	35.42	24.54	0.69
Experience: Good	47.35	51.01	1.08
Experience: Extensive	13.93	21.97	1.58

Table 2.2: Distribution of categorical variables across investors and positions. The ratio is the percentage of positions divided by the percentage of investors. Income has been split into quartiles, with quartile 1 containing those with the lowest incomes. Experience is self-reported by the investor.

portance of controlling for investor-level correlation between positions.

2.5 Comparison of marginal and frailty models

Following Feng and Seasholes (2005), the effect of a covariate on the DE can be measured by estimating a model for the hazard of a position being sold that includes an interaction term between the covariate and the paper gain indicator variable described in section 2.3. If the hazard ratio for the interaction term is greater (less) than one then the covariate makes the DE stronger (weaker). Interaction terms were initially added individually to a frailty model containing all main effects described in section 2.3. In this first stage, all covariates had a significant interaction with the paper gain indicator at the 5% level, with the exception of the SIC industry categories.

Hence only the main effects for this variable were included in subsequent models.

Both a marginal and frailty model containing all main effects and gain indicator interactions were then estimated. Table 2.3 contains hazard ratios for interaction terms that had significant coefficient estimates in both the frailty and marginal models. Significance was tested at the 5% level using the Wald statistic. As can be seen, the hazard ratios are broadly similar in both models. Table 2.4 contains hazard ratios for interaction terms that had significant coefficient estimates in the frailty model but not in the marginal model. The differences between models are larger for these covariates, and in each case the hazard ratio for the marginal model is closer to one i.e. a smaller effect. However the lack of significance itself in the marginal model is the most notable difference.[6] Had a marginal model been used exclusively then the analysis would show that gender, income and diversification did not have an effect on the DE. Full regression results for the frailty interactions are provided in section 2.5.1, and in the appendix for the main effects. The full results for the marginal model are also provided in the appendix.

	HR: frailty	HR: marginal
Gain * Experience: None	1.693	1.568
Gain * Experience: Limited	1.367	1.292
Gain * Experience: Good	1.220	1.161
Gain * December	0.504	0.503
Gain * '95/'96	0.762	0.742
Gain * Cap quintile 1	1.261	1.265
Gain * Cap quintile 2	1.171	1.197
Gain * Age	0.986	0.988
Gain * Tenure 3	1.213	1.192

Table 2.3: Hazard ratios for the interaction terms that were significant at the 5% level in both the marginal and frailty models.

[6] In addition to those shown in tables 2.3 and 2.4, there were no interactions that were significant in the marginal model but not in the frailty model. Such a result would be unlikely in the presence of any substantial correlation at the investor level.

	HR: frailty	HR: marginal
Gain * Male	1.159	1.104
Gain * Income 2	0.850	0.896
Gain * Income 3	0.902	0.931
Gain * Income 4	0.912	0.955
Gain * Diversification 2	0.924	0.983
Gain * Diversification 3	0.922	0.976
Gain * Diversification 4	0.858	0.936
Gain * Tenure 4	1.188	1.081

Table 2.4: Hazard ratios for the interaction terms that were significant at the 5% level in the frailty model but not the marginal model.

Having established that the frailty model produces materially different results than the equivalent marginal model, the next step is to assess the evidence for whether the frailty model is actually a better fit for the data, and hence that its results provide a more accurate representation of what is happening in reality. The significance of the frailty component itself can be tested using a likelihood-ratio test (LRT) comparing the frailty model with the equivalent marginal version, since they differ only by the presence of the frailty component. The null hypothesis of this test is that the frailty variance is zero, which would normally be problematic since it is on the boundary of the parameter space. However, Nielsen et al. (1992) show that the usual chi-squared distribution with one degree of freedom is still valid in this case. This LRT produces a highly significant test statistic (14,918), indicating that investor-level correlation is present, and that modelling it as shared frailties is supported by the data.

	AIC	Pseudo-R^2	Concordance
Marginal	1,084,489	0.41	0.71
Frailty	1,065,433	0.66	0.79

Table 2.5: Statistics comparing the overall adequacy of the marginal and frailty models.

Three statistics that can be used to compare the overall fit of the models are presented in table 2.5. They are the AIC, a pseudo-R^2 statistic due to Xu and O'Quigley (1999) and the concordance. The likelihood used in calculating the AIC and pseudo-R^2 for the frailty model has had the frailty component integrated out, so that it is comparable with the likelihood from the marginal model. The concordance is the proportion of all pairs of observations for which the model assigns greater hazard to the one that experiences an event first. A concordance over 70% is good for any survival model, so it is reassuring that both models are able to achieve this level. The frailty model performs better in each of these three measures, providing strong evidence that the inclusion of a frailty component produces a better fit to the data. The adjustment to standard errors in the marginal model does not affect these three statistics, hence they are the same as what would be obtained if a naive Cox model with no adjustment had been estimated.

Another important comparison to make is the degree to which the proportional hazards (PH) assumption holds in each model. If the effect of a variable appears significantly non-proportional, then the hazard ratio for that variable is not a good summary of its effect and care needs to be taken when interpreting it. The validity of the PH assumption will be checked for a particular variable by calculating residuals introduced by Schoenfeld (1982), and the graphical method for inspecting them proposed by Grambsch and Therneau (1994). This approach was discussed in detail in section 1.7.3. A non-zero slope in the smoothed curve plotted through the residuals indicates a deviation from PH, with a positive slope implying the hazard ratio for this covariate increases with survival time and a negative slope implying that it decreases.

Figures 2.1 and 2.2 show SSR plots for the interactions with the gain indicator for the 'Limited' group of the experience variable, and age respectively in the frailty model and marginal models. A horizontal dashed line is plotted in each figure (and all SSR figures below) at the parameter estimate for the variable. Clearly, there is a large departure from proportional hazards

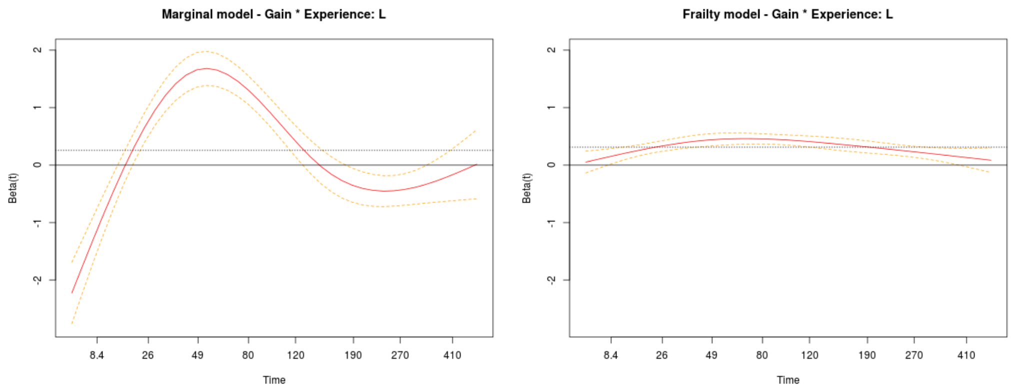

Figure 2.1: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the 'Limited' group for self-assessed experience, in the frailty and marginal models. Each curve is plotted with an approximate 95% confidence interval. A horizontal dashed line is plotted at the coefficient estimate for the variable in the respective model.

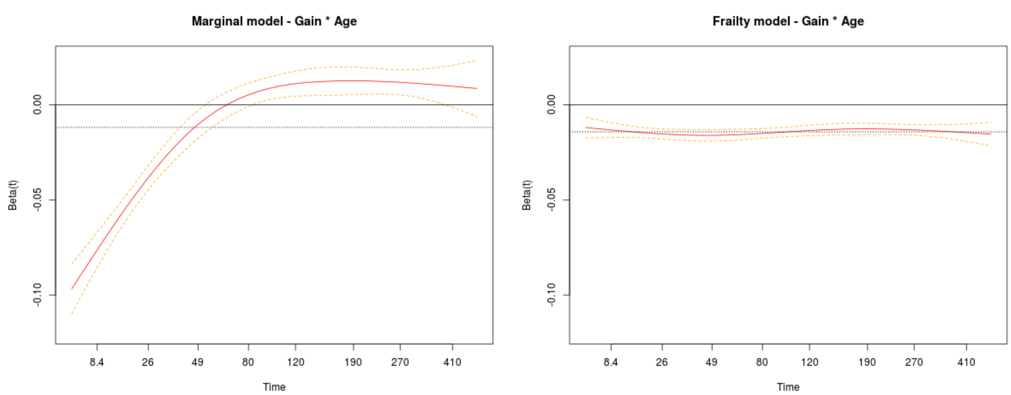

Figure 2.2: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and age, in the final frailty model and an equivalent marginal model. Each curve is plotted with an approximate 95% confidence interval. A horizontal dashed line is plotted at the coefficient estimate for the variable in the respective model.

in both plots for the marginal model. The interaction for age in the frailty model did not deviate much anyway, but for the experience group the deviation that is there is small compared to what is observed in the marginal model.

Overall adherence to the PH assumption can be assessed by computing the correlation between the SSR and the survival times for each of the interaction terms, as described in Grambsch and Therneau (1994). The median of the absolute values of these correlations for the frailty model is 0.0039, compared to 0.0193 for the marginal model. The authors also derive a test statistic for a non-zero slope that is asymptotically χ^2 under the null of no correlation. The null is rejected at the 1% level for 19/22 of the interaction terms in the marginal model, with test statistics that are at least one order of magnitude larger than their counterparts in the frailty model. In the frailty model, 8/22 have non-zero correlation significant at the 1% level. These results are displayed in full in the appendix. Since the sample size is large, highly significant test statistics are more likely. But interpretation of the effect of a covariate on the DE may not change much as a result of the PH assumption being violated. Reference to the smoothed SSR plots is therefore an important step when estimated coefficients are being interpreted, as they show the nature of the deviation from PH.

These results highlight two potential advantages of using a frailty model for measuring the effect of covariates on the DE in a dataset of this kind. Firstly, it has been possible to significantly estimate effects which would otherwise have been obscured in a marginal model. Secondly, doing so has greatly reduced the deviation from proportional hazards in the interaction terms, meaning the hazard ratios for these terms are good summaries of the effects and can be reported with confidence.

	HR	SE	p-value
Gain * Professional occupation	0.989	0.020	0.568
Gain * Male	1.159	0.039	< 0.001
Gain * Experience: N	1.693	0.068	< 0.001
Gain * Experience: L	1.367	0.028	< 0.001
Gain * Experience: G	1.220	0.023	< 0.001
Gain * Income 2	0.850	0.027	< 0.001
Gain * Income 3	0.902	0.026	< 0.001
Gain * Income 4	0.912	0.031	0.003
Gain * December	0.504	0.032	< 0.001
Gain * 95/96	0.762	0.021	< 0.001
Gain * Cap quintile 1	1.261	0.068	< 0.001
Gain * Cap quintile 2	1.171	0.045	< 0.001
Gain * Cap quintile 3	1.056	0.033	0.099
Gain * Cap quintile 4	0.981	0.025	0.435
Gain * Age	0.986	0.001	< 0.001
Gain * Diversification 2	0.924	0.024	0.001
Gain * Diversification 3	0.922	0.029	0.005
Gain * Diversification 4	0.858	0.029	< 0.001
Gain * Log sales made	0.988	0.008	0.128
Gain * Tenure 2	1.033	0.026	0.214
Gain * Tenure 3	1.213	0.025	< 0.001
Gain * Tenure 4	1.188	0.032	< 0.001

Table 2.6: Hazard ratios, standard errors and p-values for the interaction terms in the frailty model with all interaction terms included.

2.5.1 Frailty model results

This section will discuss the results for each interacted variable in the frailty model, as shown in table 2.6. Hazard ratios and Wald test p-values for the main effects are provided in the appendix. As well as the hazard ratio and its statistical significance, the presence of any non-proportionality in the effect of an interaction term will also be considered, as this may change the interpretation.

In this full model, the interactions for the professional occupation indicator and the log of number of sales made to date did not have significant estimates. In the case of the sales made variable, the estimate retains its significance in a model estimated without the self-assessed experience variable. In this model, the hazard ratio is 0.977. The inter-quartile range (IQR) for log of number of sales made is about 2.5, so an increase of this magnitude corresponds to a hazard ratio of 0.943 ($0.977^{2.5}$) and a reduction in hazard of roughly 6%. Although not strong, there is some association between the two variables. Regressing log sales on experience at the interval level[7] produces an R^2 of 0.04. Since the effect of log sales is small anyway in the model without experience, the addition of experience appears to be enough to wipe it out. The professional occupation indicator still does not have a significant estimate in a model without experience, and it is not strongly correlated with any other variable. When tested individually, the hazard ratio for the indicator was 1.09, meaning the DE was stronger for investors who worked in a professional occupation. This is contrary to what was expected based on past research. Due to its insignificance in the full model though, the conclusion made here must be that it does not have an effect when other factors are controlled for.

Moving on to variables that do have a significant effect on the DE, the

[7] Intervals are blocks of time during the holding period of a position where none of the time-varying covariates change.

hazard ratio for the gender interaction is 1.16, so on average the DE is 16% stronger for men. However this is a case where the SSR plot reveals large fluctuations of the smoothed curve for $\beta(t)$ around the estimate (indicated by the red curve and black dotted line respectively), as shown in figure 2.3. Whilst the effect is not constant over time, it is consistently positive. Hence the conclusion of a difference between genders in this regard still seems valid, although the magnitude is not constant as holding period increases.

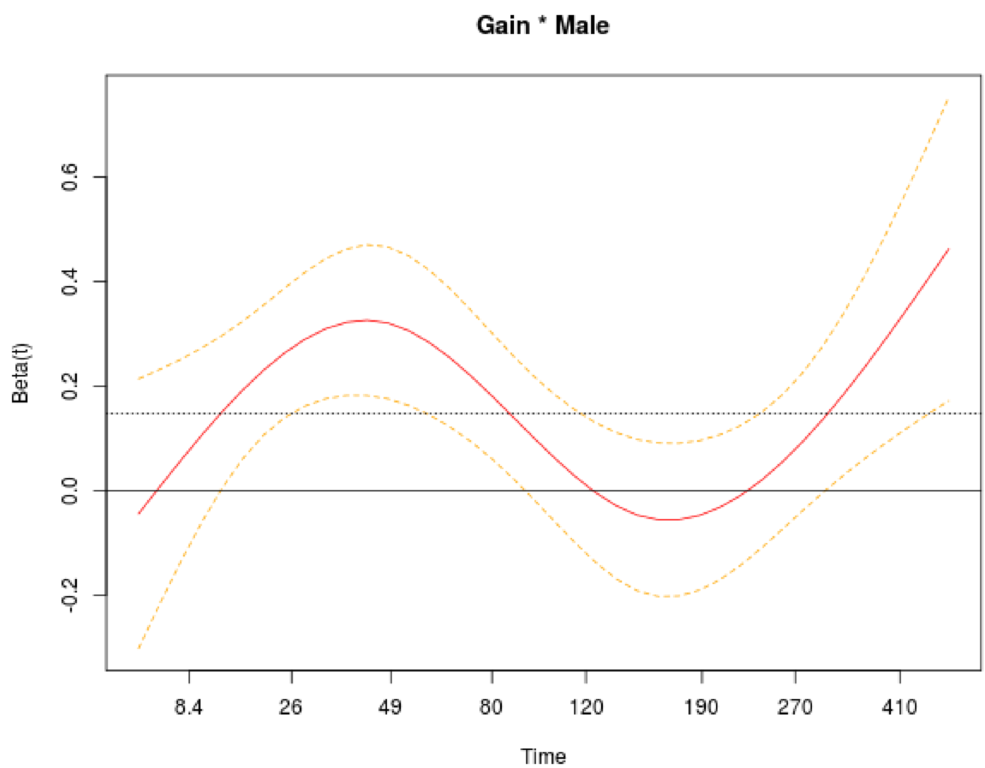

Figure 2.3: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the gender indicator (equalling one if the investor is male) in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate.

The experience variable has both highly significant main effects and interaction terms, with more experienced investors selling positions at a greater rate in general and exhibiting less of a DE. Compared to the 'extensive' group, the 'good', 'limited' and 'none' groups have a DE that is 22%, 37% and 69% stronger respectively. This is the largest effect amongst all of the categorical variables. The SSR plots show a peak in the effect after around 80 days for the 'limited' and 'good' categories, as shown in a plot of the former in figure 2.4 (the shape for the 'good' category is similar, but less severe). This means that the difference between these categories and the 'extensive' category takes some time to fully materialize, and also diminishes after the holding period has reached a certain length.

For age, an increase of one year corresponds to a 3.4% reduction in DE, and the IQR of 16 years corresponds to a 20% reduction. The SSR plot does not show any sign of non-proportionality, hence the hazard ratio estimate is a good summary of the effect. The results for the income variable show that income group 1 (lowest income) has the largest DE, as expected. But the largest reduction in DE is exhibited by income group 2 (15% reduction), rather than the two higher income groups (roughly a 9% reduction for both). For group 3, the SSR plot shows a stronger effect for the first 50 days of a holding period, during which it is closer to that of group 2, before decreasing and remaining stable for the rest of the follow-up time.

Investors with a greater level of initial portfolio diversification had a weaker DE on average, with those holding at least 12 different stocks having a 14% reduction in DE compared to those holding 1 or 2. Holding between 3 and 12 stocks lead to an 8% reduction, with no difference between those holding more or less than 7 within this range. Only the SSR plot for the third group shows cause for concern, with the effect falling to essentially zero between roughly 40 and 80 days, as shown in figure 2.5.

Only tenure groups 3 and 4 (first account opened between 4 and 7 years prior to start of data period, and first account opened at least 8 years prior respectively) have significantly different effects on the DE than group 1

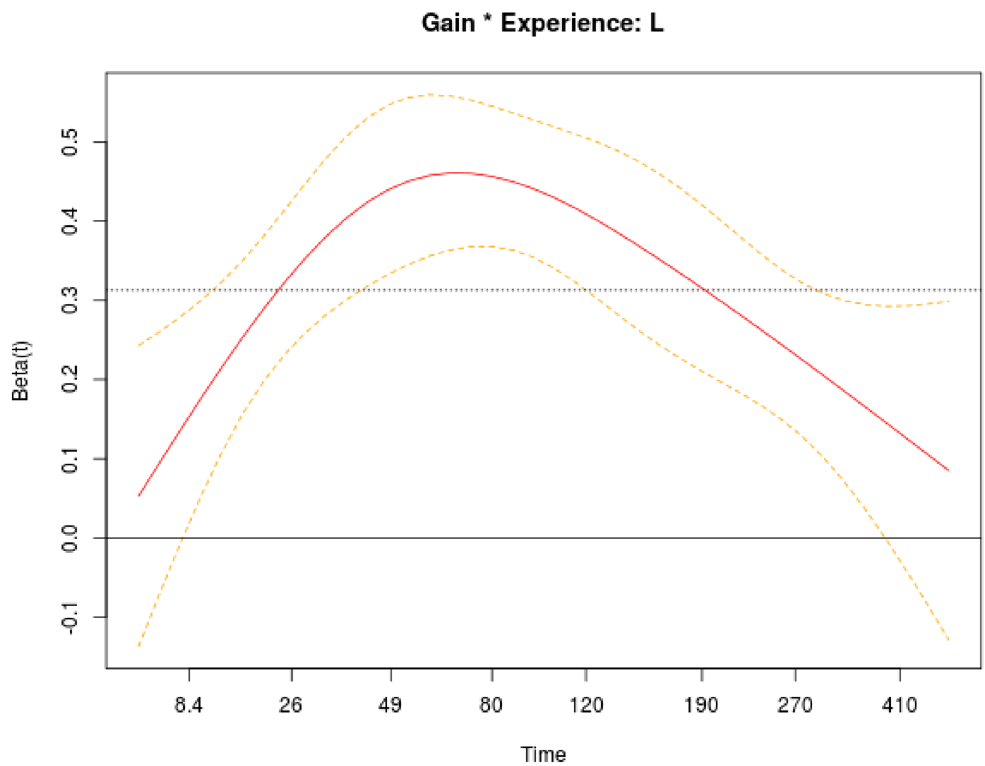

Figure 2.4: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the 'limited' experience category in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate.

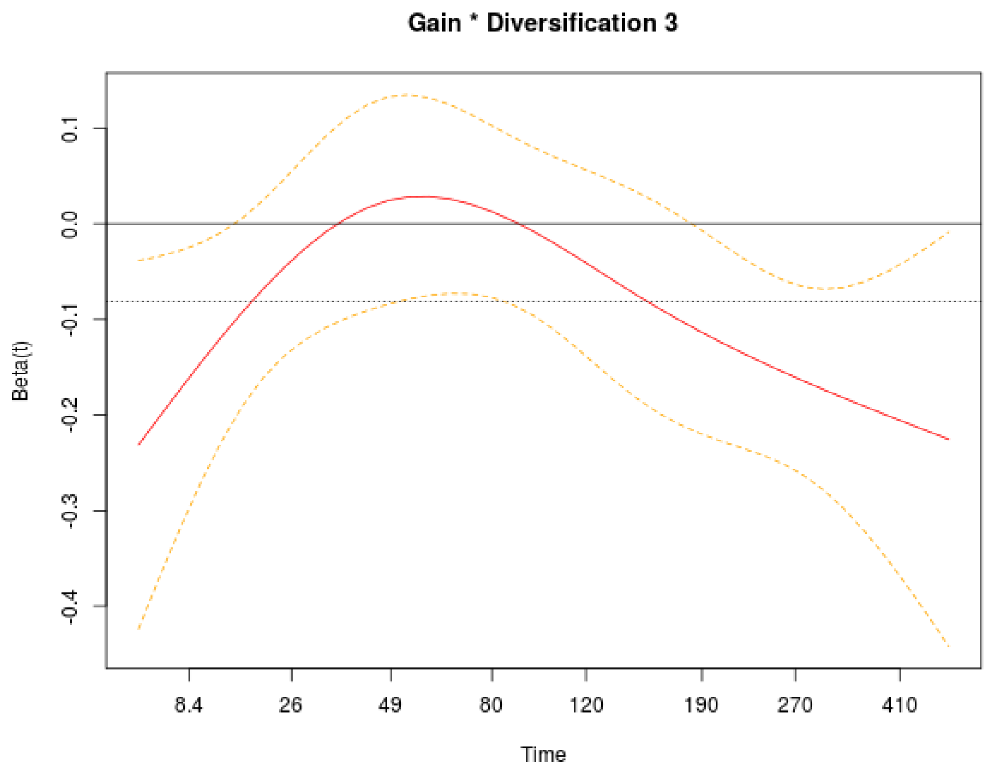

Figure 2.5: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the third diversification group (holding between 7 and 11 stocks) in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate.

(account opened between one year prior and one year following the start of the data period). Both exhibit a DE that is 20% larger than group 1, meaning that investors who have held an account at this brokerage for at least 4 years have a stronger DE than those who have held one for less. This result will be discussed further in section 2.7.

As expected, the DE is reduced in December months, in fact by 50% as shown by the hazard ratio. As shown in figure 2.6, the SSR plot shows the magnitude of the effect increasing from -0.2 (hazard ratio = 0.82) to almost -1.2 (hazard ratio = 0.3) by the end of the follow-up time. This means that the reduction in DE in December is much larger for positions that have already been held for a longer period of time. This is an important caveat to the finding of a significant December effect that would not have been revealed without inspecting the SSR plot.

Confirming the exploratory analysis, in this model the DE is significantly lower in years 1995 and 1996, with the hazard ratio indicating an average reduction of 24%. The SSR plot shows that the effect is stronger at the start of the follow-up time and weaker towards the end, but it stays close enough to the parameter estimate such that the hazard ratio is still a reasonable summary of the effect.

Stocks in the bottom two quintiles of capitalization size have an increased DE relative to those in the largest cap-size quintile. There is an increase of 26% for the first quintile and 17% for the second. There is not a significant difference between the largest quintile and quintiles 3 and 4. Figures 2.7 and 2.8 show SSR plots for these interaction terms. Importantly, the effect is not significantly above zero until roughly 100 days into the follow-up time in both cases, and continues increasing after this point. Since the interaction terms for quintiles 3 and 4 were not significant, this means that cap-size does not have an effect on the DE until roughly 100 days into a holding period.

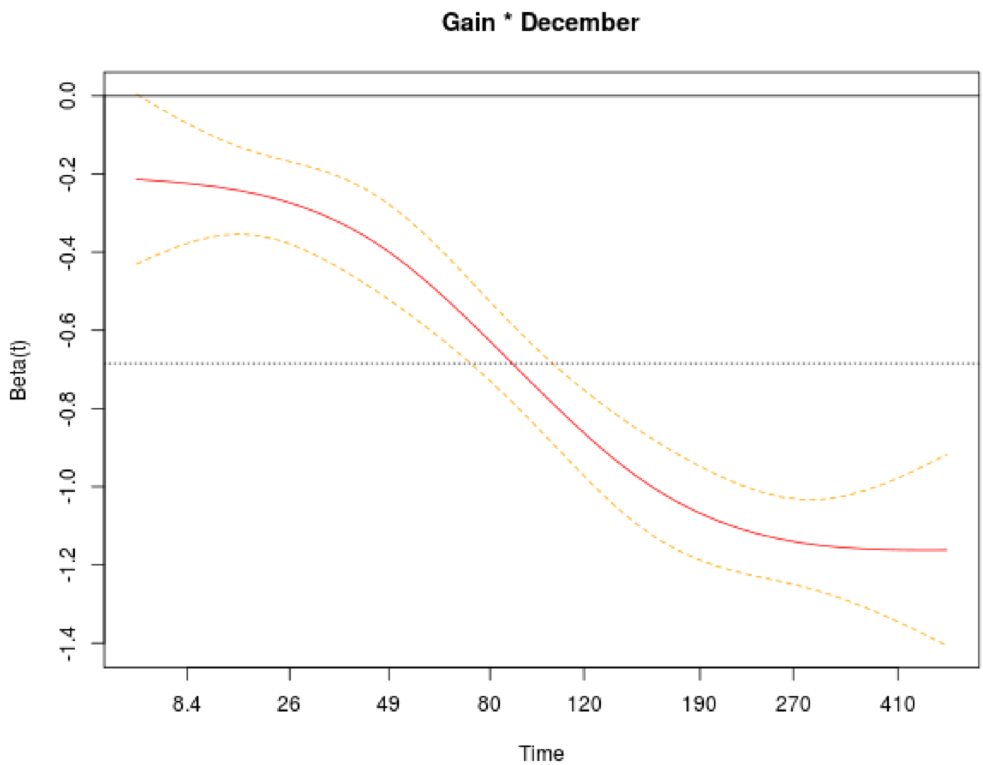

Figure 2.6: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the December indicator in the frailty model. An approximate 95% confidence interval for the curve is also plotted. A horizontal dashed line is plotted at the coefficient estimate.

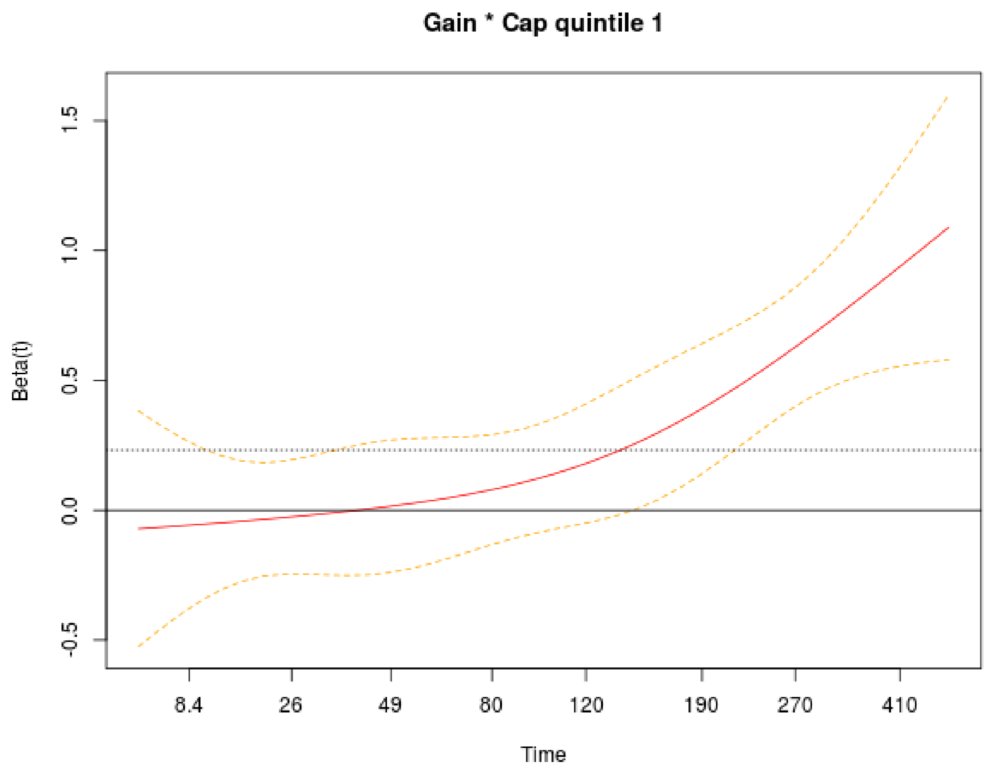

Figure 2.7: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the first capitalization size quintile, containing stocks with the smallest cap-sizes.

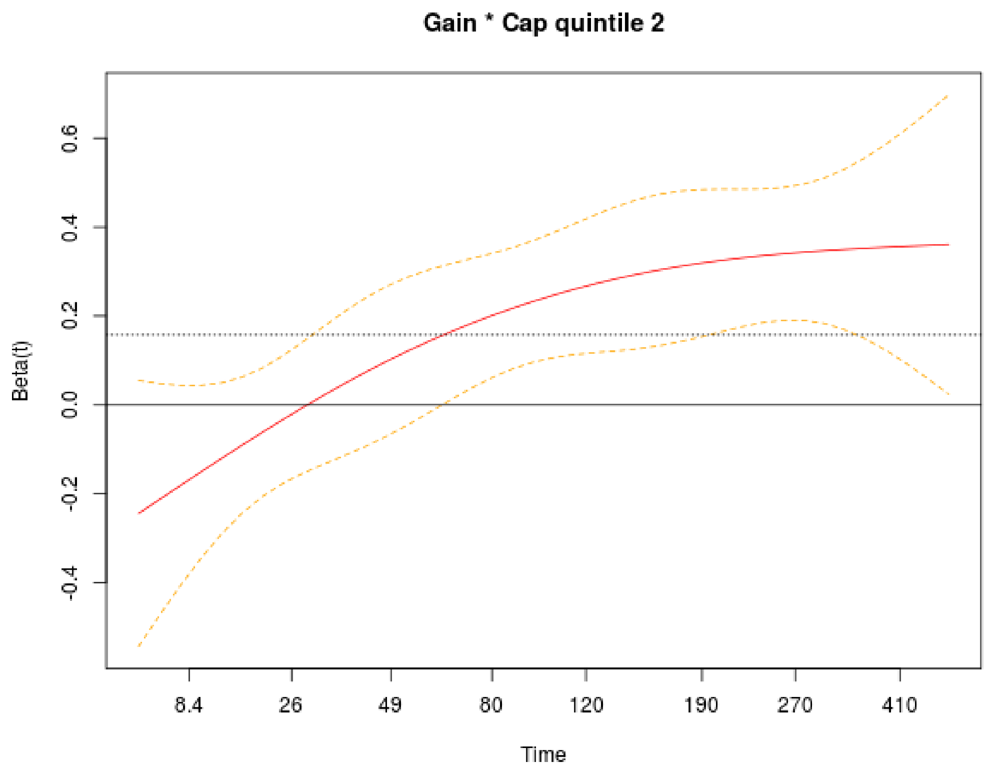

Figure 2.8: Plot of a smooth of the scaled Schoenfeld residuals for the interaction term between the paper gain indicator and the second capitalization size quintile.

2.6 Robustness checks and supplementary models

2.6.1 Model without demographic information

One thing that distinguishes the LDB dataset from others used in the literature is the large range of demographic information that is available for some investors. The improvement in the model due to the inclusion of demographic information is therefore of interest. To assess the extent of this improvement, the frailty model from section 2.5.1 was re-estimated without any of the demographic variables or their interactions. This includes age, gender, the professional occupation indicator, self-assessed experience, and income group. In untabulated results, the variables common to both models are very similar, in sign, magnitude and significance. The only substantive difference is that the sales made interaction is significant in the model without the demographic variables. As was mentioned in section 2.5.1, the log sales made variable was significant if the experience variable was excluded from the model, hence the result here is not surprising.

The two models can also be compared in their overall adequacy. As would be expected, the model with demographic variables is a better fit, judged by a few different measures, but the difference is small. The model with demographic variables has concordance of 0.791 compared to 0.790 for the model without, AIC of 1,056,616 compared to 1,057,150 and pseudo-R^2 of 0.655 compared to 0.652. The effect sizes for the demographic variables are fairly small, with the exception of self-assessed experience, so the model would not be expected to suffer greatly in their absence. But when the model was re-estimated without these variables, new frailty values were also computed, and since all the demographic variables are static over time, it is likely that the frailty component was able to do a good job of adjusting for the additional heterogeneity introduced by the their omission. This provides some evidence that the frailty model is also able to adjust for sources of heterogeneity that are truly unobserved.

2.6.2 Model without demographic variables and larger sample

In order to include demographic variables in the model, the sample had to be restricted to investors for whom full demographic information was available. This meant the sample used to fit the final model in section 2.5.1 contained 5,577 investors and 84,975 positions. This is compared to the 55,000 investors with 618,000 positions recorded in the full dataset. It is possible that the small sample with demographic variables is biased in some way that would result in different parameter estimates for the non-demographic variables compared to what would be obtained with the full sample. To check for this, the model from the previous section was re-estimated but with no demographic variables and the largest possible sample. Due to memory constraints, this contained 364,664 randomly sampled positions.[8]

Table 2.7 shows the hazard ratio and estimate significance for each interaction term that is common to both the model from section 2.5.1, which also contained demographic variables, and the model estimated using the much larger sample. The hazard ratios for the December and '95/'96 indicators are very similar and both highly significant. For cap-size quintiles the pattern in hazard ratios is similar with positions in stocks from smaller cap-size quintiles having a stronger DE. However the differences between quintiles 3, 4 and 5 were significant in the larger sample size model, but were not in the small sample size model.

[8]Although the full dataset of positions itself is only 1.3GB in size, the procedure for fitting a frailty model is very memory intensive. A machine with 16GB of memory was able to fit a model using half of the full dataset, hence half of all positions were randomly sampled. Fitting a model using this sample took 22 hours.

	HR (S)	HR (L)	p-value (S)	p-value (L)
Gain * December	0.504	0.540	< 0.001	< 0.001
Gain * 95/96	0.762	0.812	< 0.001	< 0.001
Gain * Cap quintile 1	1.263	1.433	0.001	< 0.001
Gain * Cap quintile 2	1.170	1.207	< 0.001	< 0.001
Gain * Cap quintile 3	1.055	1.106	0.104	< 0.001
Gain * Cap quintile 4	0.980	1.078	0.419	< 0.001
Gain * Diversification 2	0.922	0.904	0.001	< 0.001
Gain * Diversification 3	0.919	0.908	0.003	< 0.001
Gain * Diversification 4	0.858	0.790	< 0.001	< 0.001
Gain * Log sales made	0.989	0.941	0.143	< 0.001
Gain * Tenure 2	1.035	1.033	0.178	0.007
Gain * Tenure 3	1.218	1.061	< 0.001	< 0.001
Gain * Tenure 4	1.190	1.099	< 0.001	< 0.001

Table 2.7: Hazard ratios for interaction terms in a frailty model that does not include any demographic information and estimated using a much larger sample (364,000 positions compared to 85,000), denoted by (L). Compared with hazard ratios for the same interactions as estimated in the full frailty model from section 2.5.1, which used the smaller sample, denoted by (S).

There is again close agreement for diversification, with groups 2 and 3 having a small decrease in DE relative to group 1, and group 4 having a larger relative decrease. As expected from the supplementary models estimated already, log of sales made is highly significant in the larger sample model, with a hazard ratio of 0.941. For an increase of 10 sales made this translates into a reduction in DE by 13%, and a reduction by 17% for an increase of 20 sales made. For investor tenure, the large sample model again finds that it is investors who have held an account for the least amount of time (tenure group 1) that have the lowest DE, with tenure groups 2, 3 and 4 having a DE that is 3%, 6% and 10% than group 1 respectively. In the small sample model the increase for groups 3 and 4 was much larger, around 20% in both cases. As discussed previously, the increase in DE for investors who have held accounts for longer is the opposite of what was expected, and the reason for it remains unknown. It could be that tenure is capturing something else, such as an investor's willingness to 'shop around' for the best terms when choosing which brokerage to trade with. Shorter tenure may suggest a greater willingness to switch for a better deal, which may go along with a more sophisticated approach to investing, and hence a lower DE.

2.6.3 Recurrent positions

The frailty model estimated in section 2.5 assumes that the survival times of positions held by the same investor are independent conditional on the covariates and the investor's frailty. One possible deviation from this is the correlation that may exist between positions an investor takes in the same stock over time. Of the 84,975 positions recorded in the sample where full demographic information is available, 87% are positions where it is the first time the investor has held the stock during the data period. 9% are positions where it is the second time, and 2% where it is the third. This type of situation is common in medical applications of survival analysis, with one example being a model for recurrences of a chronic condition.

All positions which are not the first time an investor has held the stock shall collectively be referred to as recurrent positions. Since a minority of positions are recurrent, it is unlikely that any difference in an effect for recurrent positions would effect the overall result. But it is still of interest to check if any such difference can be detected. The simplest way to do this is to fit models using only first and only recurrent positions, and seeing if the parameter estimates differ. The results from these models can also be compared to those from the model fitted on all positions together. The hazard ratio and estimate significance for each interaction term in the three models is given in table 2.8

As can be seen, agreement between the three models is close in general. When an estimate is significant in each model, the sign[9] and magnitude are similar. There is a lack of significance in all but the third cap-size quintile interaction in the recurrent positions model, although only the first and second quintiles were significant in the full data model anyway.

There is also disagreement when it comes to the significance of the diversification group interactions. All three were significant in the full data model, whereas the first two are not in the first position model and the third is not in the recurrent position model. These results raise the possibility of the DE-reducing effect of increased diversification only really coming in to play for stocks an investor trades multiple times. This is something that could be investigated in more detail in future research. Also of interest is the significance of the log of number of sales made interaction in both the first and recurrent position models, despite it not being significant in the full data model. Since the effect is stronger in both the first and recurrent position models compared to the full data model, it does not seem as if it

[9]Or if hazard ratios are being compared, as they are in the table, whether the hazard ratios are above or below one.

	HR (1)	HR (2)	HR (3)	p-value (1)	p-value (2)	p-value (3)
Gain * Professional occupation	0.989	0.997	0.932	0.568	0.901	0.155
Gain * Male	1.159	1.127	1.236	< 0.001	0.004	0.038
Gain * Experience: N	1.693	1.573	1.946	< 0.001	< 0.001	< 0.001
Gain * Experience: L	1.367	1.378	1.248	< 0.001	< 0.001	0.001
Gain * Experience: G	1.220	1.179	1.423	< 0.001	< 0.001	< 0.001
Gain * Income 2	0.850	0.894	0.756	< 0.001	< 0.001	< 0.001
Gain * Income 3	0.902	0.924	0.835	< 0.001	0.007	0.006
Gain * Income 4	0.912	0.938	0.881	0.003	0.057	0.097
Gain * December	0.504	0.485	0.643	< 0.001	< 0.001	< 0.001
Gain * 95/96	0.762	0.783	0.732	< 0.001	< 0.001	< 0.001
Gain * Cap quintile 1	1.261	1.348	1.211	0.001	< 0.001	0.44
Gain * Cap quintile 2	1.171	1.203	1.489	< 0.001	< 0.001	0.006
Gain * Cap quintile 3	1.056	1.103	1.107	0.099	0.006	0.315
Gain * Cap quintile 4	0.981	0.999	1.059	0.435	0.964	0.398
Gain * Age	0.986	0.986	0.989	< 0.001	< 0.001	< 0.001
Gain * Diversification 2	0.924	0.957	0.823	0.001	0.104	0.001
Gain * Diversification 3	0.922	0.959	0.853	0.005	0.183	0.026
Gain * Diversification 4	0.858	0.877	0.931	< 0.001	< 0.001	0.332
Gain * Log sales made	0.988	0.954	0.913	0.128	< 0.001	< 0.001
Gain * Tenure 2	1.033	1.047	0.909	0.214	0.108	0.145
Gain * Tenure 3	1.213	1.190	1.193	< 0.001	< 0.001	0.005
Gain * Tenure 4	1.188	1.136	1.309	< 0.001	< 0.001	0.001

Table 2.8: Hazard ratios and p-values for interaction terms in the model estimated with: all data (1), first positions only (2), and repurchased positions only (3). First positions occur the first time an investor purchases a particular stock during the data period and all other positions repurchased positions.

is a case of two opposite effects cancelling each other out in the full data model. Again this is an issue for future investigation to explore.

Since the goal here is to test the effect of covariates on the DE, a model can be fitted where recurrence is controlled for and the estimates in this model checked to see if they differ from those produced by the final model in section 2.5.1. The most general way to control for a fixed effect of recurrence is to assign first and recurrent positions to different strata. This allows the set of first positions to have a different baseline hazard function to the set of recurrent positions. Hence any fixed difference between survival in the two groups will be absorbed into the baseline hazard, preserving the effect of covariates that is constant across the groups. In untabulated results, fitting a model stratified in this way produces results very close to those in section 2.5.1. Hence the conclusion is that the results in that section are robust to controlling for recurrence, but as discussed above there are some differences with recurrent positions that may be worth further investigation.

2.7 Summary and discussion

Section 2.5 showed that the addition of a frailty component significantly improved the Cox model for position lifetimes, in terms of both goodness-of-fit and adherence to the PH assumption. By isolating the effect of covariates and their interactions from the unobserved frailty unique to each investor, many more of these effects could be significantly estimated in the frailty model. This is of particular importance in a sample like this where the distribution of positions amongst investors is extremely imbalanced.

That the frailty model is able to control for unobserved heterogeneity is supported by the results of section 2.6.1, where a model was fitted using the same sample as the full interaction model, but without any demographic variables. The parameter estimates for variables common to both models did not change much, which would be a concern when variables known to be important are omitted. The overall performance of the model did also not greatly suffer, with concordance and pseudo-R^2 values close to the full model. This suggests the frailty terms were able to 'absorb' some of the static differences explained by the demographic variables when the model was re-estimated without them.

Graphically checking the PH assumption proved to be an important step in the modelling process. Deviations from PH substantially changed the interpretation of some effects compared to what would have been concluded if only the hazard ratios were considered. For example, the reduction in DE in December was shown to be much stronger for positions that had already been held for a long period of time. This adds importance nuance to the evidence on this topic.

2.7.1 Model results

This section will summarise the results of the analysis, making comparisons to those in Feng and Seasholes (2005) (F&S) and (Dhar and Zhu, 2006) (D&Z) when possible. A description of the methods used in these papers can be found in section 1.3.2.

Demographic covariates

The DE for men was 16% stronger on average in this sample. This is in contrast to F&S who find a DE that is 30% stronger for women, with a much more even gender balance: 51% men in their dataset from China compared to 80% in the LDB dataset. This suggests that gender differences in the prevalence of the DE vary between countries, but more research on this specific topic would be needed to understand the nature and causes of this variation. Despite a significantly weaker DE being found amongst those working in a professional occupation by D&Z, no such difference was found here. This lack of significance persists when the self-assessed experience variable is removed from the model.

For age, an increase of one year corresponds to a decrease in DE by 1.5%, or for the IQR of 16 years a decrease by 20%. D&Z also find that older investors suffer less from the DE, although comparing the magnitude of the effects is difficult due to the difference in methodology. Rather than greater age being associated with increased investing sophistication, F&S hypothesise, and confirm in their results, that younger investors in China will be more sophisticated due to older investors having grown up in a radically different economic system.

Those with incomes in the lowest of the sample quartiles have the strongest DE, but there is not a monotonic decrease for higher quartiles. The largest decrease in DE relative to the lowest income group was actually for the second lowest, and there was no difference between the top two quartiles relative to the lowest quartile. Assuming more sophisticated investors will exhibit a lower DE, these results show that income is not a reliable measure of sophistication when other factors are controlled for. D&Z split investors into three income groups, and find a 10% reduction in DE for the highest compared to the lowest. Their low and high groups correspond closely to the lowest and highest quartiles used here, for which a 9% difference was found.

Other investor-level covariates

Self-assessed experience proved to have a lot of explanatory power when it came to differences in the disposition effect. Investors with the lowest level of experience had a DE that was 69% stronger than that of the group with the highest level. It is interesting that investors' assessment of their own ability matched up so well with the extent to which they committed what is generally considered to be an investment mistake i.e. selling winners too soon and holding losers too long. This raises the question of why investors who consider themselves to be inexperienced trade at all.

It may be that they hope to learn by trading and improve their performance as they gain in experience. The results here provide mixed evidence for this being possible. When self-assessed experience is not included in the model, an increase in the number of trades an investor has made reduces their DE, with an increase of 10 trades corresponding to a reduction in DE by 13%. Importantly, only trades made during the data period can be counted, and for some investors this will represent only a small portion of their investing career. However, when self-assessed experience is included in the model, the number of sales an investor has made no longer has a significant effect. This suggests that the number of sales made is actually capturing differences that existed before the start of the data period, and not learning that happens as a result of the trading that's observed.

One explanation for this is that the investors who rate themselves as having high experience hold and sell more positions during the data period. In fact the group with the highest self-assessed experience contains 14% of investors, but account for 22% of all positions. The number of sales made variable will increase faster for investors in this group, and throughout the data period larger values for number of sales made will typically indicate an investor has a higher level of self-assessed experience. F&S and D&Z also found that trading more reduced an investor's DE, but did not have a similar measure of self-assessed experience in their models. This provides some new evidence on the question of whether investors learn from trading, and shows that the number of trades an investor has made is perhaps not a good measure of gained experience by itself.

Investors with a greater level of initial diversification had a weaker DE. It was reduced by 14% for those holding at least 12 stocks initially, compared to those holding less than 3. Investors initially holding 3-11 stocks had a DE 8% lower than those holding less than three. F&S found that holding at least two stocks corresponded to a 16% reduction in DE. It seems likely that the effect of diversification is weaker in this model since a greater range of variables, particularly self-assessed experience, are being controlled for in the model relative to that of F&S.

Somewhat surprisingly, those who had held an account for the shortest time had the weakest DE. The increases in DE for groups who had held an account for longer were smaller, but still present, in the model estimated using the much larger sample. This is contrary to what was expected, with one explanation being that account tenure is not measuring investor experience or sophistication in this dataset. Tenure is positively correlated with age, but not strongly so (Pearson's correlation coefficient of 0.15), and the mean value of tenure is only 4 years. This suggests there is not a strong relationship between account tenure and the amount of time an investor has been investing in total. If there was, we would expect more large values of tenure, and large values would only be possible for older investors, hence stronger correlation with age.

This result is also the opposite of what would be expected if investors with lower investing ability were dropping out as a result of their poor performance. Assuming that better investors will suffer less from the DE, and that poor performance makes an investor more likely to close their account[10], then the investors who have held accounts the longest should have a lower DE as a group. The decision to cease trading or even close an account is likely more complicated than whether an investor's trading performance has been good or not, which itself is affected by more than just the extent to which they have exhibited the DE. Further research would be needed to untangle the different factors in this issue.

[10]Or at least stop actively trading common stocks, in which case they would not appear in the sample used here, even if they do not close their account at the brokerage entirely.

Stock-level and calendar covariates

The DE was greater for positions in stocks with a smaller cap-size. In the model with demographic variables there was not a significant difference between cap-size quintiles 3-5, but a difference between these groups was detected in the model estimated with a much larger sample size in section 2.6.2. In this latter model, the increases in DE relative to the fifth quintile (stocks with the largest cap-sizes) were 43%, 21%, 11% and 8% respectively for quintiles 1-4. It was important in this case to check the SSR plots, as they showed that the effect for all quintiles does not really take effect until 100 days into the follow up time for a position. This is evidence that investors consider different characteristics of stocks they have positions in at different times when deciding whether to sell the position or not.

Smaller cap-size stocks tend to be more volatile and Kumar (2009b) hypothesises that more volatile stocks are harder for investors to value, which makes them more likely to exhibit a DE when trading small cap stocks. So one possibility is that the volatility of a stock only begins to affect an investor's decision making after they have experienced it first-hand and seen the fluctuations in price as they hold it. Unlike cap-size, the industry group of the stock being held did not have an effect on the DE. There were differences between groups in terms of their hazard of being sold, which merited this variables inclusion as a main effect. But the interactions with the gain indicator were not significant, even when tested in a model without interactions for any other variables.

As expected, the DE was much weaker during December months, being reduced by half compared to other times of the year. The SSR plot showed that the effect increased dramatically over time, from -0.2 (hazard ratio = 0.82) to almost -1.2 (hazard ratio = 0.3) by the end of the follow-up time. This means that, when wanting to sell a losing position in December for tax purposes, investors sell 'old' losses at a greater rate than losing positions they have not held for very long. One explanation is that investors are more optimistic about the possibility of a price reversal in positions that are still relatively new. This result adds nuance to the understanding of the December effect on the DE, and provides evidence that the holding period of a position is a factor, conscious or otherwise, when investors are deciding whether to sell it or not. The result also raises questions that cannot be answered by this model. It is not possible to tell whether investors are actively choosing to sell older losses since this would require knowing if they have a younger loss available to sell as well. It may also be the case that investors prefer to sell positions that are further from the purchase price when they have a choice, and older positions are more likely to have fallen further in price.

In this dataset the DE is significantly lower in years 1995 and 1996, with the hazard ratio indicating an average reduction of 24%. The SSR plot shows that the effect is stronger at the start of the follow-up time i.e. for stocks that have only recently been purchased, and weaker towards the end, but it stays close enough to the parameter estimate such that the hazard ratio is still a reasonable summary of the effect. These results establish that the fall in aggregate DE during these years, as observed in Kumar (2009b), cannot be entirely explained by other factors, such as the characteristics of investors trading in these final two years of the data period, since they are controlled for in the model. The decrease in DE is therefore likely due to changes in market conditions during the data period, as discussed in section 1.5.1.

Chapter 3

Using a mixed-effects logistic model to analyse the determinants of lottery stock purchases

3.1 Introduction

Beginning with Odean (1999), studies of individual investor trading records have found that many investors earn poor returns on their trades even before transaction costs have been accounted for. This suggests that these investors have poor stock selection ability. In a dataset from Taiwan, Barber et al. (2008) find that, in aggregate, individual investors underperform the market portfolio by 3.8 percentage points annually. They attribute roughly one third of this gap to poor stock selection ability, with the rest being due to commissions and tax. Using the LDB data, Korniotis and Kumar (2013) identify low cognitive ability investors using demographic information and find that they underperform the market by 3.6% annually. The authors attribute half of the shortfall to stock selection ability and half to transaction costs.

Due to the welfare implications of adverse performance, understanding the process by which investors choose which stocks to purchase is very much of interest. Kumar (2009a) identifies one particular category of stock that is over-represented in the portfolios of investors in the LDB dataset relative to the market portfolio, which he calls lottery stocks. These are stocks that are low in price and have returns distributions that are high in volatility and positive skewness. This builds on the cumulative prospect theory of preferences, introduced by Tversky and Kahneman (1992), which predicts investors will prefer skewed payoff distributions since they overweight the small probability of a large gain.

Kumar identifies demographic groups with a stronger preference for lottery stocks using an investor level regression and a measure of preference that aggregates over the whole data period. This chapter extends this approach by estimating a logistic regression model at the level of individual stock purchases, with the response being the odds of a purchase being of a lottery stock. This allows the inclusion of both static demographic covariates and time-varying information about the investor's portfolio and behaviour. Due to the likely presence of correlation between purchases at the investor level, a mixed effects model is used in order to control for each investor's idiosyncratic preference for lottery stocks.

Of the demographic covariates included, only age and income had a significant effect on the odds of a lottery stock being purchased when other factors were controlled for. Increases in these variables tended to lower the odds of a lottery stock purchase, although a non-linear effect was found for age. The odds were lower for more well diversified investors and those who traded less frequently, and higher for investors who had displayed a stronger preference for lottery stocks in the past. Whilst the return of the investor's recent lottery sales and the paper return of their current lottery positions did not have much explanatory power, stronger recent performance of lottery stocks in the market was found to significantly increase the odds of lottery stock purchases. A model with interaction terms showed that this latter effect was weaker for more well diversified investors.

This chapter is organised as follows. Section 3.2 motivates the study of lottery stocks with reference to cumulative prospect theory, and introduces Kumar's formal definition of them. Section 3.3 describes the purchase-level logistic regression model that will be used for the analysis and discusses the importance of controlling for investor-level correlation using random effects. Section 3.4 describes the covariates that will be included in the model. Section 3.5 presents the main results and provides a detailed interpretation of them. Section 3.6 presents models that supplement the main results and checks their robustness. Section 3.7 provides a summary of the chapter's findings.

3.2 Stocks as lotteries

One important deviation from expected utility theory that has been observed in a variety of settings is the tendency for people to overweight small probabilities. Building upon their earlier work, Tversky and Kahneman (1992) proposed a model for preferences, called cumulative prospect theory (CPT), which incorporates this idea of probability weighting. They introduce a weighting function which is applied to the probabilities of risky gambles that the investor faces.

Barberis and Huang (2007) note that CPT preferences predict a preference for positively skewed payoff distributions, and extend Tversky and Kahneman's model to a setting where the investor must assess a continuous payoff distribution rather than discrete. A positively skewed payoff distribution will be highly desirable to the investor since they will overweight the small tail probability of a very large payoff in this distribution. In a simple market of financial assets with CPT investors and payoffs which are normally distributed, the authors show that a new asset with positively skewed payoffs can become overpriced and have a negative excess return as a result. Since CPT investors would like the payoff distribution of their whole portfolio to be positively skewed, the skewed asset is very useful to them, and hence are willing to pay a large premium to hold it. Compared to the standard scenario of maximizing risk-adjusted return, CPT preferences can produce a more 'lottery-like' approach to investing, where investors prefer to have a small chance of a very large gain in exchange for a high chance of moderate losses.

Henderson et al. (2017) solve an asset liquidation problem in continuous time for an investor who has both the value function and probability weighting components of CPT. In agreement with Barberis and Huang (2007) they find that the optimal prospect for such an investor is indeed positively skewed. They note that probability weighting discourages the investor from selling when their asset is trading at a gain, since they are overweighting the probability that the price will increase to an even higher level. Importantly however, their model is able to predict both a preference for positively skewed assets and a disposition effect that matches the magnitude found by Odean (1998)[1] for realistic values of the CPT parameters. This is because the desire to wait for higher gains resulting from probability weighting is offset by the desire to realize gains that results from the shape of the value function.

Using the LDB data, Mitton and Vorkink (2007) find that investors who hold poorly diversified portfolios are able to achieve a much higher level of positive skewness in the returns distribution of their portfolios than well diversified investors. As a result of this, many of the investors who achieve the highest return over the course of the data period hold poorly diversified

[1] i.e. that opportunities to sell for a gain are taken at a rate that is roughly 50% greater than the rate at which opportunities to sell for a loss are taken.

portfolios. The authors also document a consistent trade-off between the Sharpe ratio (excess return divided by standard deviation) and skewness of investor portfolios, implying that some investors sacrifice risk-adjusted returns in order to increase the positive skewness they are exposed to.

Kumar (2009a) extends the idea and formally defines a category of stock that he refers to as lottery stocks. With reference to the key features of actual lotteries, he argues that the stocks with low price, high idiosyncratic (positive) skewness and high idiosyncratic volatility are most likely to be seen as lottery-like by investors. Lottery tickets have a very low price relative to the potential payoff, hence stocks with a low price are more amenable to being treated like lotteries, particularly for the many investors in the LDB dataset who invest relatively small amounts at a time. Stocks with higher volatility will be more attractive since they increase the chance of an extreme return.

3.2.1 Empirical definition

Kumar's definition of lottery stocks is as follows. For a particular month t, he labels each stock as either a 'lottery', 'non-lottery' or 'other' type stock, based on the previous six months of daily returns data, $t-1$ to $t-6$, which is taken from the CRSP database. The volatility and skewness are calculated using this set of daily returns using methods described below, and the price is taken as the closing price at the end of month $t-1$. This forms empirical distributions for the three components in month t across all stocks in the CRSP database. A stock is in the lottery category in month t if it is in the top half of the distribution for both volatility and skewness, and in the bottom half of the distribution for price. A non-lottery stock is in the top half of the distribution for price, and in the bottom half for both volatility and skewness. All remaining stocks are in the 'other' category. Lottery stocks are the primary focus of the analysis conducted in this chapter, but comparisons will be made with the non-lottery category in order to provide a point of reference for the results.

The volatility of a stock in month t is the volatility of the residuals obtained from fitting a four-factor model to the daily excess returns of the stock in months $t-1$ to $t-6$, where the factors are:[2]

- **Excess market return**
- **SMB**: The historic excess return of small (in terms of market capitalization) stocks over big
- **HML**: The historic excess return of stocks with a high book-to-market ratio over those with a low ratio
- **Momentum**: The historic excess return of stocks which have performed well in the recent past, relative to those that have not

Following the method of Harvey and Siddique (2000), the skewness is that of the residuals obtained from fitting a two-factor model to the same returns series, where the factors are the excess market return and its square. These variables were obtained from Kenneth French's data library [3].

3.2.2 Past findings

Defined in this way, Kumar finds that lottery stocks make up 1.25% of the market portfolio[4] during the period 1991-96, but make up 3.74% of the aggregate portfolio held by investors in the LDB dataset. In contrast to this, only 0.76% of weight is assigned to lottery stocks in the aggregate portfolio of

[2] The excess return is the return minus the return on one-month U.S. treasury bills over the same period.

[3] `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[4] A portfolio where stocks are weighted by their market value as a proportion of the total value of the market.

institutional investors. Using a variety of measures, Kumar finds that lottery stocks under perform the other two stock categories by at least 4 percentage points annually, and that on average, LDB investors would have achieved 2.84 percentage points higher annual returns if they had replaced the lottery component of their portfolio with the non-lottery component.

Part of the initial work on lottery stocks by Kumar (2009a) was the estimation of a regression model for the lottery stock preference of individual investors. Preference is calculated for each investor at the end of every month, and the average across all months is used in the model.[5] This model provides estimates for the importance of different characteristics of investors whilst controlling for other factors, such as how well diversified their portfolios are on average. He finds that investors who are poor, young, less-educated, have a non-professional job and live in an urban area have a stronger preference for lottery stocks. Importantly, these groups have also been found to spend more on actual lotteries in the literature on state lottery participation.[6]

However, using each investor's preference averaged over the whole data period precludes the inclusion of variables that are time-varying, for example the recent performance of lottery stocks in the market, or information about the current state of the investor's portfolio. Kumar also estimates a time series model for the aggregate lottery stock preference across all investors, including information about the market and wider economy as covariates. But this has the reverse problem of losing the ability to describe differences between individual investors.

[5] Kumar uses five different measures of monthly preference, but finds the model results are similar in each case.

[6] See the references provided by Kumar for details.

3.3 Modelling approach

This chapter will seek to combine these approaches and estimate a single model that will allow comparison between static and time-varying covariates, in terms of their effect on an investor's decision to purchase a lottery stock. There are two aspects to the decision to buy a stock: which stock to buy and when to buy it. The decision to buy a lottery stock versus another type of stock is much more of interest than the decision to buy a stock on a particular day versus another. These two aspects are clearly related; investors in general will not first decide to buy a stock and then separately decide which one to buy. Their decision to purchase or not depends on the price, and other factors, and their evolution over time. But to simplify the analysis and maximize the amount that can be deduced about an investor's decision between stocks, the likelihood of a lottery stock purchase will be modelled conditional on a purchase having been made.

This can be framed as a prediction problem: when an investor buys a stock, can the type of the stock be predicted using only information that could be known in the instant before the purchase was made. Logistic regression is a natural fit for this problem, where purchases are categorised as either being of a lottery stock or not. Time-varying covariates can be included by recording their state at the time of each purchase. This formulation will allow the comparison of these time-varying factors with static things such as an investor's gender or occupation in terms of their effect on the investor's decision to buy a lottery stock.

Since logistic regression is closely related to proportional hazards (PH) models, as used in chapter 2, it is worth discussing why a PH model is not applicable here. In a PH model, the observational unit is a lifetime, which in chapter 2 was the holding period of a stock from purchase until sale. When a position is sold, it is compared to positions that had been held for the same period of time but had not yet been sold. For lottery stocks, a lifetime could be defined as ending when the investor buys a lottery stock. But there would be no clear start point for this lifetime. It could be the time of the investor's last purchase of a lottery stock, their last sale or the start of the data period. But purchase lifetimes that are at the same point on one of these timescales do not have anything obvious in common when considering the likelihood of the investor making a lottery stock purchase. Hence imposing this additional structure on the data when it is not needed would only hinder learning about the reasons why an investor chooses to buy a lottery stock.

3.3.1 Investor-level correlation

As with the PH model used in chapter 2, the issue of investor-level correlation must be considered here too. It is reasonable to think that the propensity to purchase a lottery stock could differ substantially between investors, and in ways that cannot be captured by the information that is available in the dataset. Similarly to position sales, the distribution of purchases across investors is extremely imbalanced. In the LDB dataset there are 55,052 investors who purchase at least one common stock and 976,934 common stock purchases in total. The median number of purchases per investor is 7, whereas at the extreme end of the distribution there are 8 investors with over 1000 purchases and the top 10% of investors in this distribution account for 52% of all purchases. Clearly, if this grouping at the investor-level was ignored then the results of a model where the observational units were individual purchases would be biased towards the behaviour of these most active investors.

The solution that will be used here is to introduce investor-level latent variables into the model that will control for the component of each investor's idiosyncratic preference for lottery stocks that is not captured by the model covariates. These latent variables are called random effects, and their inclusion places the model in the family of mixed models i.e. a model including both fixed effects (the model covariates) and random effects. Specifically, since a logistic regression is being used, the model is a generalized linear mixed model (GLMM). The random effects are the direct analogue of the frailties used in chapter 2. Here they model an investor's propensity to purchase a lottery stock, conditional on them having made a purchase, and in the aforementioned chapter they modelled an investor's propensity to sell a position that they are currently holding. Including the investor grouping factor as a random effect rather than a fixed effect makes sense since the goal is to control for the unobserved heterogeneity between investors, rather than test hypotheses about differences between specific investors.

3.3.2 Mixed-effects logistic regression

In a mixed-effects logistic model, the probability of the i-th purchase by the j-th investor being of a lottery stock is denoted as p_{ij} and defined by

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta^{\top} X_{ij} + Z_j$$

where X_{ij} is the row of the design matrix (containing an intercept and the covariates) corresponding to the purchase, β is the vector of coefficients and Z_j is the investor-specific random effect. The random effects are assumed to be drawn from a normal distribution with $\mathbb{E}(Z) = 0$ and $Var(Z) = \theta_Z^2$, with estimation of this variance being an important part of the model fitting process.

Maximum likelihood estimation of the parameters in β and of θ_Z^2 involves integrating the random effects Z out of the joint likelihood. This integral does not have a closed form solution and hence analytical methods for likelihood maximisation cannot be used. Different approaches to the problem are discussed in Tuerlinckx et al. (2006). In the 'lme4' package (Bates et al.,

2015) in R, the integral is approximated using the Laplace method. This allows maximisation of the approximate marginal likelihood via the use of a standard nonlinear optimizer. In the analysis presented below the BOBYQA algorithm, introduced by Powell (2009), is used, which has the advantage of not requiring derivatives of the function to be computed.

As discussed in Bolker et al. (2009), the Wald statistic is preferred for testing the significance of fixed-effect parameter estimates. A LRT comparing models with and without a random effect is recommended for determining whether the random effect is necessary. An adjustment to the critical value for the test statistic is needed since the null value of the parameter being tested is on the boundary of its allowable range. Zhang and Lin (2008) show that in the simple case of testing the significance of a single random effect, the asymptotic null distribution of the test statistic is in fact a 50:50 mixture of χ^2_0 and χ^2_1.

3.4 Covariates

The key advantage of using the individual buys as the observations in a logistic regression is that time-varying factors can be included in the model alongside static factors that were known at the start of the data period. The covariates that will be included in the analysis can be divided into these two categories. Note that time-varying covariates are recorded at the start of the day on which the purchase occurs. This is because the dataset only records the date on which a trade occurred, so the ordering of trades made on the same day is unknown. This section will describe each covariate that will be included in the analysis, along with justification for their inclusion and some summary statistics.

3.4.1 Time-varying covariates

A natural hypothesis is that having recently sold a lottery stock for a gain would encourage the investor to purchase another stock that is similar i.e. also in the lottery category. The mean excess return of lottery stock sales made by the investor in the year to date was used to test this, with each return being weighted by the dollar value of the sale.[7] An indicator for whether the investor had sold any lottery stocks in the past year was also included in order to differentiate between investors who had sold lottery stocks for a return close to or exactly zero, and those who had not sold any. For 30% of all purchases, the investor had sold at least one lottery stock in the year to date.

Similarly to the return of their recent sales of lottery stocks, strong performance of lottery stocks the investor currently holds may encourage them to purchase more. The value-weighted mean return of any currently held lottery stocks is included as a covariate, along with an indicator for whether the investor has any such holdings. The lottery category of the stock at the time the investor purchased it is used, rather than the possibly different category at the time the current return is being calculated. The investor currently holds a lottery stock for 32% of purchases.

Rather than encouraging an investor to purchase lottery stocks in general, a recent positive experience with a lottery stock may encourage the investor to repurchase the exact same stock. Repurchases are a common occurrence, accounting for 36% of all purchases in the dataset. This issue will be addressed in section 3.6.3 by re-estimating the model without repurchases, and also exclusively with repurchases

As well as recent good performance of stocks they own, investors may also

[7]For the return of each sale, the return of the market during the holding period of the position was subtracted, and the value-weighted mean for these quantities was then calculated.

	% of purchases
Lottery sale in past year	30.3%
Currently holds lottery stock	32.3%

Table 3.1: Percentage of stock purchases where the investor making the purchase had sold at least one lottery stock in the year prior to the purchase, and the percentage where the investor held at least one lottery stock at the time of the purchase.

	Mean	Median	IQR
Return of lottery sales in past year	15.1	7.7	36.7
Paper return of current lottery stock positions	15.7	0.1	36.2
Market return of lottery stocks in previous month	4.8	4.3	5.5

Table 3.2: Summary statistics for the time-varying retuns covariates, calculated across purchases. This includes the value-weighted mean return of the investor's lottery sales in the past year, conditional on at least one such sale having been made; the value-weighted mean return of the investor's currently held lottery stock positions, conditional on them having at least one such position; and the value-weighted mean return of all lottery stocks in the market in the previous calendar month

be more likely to purchase lottery stocks if they are performing well in the market. To test this, the mean excess return of lottery stocks in the previous calendar month, weighted by their market value, was included in the analysis. In section 3.6.1 this variable is interacted with others to test whether different groups are effected by it to different extents. Summary statistics for these indicator and return variables are presented in tables 3.1 and 3.2 respectively.

The next set of covariates contain information about the current state of the investor's portfolio and their behaviour since the start of the data period. The logarithm of the current number of positions the investor holds is included as a measure of how well diversified their portfolio is. The proportion of their total portfolio value that is due to their largest stock holding is included as a measure of portfolio concentration. Better diversified investors who hold less concentrated portfolios are expected to be less likely to purchase lottery stocks. In terms of the investor's behaviour, the number of actions (buy or sell) they have made so far divided by the number of days since the start of the data period is included to control for the investor's general level of activity. How frequently an investor trades is an important component of their investing style i.e. whether they are trading actively or following a more passive 'buy and hold' strategy. Some summary statistics for these variables are presented in table 3.3.

	Mean	Median	IQR	Transform
Lottery buy proportion	0.1	0.0	0.1	Square root
Number of current positions	11.0	7.0	11.0	Log
Number of days per action so far	56.0	24.1	48.6	Log

Table 3.3: Summary statistics for the time-varying covariates containing information about the investor's behaviour, calculated across purchases. Includes the proportion of the value of purchases made so far that the investor spent on lottery stocks, the number of positions the investor currently holds and the number of actions they have taken so far (buy or sell) divided by the number of days since the start of the data period. Also describes the transformation applied to each variable before being entered into the model.

The investor's preference for lottery stocks in the past is likely to be a strong predictor of their preference in the future. By controlling for this, the effects of other variables will describe differences between investors who have displayed a similar preference for lottery stocks in the past, which is a more interesting comparison. An investor's past preference for lottery is measured using the proportion of the total value of all purchases they have made so far that is due to purchases of lottery stocks. A square root transformation was applied to this variable, as suggested by the plot in figure 3.2. Finally, calendar year dummies were included for 1992-96, with 1991 being the reference category. The aggregate preference for lottery stocks across all investors in the sample increases substantially over time, as shown in figure 3.1. Including these dummy variables will isolate the effect of other covariates from this time trend. For example, the number of positions an investor currently holds will tend to be larger in later years since there are more buys than sells in the sample.

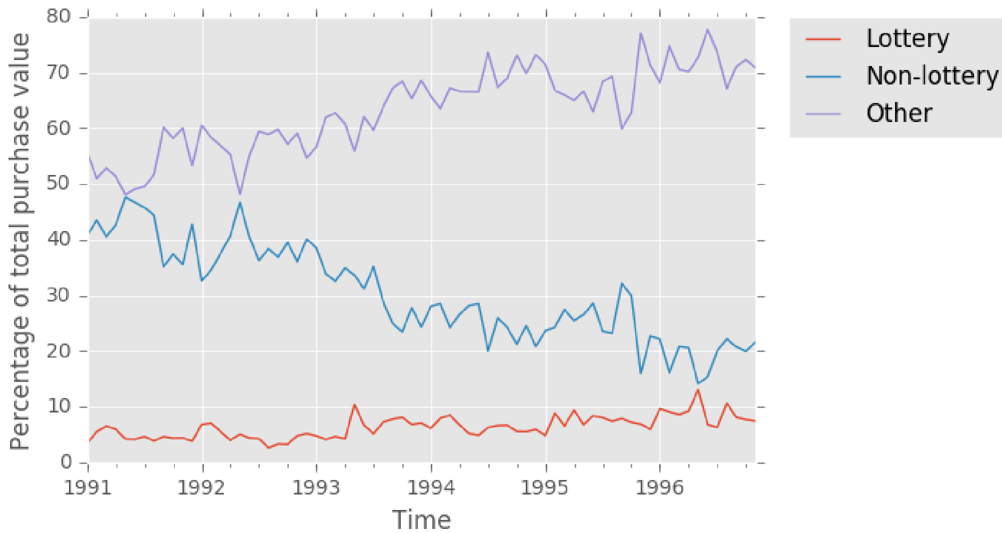

Figure 3.1: Monthly time series of aggregate preference for each of the three stock categories: lottery, non-lottery and other. The aggregate preference for a category is the percentage of the total value of all purchases made during the month that is due purchases of stocks in that category.

3.4.2 Static covariates

Based on the results of Kumar's cross-sectional analysis, a variety of demographic variables will be used. This includes indicators for the investor's gender, their marital status, whether they are retired and whether they have a professional occupation or not. Age was included along with its square, as a smoothed plot of the response (on the logit scale) against age suggested a non-linear effect that could be accounted for by the addition of a squared term. Kumar finds that younger, single, male investors who have non-professional occupations have a stronger preference for lottery stocks.[8] Some summary statistics for these categorical variables are shown in table 3.4.

	% of investors	% of purchases
Female	7.0	6.1
Male	93.0	93.9
Non-professional	36.7	37.6
Professional	63.3	62.4
Non-retired	84.9	83.2
Retired	15.1	16.8
Married	80.8	82.2
Single	19.2	17.8

Table 3.4: Distribution of static categorical covariates across investors and purchases

The natural logarithm of income will also be included, with investors who have lower incomes expected to have a stronger preference for lottery stocks. As well as the number of stocks the investor holds at the time of the purchase, as mentioned in the previous section, the number they held at the beginning of the data period is also included. This latter variable is referred to as the investor's initial diversification. Both variables being significant in the model will indicate that deviations away from the investor's initial diversification have some predictive power for lottery stock purchases. Also included is the time in years from the date the investor opened their first account at the brokerage until the start of the data period, which is referred to as the investor's tenure. Negative values occur for investors that opened their account after the start of the data period. The results in chapter 2 suggested investors who opened an account most recently may actually be the most sophisticated, contrary to what was expected. Including it here will provide more evidence on the issue. A summary of these continuous static covariates

[8] Whilst age and some of the other covariates in this section can (or do, in the case of age) change over time, all of them were recorded only once in the dataset, and as such are entered into the model as they were at that point in time.

is presented in table 3.5

	Mean	Median	IQR	Transform
Age	50.9	48.0	16.0	With square term
Income ($)	171100.0	75000.0	62500.0	Log
Tenure (years)	4.0	3.8	5.5	No transform
Initial diversification	5.3	3.0	4.0	Log

Table 3.5: Summary statistics for static continuous covariates. Calculated across investors, rather than stock purchases. Also describes the transformation, if any, applied to the variable before being entered into the model. Initial diversification is the number of stocks the investor held at the start of the data period.

In the LDB dataset there are 976,934 stock purchases where there is sufficient data in the CRSP database to categorise the stock as either lottery or not at the time of the purchase. However, as with the model in chapter 2, missing data amongst the demographic covariates means that only a relatively small proportion of the sample can be used to estimates a model that contains these variables. Age, gender, income and marital status information is missing for 40-55% of purchases. Since different variables are missing for different investors, collectively this means that only 92,975 have complete information for the model containing all covariates mentioned in this section. This is still a large sample, particularly for a mixed-effects model of the kind described in section 3.3.2. So if they exist, it should be possible to detect even small effects. As was done in chapter 2, a model without demographic covariates was also estimated using a much larger sample, to see if the estimated effects of the other covariates are materially different. This is the subject of section 3.6.4. Of the remaining purchases in the sample with full demographic information, 9.1% are purchases of lottery stocks.

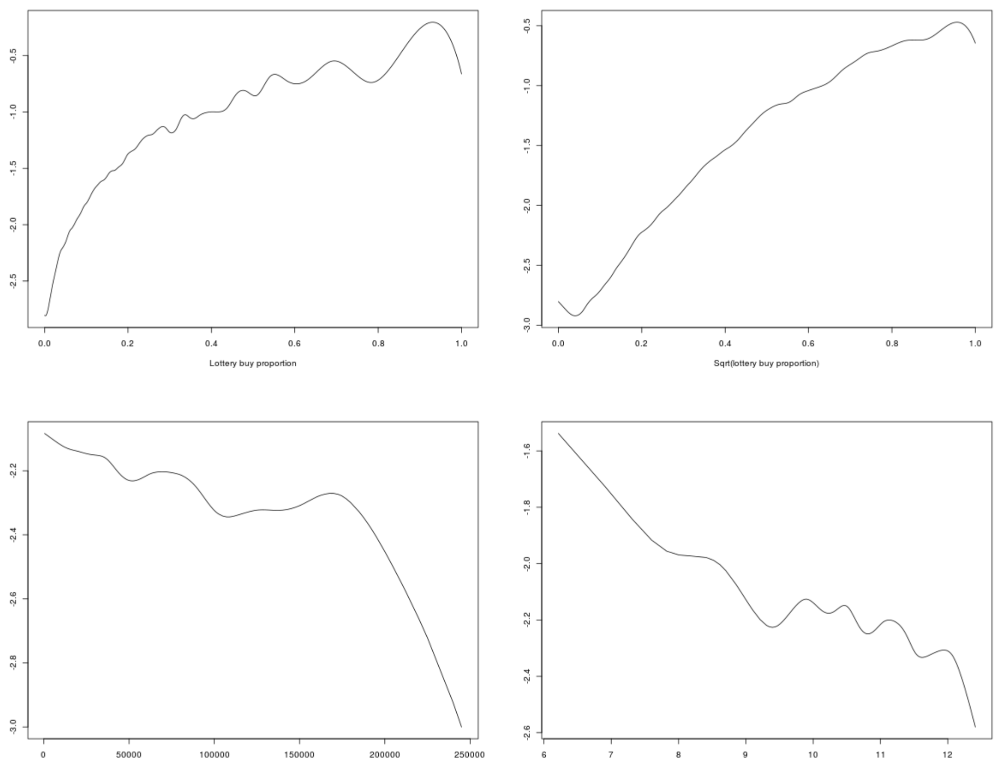

Figure 3.2: Cubic smoothing spline fits for the response variable (1 if the buy was of a lottery stock and 0 otherwise) plotted on the logit scale against covariate values. The regression model assumes a linear relationship on this scale, hence the plots for the transformed variables on the right demonstrate closer adherence to this assumption.

3.5 Model estimation results

3.5.1 Testing significance of the random effect component

Table 3.6 presents regression results for the mixed effect logistic model containing all the covariates described in section 3.4. The first step in interpreting this model is to establish that the random effect component improves the fit of the model. With reference to Bolker et al. (2009), the preferred method is to use a LRT comparing the model with the corresponding GLM i.e. the same model without random effects. This produces a test statistic of 985.82, which is highly significant (p-value $\ll 0.001$) compared to the reference distribution under the null of no significant difference, which is a 50:50 mixture of the χ_0^2 and χ_1^2 distributions. The estimate for the random effect standard deviation is 0.781, with a 95% profile likelihood confidence interval given by (0.723, 0.841). These results provide strong evidence that a model including investor-level random effects is supported by the data.

3.5.2 Residuals

As is standard in regression analysis, the adequacy of the estimated model can be assessed using a residual quantity. For GLMs where the response can only take on a small number of values, the residual produced by subtracting the fitted value from the response is also restricted to the same small number of values. This limits the amount of variation that is visible in standard diagnostic plots of the residuals and hence makes it difficult to identify model inadequacies. As an alternative, the randomized quantile residuals (RQRs), introduced by Dunn and Smyth (1996), will be used to assess the model. The idea of these residuals is to add random noise to the value of the theoretical cumulative distribution function (CDF) of each observation, which produces a set of residuals that can be treated as continuous. If the model is true then these residuals should constitute an i.i.d. sample from the standard uniform distribution, a property that can be easily tested.

Let $F(y_{ij};\mu_{ij})$ be the CDF of the i-th purchase made by the j-th investor, where $\mu_{ij} = E(y_{ij})$. μ_{ij} is a function of the covariates X_{ij}, the parameters β and the random effect Z_j. Let $a_{ij} = \sup_{y<y_{ij}} F(y, \hat{\mu}_{ij})$ and $b_{ij} = F(y_{ij}, \hat{\mu}_{ij})$, then the RQR for purchase ij is defined by

$$r_{ij} = \mathcal{U}(a_{ij}, b_{ij})$$

which denotes a uniform random variable on the interval $(a_{ij}, b_{ij}]$. If the model is true, meaning each observed y_{ij} was in fact generated by the distribution defined by $F(y_{ij};\mu_{ij})$, then these residuals will constitute a sample from the standard uniform distribution, $\mathcal{U}(0,1)$.[9]

For a logistic model this simplifies to

$$r_{ij} = \begin{cases} \mathcal{U}(0, 1-\hat{p}_{ij}) & \text{if } y_{ij} = 0 \\ \mathcal{U}(1-\hat{p}_{ij}, 1) & \text{if } y_{ij} = 1 \end{cases}$$

where $\hat{p}_{ij}$ is the fitted value for purchase ij. A QQ plot can be used as a first check of whether the residuals have the expected distribution. For this purpose, Dunn and Smyth recommend converting the residuals to quantiles of the standard normal distribution by applying the inverse CDF of the standard normal distribution, as QQ plots for the standard normal distribution are more familiar. A normal QQ plot for one realization of the RQRs is shown in figure 3.3. On the evidence of this plot, the residuals adhere very closely to the expected distribution.

To further check the integrity of the model, the residuals can be plotted

[9] See Feng et al. (2017) for a proof of this.

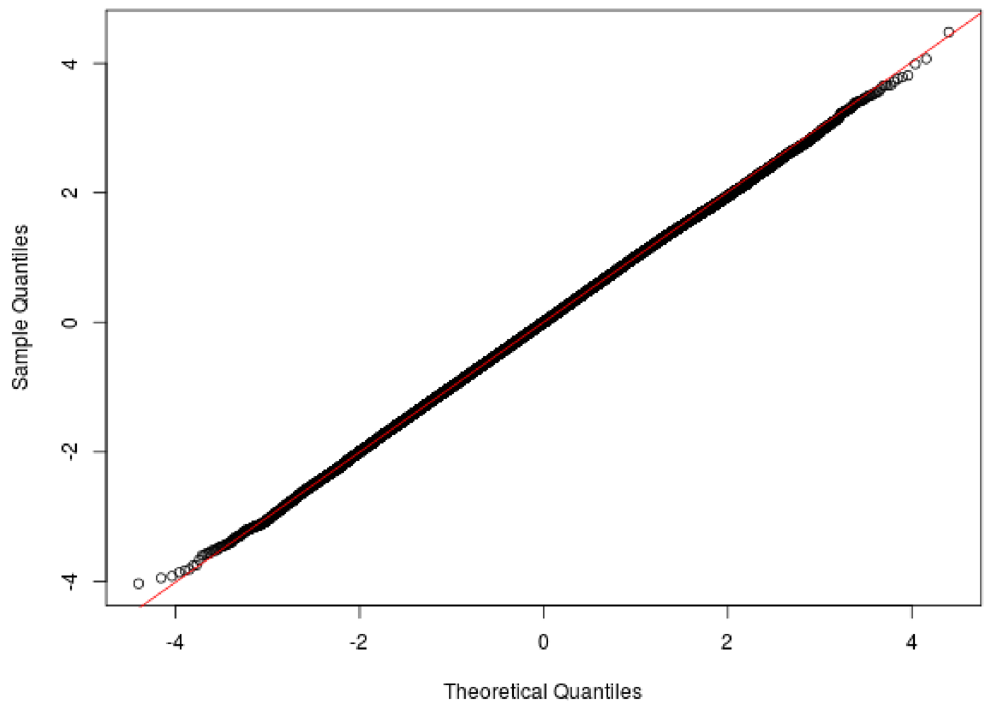

Figure 3.3: QQ plot comparing a realization of the RQRs for the mixed-effects logistic model with the standard normal distribution. The RQRs have been converted into quantiles of the standard normal distribution by applying the inverse CDF of this distribution to them.

against the fitted values. The original RQRs, which should be uniformly distributed, are recommended for this plot since it is easy to assess whether the mean is close to the expected value of 0.5 at all magnitudes of the fitted values. Due to the random element inherent to the process, Dunn and Smyth recommend generating four realizations of the residuals and discounting any apparent patterns that are not common to all of them. Figure 3.4 contains plots against the fitted values for four such realizations. The first, in the top left of the plot, is the same realization as was used for the QQ plot in figure 3.3. A smoothing spline has been added to each plot, which can be compared to the horizontal line indicating the expected value of 0.5. These plots reveal a clear tendency for the residuals to be larger than expected for purchases with large fitted values. This means that, amongst this group of purchases, there are more lottery purchases than would be expected if the model was true. This suggests the model is missing a factor that could explain the very strong preference for lottery stocks of some investors.

For comparison, figure 3.5 contains four realizations of the RQRs for the corresponding model without random effects. The same pattern is visible in these plots, and to a much greater extent. This provides reassurance that the model is significantly improved by the inclusion of random effects, and hence also that the deviation from uniformity in the scatter plots for the mixed-effects model is small compared to what results from a major misspecification of the model. So whilst these plots suggest the current model could be improved, the pattern in the residuals is not severe enough to invalidate the conclusions drawn from the model results. The analysis can therefore proceed to a detailed interpretation of these results, which is the focus of the next section.

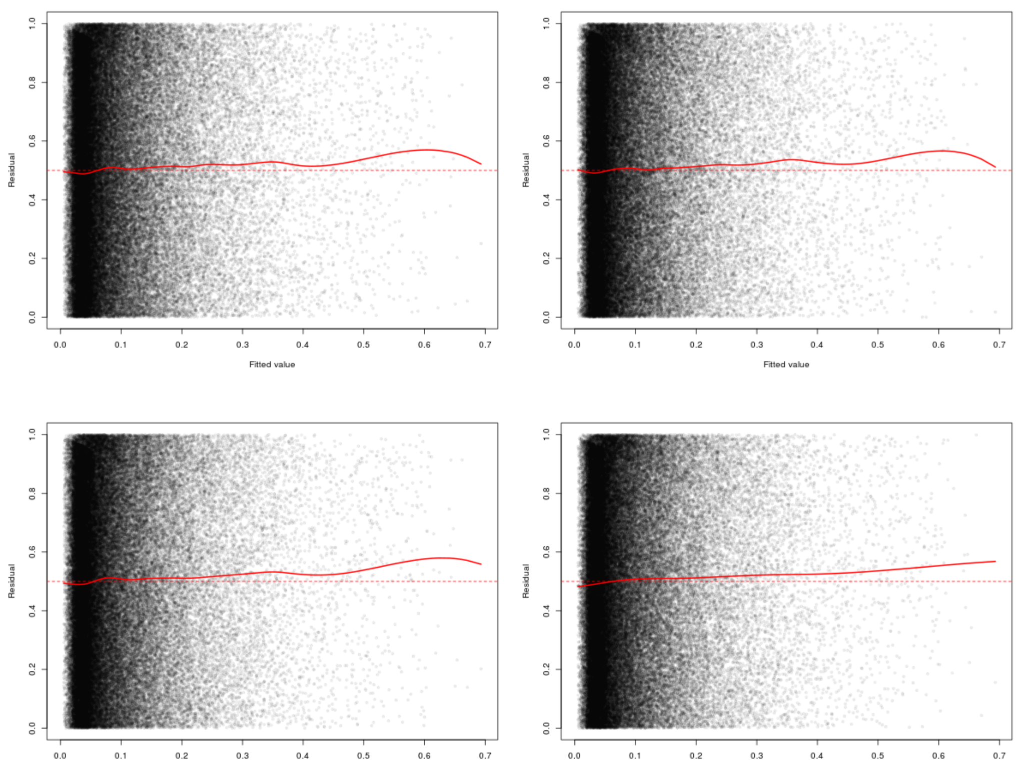

Figure 3.4: Four realizations of the RQRs for the mixed-effects logistic model, plotted against the fitted values of the model. A smoothing spline has been added to each plot to help identify deviations from the expected value of 0.5.

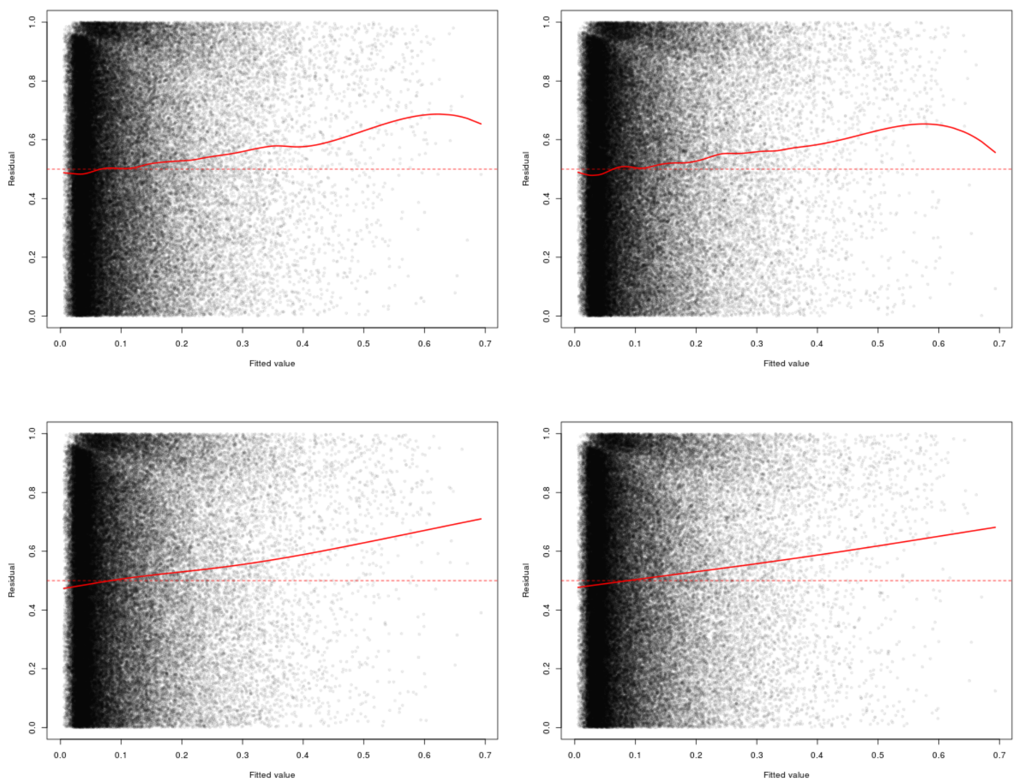

Figure 3.5: Four realizations of the RQRs for the logistic model without random effects, plotted against the fitted values of the model. A smoothing spline has been added to each plot to help identify deviations from the expected value of 0.5.

3.5.3 Interpretation of full model

As mentioned in section 3.4, many of the continuous covariates were centred when entered into the model (mean subtracted and divided by standard deviation) as this is known to help with convergence in GLMMs.[10] Other transformations, such as the natural logarithm and square root, were applied to some variables prior to them being centred in order to bring the relationship with logs odds closer to being linear. This means that the effects of these variables is not obvious from their coefficient estimates in table 3.6, and care needs to be taken to reverse the transformations in order to produce a meaningful interpretation. Two things in table 3.6 that are informative are the sign of the coefficient estimate, with a positive value indicating that a purchase being of a lottery stock becomes more likely as the covariate increases, and the Wald test p-values, which can be interpreted in the normal way. The remainder of this section will discuss these regression results in detail.

Throughout the following, covariates will be described as either raising or lowering the odds of a lottery stock purchase as their values change. As was discussed in section 3.3, the model is actually estimating the effect of covariates on the odds of a particular purchase being of a lottery stock, conditional on that purchase being made. This less precise language is used purely for convenience and the conditional nature of the model should be kept in mind.

Time-varying information about the investor's portfolio and trading history

An investor's past preference for lottery stocks is a strong predictor of their future preference, as shown by the lottery buy proportion variable, which is

[10] See for example the 'convergence' section of the lme4 documentation.

	Coefficient	SE	p-value
Intercept	-3.355	0.104	< 0.001
Lottery sale in past year	-0.024	0.036	0.516
Return of lottery sales in past year	0.027	0.012	0.026
Currently holds lottery stock	0.365	0.035	< 0.001
Paper return of current lottery stock positions	-0.006	0.014	0.64
Portfolio concentration	0.142	0.071	0.045
Log(Number of current positions)	-0.210	0.033	< 0.001
Log(Number of days per action so far)	-0.136	0.032	< 0.001
Sqrt(Lottery buy proportion)	1.083	0.098	< 0.001
Return of lottery stocks in previous month	0.085	0.013	< 0.001
Gender: Male	0.071	0.086	0.413
Age	0.396	0.163	0.015
Squared(Age)	-0.490	0.167	0.003
Marital status: Single	0.024	0.054	0.653
Log(Initial diversification)	-0.019	0.027	0.487
Tenure	-0.015	0.022	0.496
Log(Income)	-0.090	0.022	< 0.001
Professional occupation	0.056	0.052	0.281
Retired	0.094	0.079	0.234
Year: '92	-0.019	0.047	0.683
Year: '93	0.210	0.050	< 0.001
Year: '94	0.338	0.057	< 0.001
Year: '95	0.563	0.058	< 0.001
Year: '96	0.657	0.061	< 0.001

Table 3.6: Results from estimating a logistic regression model with investor-level random effects for lottery stock purchases.

the proportion of the total value of the investor's purchases made so far that was spent on lottery stocks. Exploratory analysis suggested a square root transformation should be applied to this variable, and as such the effect of a change in it on the log odds depends on the specific values as well as the difference between them since the transformation is non-linear.

Table 3.7 provides some examples of changes in lottery buy proportion and their effect on the odds of a purchase being of a lottery stock. The effect of this variable is large compared to others in the model. An investor for whom lottery stocks have made up 90% of their purchase value so far is almost twice as likely to purchase a lottery stock than an investor for whom 10% of their purchase value has been of lottery stocks. That this covariate is an important predictor of lottery stock purchases is expected, and shows that it is fulfilling its primary purpose in the model, which is to control for the investor's past preference when estimating the effect of other, more interesting variables.

Increase in buy proportion	% increase in odds
$0.25 \to 0.5$	25.2%
$0.5 \to 0.75$	19.0%
$0.1 \to 0.9$	98.4%

Table 3.7: Change in odds for some example increases in the proportion of the total value of their purchases to date that an investor has spent on lottery stocks.

One of the main hypotheses in this analysis was that positive past experience with lottery stocks would encourage an investor to purchase more in the future. To test this, the excess mean return of any lottery stock sales the investor made in the previous year was included in the model, along with an indicator for whether the investor did indeed sell any lottery stocks in the previous year or not. The coefficient estimate for the indicator variable was not significant, but was for the mean return, although only at the 5% level which is fairly weak compared to many of the other effects in the model.

Reversing the scaling on the coefficient reveals that an increase in return of 10 percentage points only increases the odds of a lottery stock purchase by 0.7%, and an increase of 30 percentage points (which is the IQR across all purchases) increases it by 2.3%. So whilst these results show that a greater return on recent lottery stock sales increases the chance an investor will buy a lottery stock in future, the effect is small compared to others in the model.

Also of interest was the mean paper return of lottery stocks positions that the investor currently holds. As well as this return, an indicator for whether the investor did hold any lottery stocks at the time was also included. In this case, the estimate for the return variable was not significant, whereas it was for the indicator. As would be expected, currently holding a lottery stock makes a purchase of another more likely, with the odds increasing by 44%. This indicator will clearly be capturing similar information as the lottery buy proportion variable, but the fact that both are significant shows that an investor who currently holds a lottery stock is more likely to buy another relative to an investor that does not, even if both have displayed a similar preference for lottery stocks in the past.

Moving on to more general information about the investor's behaviour and trading style, the results show that investors with more concentrated portfolios, where a larger proportion of their portfolio value is assigned to a single stock, are more likely to purchase lottery stocks. The coefficient estimate is only significant at the 5% level however, and the effect is small. Holding 3 equally weighted stocks rather than just 1 (which lowers concentration from 1 to 0.33) decreases the odds of a lottery stock purchase by 9%, and similarly holding 10 rather than 3 decreases the odds by 3%. So portfolio concentration is informative about an investor's tendency to buy lottery stocks, but not strongly so when the other factors in the model are controlled for.

Related to portfolio concentration is the number of positions the investor currently holds in total. Since this variable is highly skewed, its natural logarithm was entered into the model, with 1 added to the total prior to the transformation in order to avoid taking the logarithm of zero. This variable was highly significant and had a fairly large effect for differences that would be common to observe in the data. An increase from 1 to 3 positions reduced the odds of a lottery stock purchase by 13%, an increase from 3 to 12 reduced the odds by 21% and an increase from 12 to 20 positions reduced the odds by 9%. Holding more positions means the investor is better diversified, and this is generally thought to indicate a more sophisticated approach to investing. More sophisticated investors being less likely to purchase a lottery stock is in line with both the empirical findings in Kumar (2009a), and the accompanying theory about which groups of investors will find lottery stocks most appealing. It is also worth noting that the fluctuations inherent to high volatility stocks will be more damaging for investors who do not have the 'cushion' of a well diversified portfolio to absorb potential losses.

Another important dimension to an investor's trading behaviour is how frequently they trade. This needed to be controlled for so that the effect of other covariates that could proxy for it was not obscured. It was included in the model by dividing the number of trading days from the start of the data period to the time of the purchase by the number of actions, buy or sell, that the investor took during that time. This quantity was then logged and centred. A higher value indicates that the investor takes actions less frequently than an investor with a lower value. The results show that the odds of buying a lottery stock are 13% lower for an investor who has traded once per month (21 days per action on average) compared to one who has traded once per week (5 days per action). Similarly, the odds are 22% lower for an investor who has traded once per year (250 days per action on average) compared to one who has traded once per month. Lottery stocks naturally lend themselves to a more active trading style due to their volatile nature. The median holding period for lottery stocks positions is 114 days, compared to 189 for non-lottery stocks. However, trading frequently can also result from overconfidence, as explored by Barber and Odean (2000), who find that frequent trading typically leads to poorer performance. Less frequent trading could therefore indicate a more sophisticated investor.

Static investor-level covariates

Amongst the static covariates, only age and income had significant parameter estimates. Despite explaining variation in the preference for lottery stocks when aggregated over the whole data period, as was done by Kumar, the remainder of these variables had no effect in this model when other factors were controlled for. This includes indicators for the investor's gender, marital status, whether they are retired, whether they are in a professional occupation and their account tenure in years[11]. The investor's diversification at the start of the data period was also not significant, although as would be expected this variable is correlated with the number of positions the investor holds at the time of the purchase (Pearson correlation = 0.52), so it is likely that it would be significant if that variable was not included.

For age, the coefficient estimate for age itself is positive, implying lottery stock purchases become more likely as age increases, but the coefficient is negative for the square of age. Since squaring is a non-linear transformation, the effect on the odds depends on both the difference in age and the magnitude of the two ages being compared. Table 3.8 shows the change in odds for some example differences in age. The median age of investors in this sample is 48, so for investors in the top half of the age distribution, older investors are less likely to purchase lottery stocks relative to younger ones, provided the age difference is at least a few years. However the opposite will usually be true for investors below the median age. This adds important nuance to understanding the relationship between age and preference for lottery stocks.

[11] Kumar calls this variable 'investment experience'.

Difference in age	% change in odds
30 → 40	6.6%
50 → 60	-7.0%
25 → 50	12.3%
50 → 75	-27.2%

Table 3.8: Change in odds for some example differences in age

As expected, investors with higher income are less likely to purchase a lottery stock. A doubling of income, which is roughly the difference between the 25th and 75th percentiles, results in a 7.4% reduction in odds. Income at the 90th percentile is 4.3x larger than at the 10th percentile, which equates to a 15% reduction in odds.

Exogenous covariates

In agreement with Kumar's aggregate results, better recent performance of lottery stocks in the market increased the odds of lottery stock purchases. An increase in the excess return of lottery stocks by one percentage point in the previous calendar month raised the odds of a lottery stock purchase by 1.7%. An increase of 5 percentage points, which is roughly equal to the IQR across all months in the dataset, raised the odds by 8.6%. An increase of 10 percentage points, which is roughly the difference between the 10th and 90th percentiles, raised the odds by 17.9%. Two immediate questions are whether some groups of investors are more or less sensitive to this effect, and if the effect is particular to lottery stocks or would be present for another group as well. These questions are addressed in sections 3.6.1 and 3.6.2 respectively.

The calendar year dummies show that the odds of a lottery stock purchase are much greater for later years, even after the other factors in the model have been controlled for. The odds are 23%, 40%, 76% and 93% large in years 1993-96 respectively, compared to those in 1991, with no significant difference between 1991 and 1992. The magnitude of this change shows the importance of controlling for it when measuring the effect of other covariates. As discussed in section 1.5.2, the period of time covered by this dataset was one of sustained growth in the U.S. stock market, which was accompanied by increased interest and confidence amongst the general public. However, in the same dataset Kumar finds evidence that aggregate demand for lottery stocks is greater during economic downturns, for example when the unemployment rate is higher. Further investigation would be required to discover what was driving the aggregate trend of increasing demand for lottery stocks, and whether the trend continued beyond the end of the data period.

3.6 Supplementary models

3.6.1 Is the recent performance of lottery stocks less important for some groups of investors?

The importance of the recent performance of lottery stocks in the market, specifically their return in the previous calendar month, raises the question of whether some groups of investors may be more or less sensitive to this effect. This was tested by re-running the model with the addition of interactions between the lottery returns variable and a set of the other covariates that may capture differences in sophistication or investment style. These interacted variables include age, gender, income, portfolio concentration, the number of currently held positions, the number of days per action so far, the investor's initial diversification and the indicators for if the investor is retired, has a professional occupation and their marital status. Regression results for these interaction terms are presented in table 3.9.

Of these interactions, only the one with the number of currently held posi-

	Coefficient	SE	p-value
RLPM * Age	-0.104	0.097	0.285
RLPM * Squared(Age)	0.106	0.098	0.283
RLPM * Log(Income)	-0.024	0.012	0.051
RLPM * Portfolio concentration	-0.057	0.060	0.345
RLPM * Log(Number of current positions)	-0.081	0.020	< 0.001
RLPM * Log(Number of days per action so far)	-0.033	0.018	0.071
RLPM * Gender: Male	-0.026	0.051	0.613
RLPM * Professional occupation	0.056	0.031	0.071
RLPM * Log(Initial diversification)	-0.010	0.016	0.552
RLPM * Retired	0.038	0.046	0.411
RLPM * Marital status: Single	-0.064	0.031	0.04

Table 3.9: Results for interactions with the variable recording the return of lottery stocks in the previous calendar month (RLPM) when added to the logistic regression model with investor-level random effects.

tions had a significant coefficient estimate at the 1% level. The coefficient for this interaction was negative, meaning that increases in recent lottery performance are less effective at raising the odds of a lottery stock purchase for investors who hold a larger number of positions. Table 3.10 shows the change in odds when the market return of lottery stocks for the previous calendar month increases by 5 and 10 percentage points, for investors who hold 1, 5, and 10 positions but are otherwise identical. Whilst an increase in lottery return raises the odds of a lottery stock purchase for an investor who holds only 1 position, for an investor who holds 10 positions such an increase actually decreases the chance of them purchasing a lottery stock.

3.6.2 Equivalent model for non-lottery stocks

A model like the one in section 3.5.1 could in theory be estimated for any category of stock in order to see which factors are important in the decision to purchase them. To demonstrate this, and as a comparison to the results of the model for lottery stocks, an equivalent model was estimated for non-

Number of positions	Increase in lottery returns (ppts)	Change in odds
1	5	11.9%
	10	25.3%
5	5	3.0%
	10	6.2%
10	5	-1.6%
	10	-3.1%

Table 3.10: Examples of the effect on the odds of a lottery stock purchase for increases in the return of lottery stocks in the market in the previous calendar month (denoted in percentage points) for investors who currently hold 1, 5, and 10 positions.

lottery stocks. This is a category of stocks also defined by Kumar that are essentially the opposite of lottery stocks. In a particular month, non-lottery stocks are those that are in the bottom half of the distribution for volatility and skewness, and in the top half for price. They tend to be large-cap stocks and hence account for a greater proportion of purchases in the dataset than lottery stocks: 29% compared to 9%. The estimated model contained all the same covariates as in section 3.5.1, but with the category of stock changed where appropriate. For example, the paper return of the investor's currently held non-lottery stock positions was included, as well as the market return of non-lottery stocks in the previous month. The results are shown in table 3.11

Since non-lottery stocks can be thought of as the opposite of lottery stocks, covariates capturing information about investment style and sophistication might be expected to have the reverse of the effect they had in the lottery model. This is true in some cases: investor's who trade more actively (smaller number of days per action) and hold more concentrated portfolios have lower odds of purchasing a non-lottery stock. Investors with greater initial diversification, larger incomes, a non-lottery stock currently in their portfolio and who had purchased more non-lottery stocks in the past had increased odds. As expected from figure 3.1, the calendar year indicators

	Coefficient	SE	p-value
Intercept	-0.785	0.074	< 0.001
Non-lottery sale in past year	-0.026	0.023	0.254
Return of non-lottery sales in past year	0.008	0.008	0.325
Currently holds non-lottery stock	0.175	0.029	< 0.001
Paper return of current non-lottery stock positions	-0.133	0.033	< 0.001
Portfolio concentration	-0.148	0.047	0.002
Log(Number of current positions)	-0.104	0.022	< 0.001
Log(Number of days per action so far)	0.223	0.022	< 0.001
Sqrt(Non-lottery buy proportion)	0.447	0.051	< 0.001
Return of non-lottery stocks in previous month	-0.017	0.009	0.045
Gender: Male	-0.098	0.060	0.1
Age	-0.077	0.112	0.491
Squared(Age)	0.212	0.113	0.061
Marital status: Single	-0.027	0.038	0.472
Log(Initial diversification)	0.097	0.019	< 0.001
Tenure	0.006	0.016	0.715
Log(Income)	0.044	0.016	0.006
Professional occupation	-0.022	0.037	0.553
Retired	0.092	0.054	0.092
Year: '92	-0.204	0.028	< 0.001
Year: '93	-0.413	0.032	< 0.001
Year: '94	-0.714	0.036	< 0.001
Year: '95	-0.685	0.039	< 0.001
Year: '96	-0.868	0.041	< 0.001

Table 3.11: Results from estimating a logistic regression model with investor-level random effects for non-lottery stock purchases.

confirm that purchases of non-lottery stocks are much less likely in later years. All else being equal, the odds of a purchase being of a non-lottery stock purchase are 58% lower in 1996 relative to 1991.

Some of the results are more surprising. Despite its importance in the lottery stock model, neither age nor its square were significant predictors of non-lottery stock purchases. Likewise, the indicator for whether the investor had sold a non-lottery stock in the past year, and the return of any such sales were both insignificant at the 5% level. Whereas greater initial diversification increased the odds of an investor purchasing a non-lottery stock, their level of diversification at the time of the purchase had the opposite effect. Only the latter variable was significant in the lottery model (odds decreasing as current level of diversification increases), whereas they both are in this model. The two effects are similar in magnitude however, so for example the odds of purchasing a non-lottery stock for an investor who held 2 stocks at the start of the data period and holds 2 now are not much different from an investor who held 10 stocks at the beginning and holds 10 now. One interpretation of these results is that, for two investors who had the same level of initial diversification, the investor with a lower level of current diversification is more likely to purchase a non-lottery stock. If an investor feels that their current level of diversification is not sufficient, then purchasing a non-lottery stock probably makes more sense than purchasing any other type. Hence the odds of a non-lottery stock purchase increase as the number of stocks currently held falls. This is one plausible explanation, but it is hard to rule out others on the basis of these results alone.

Whilst currently holding a non-lottery stock did make an investor more likely to purchase another, conditional on this, investors whose current non-lottery holdings were performing better were actually less likely to purchase another non-lottery stock. In agreement with this latter result, strong performance of non-lottery stocks in the market during the previous calendar month also reduces the odds of non-lottery stock purchases. This could be due to mental accounting[12], and would be consistent with the behavioural portfolio theory of Shefrin and Statman (2000) in which investors construct their portfolios in layers which have different purposes. If an investor has a non-lottery component to their portfolio that is intended to provide security, then strong performance of this component may make the investor feel more able to expand the other components of their portfolio and therefore be less likely to purchase additional non-lottery stocks.

This suggests that strong performance of non-lottery stocks might increase the odds of lottery stock purchases. However, in untabulated results, the indicator for currently holding a non-lottery stock and the paper return of an investor's current non-lottery holdings are not significant when added to the lottery stock model of section 3.5.1. This means that when non-lottery stocks are performing well in the market and the investor's own portfolio, they are choosing to purchase stocks that are somewhere in between the two extremes of the lottery and non-lottery categories.

3.6.3 The importance of repurchases

The results in section 3.5 provide strong evidence (p-value < 0.001) that investors who have purchased more lottery stocks in the past are more likely to purchase them in the future, and weaker evidence (p-value $= 0.026$) that an increased return on the investor's lottery stock sales in the past year also makes them more likely to purchase another lottery stock. These results could be explained by the tendency for investors to repurchase the same stocks they have held in the past, and particularly stocks they have had a positive prior experience with. Strahilevitz et al. (2011) find that, amongst stocks they have previously held, investors are more likely to repurchase stocks that they have previously sold for a gain and which have decreased in price subsequent to the sale. Jiao (2015) finds that more experienced

[12] First introduced by Thaler (1985)

and sophisticated investors are less biased against stocks they have previously sold for a loss, but are still more likely to repurchase stocks they have previously sold for a gain.

	Coefficient	SE	p-value
Intercept	-3.241	0.110	< 0.001
Lottery sale in past year	-0.006	0.047	0.897
Return of lottery sales in past year	0.026	0.016	0.099
Currently holds lottery stock	0.239	0.044	< 0.001
Paper return of current lottery stock positions	-0.001	0.016	0.942
Portfolio concentration	0.187	0.082	0.023
Log(Number of current positions)	-0.218	0.037	< 0.001
Log(Number of days per action so far)	-0.217	0.035	< 0.001
Sqrt(Lottery buy proportion)	0.645	0.119	< 0.001
Return of lottery stocks in previous month	0.099	0.016	< 0.001
Gender: Male	0.125	0.089	0.159
Age	0.438	0.168	0.009
Squared(Age)	-0.534	0.172	0.002
Marital status: Single	0.003	0.055	0.964
Log(Initial diversification)	-0.033	0.028	0.239
Tenure	-0.011	0.023	0.623
Log(Income)	-0.078	0.023	< 0.001
Professional occupation	0.058	0.053	0.277
Retired	0.073	0.081	0.365
Year: '92	0.062	0.055	0.265
Year: '93	0.315	0.060	< 0.001
Year: '94	0.382	0.070	< 0.001
Year: '95	0.637	0.070	< 0.001
Year: '96	0.742	0.074	< 0.001

Table 3.12: Results from estimating a logistic regression model with investor-level random effects for lottery stock purchases. Repurchases were removed, so the sample only contained the first purchase an investor made of any particular stock.

In the LDB dataset, 37% of purchases are of a stock that the investor has previously held. The percentage amongst the sample used to estimate the model in section 3.5 is very similar. This is an underestimate of the true value since many of the investors will already have been trading for a long time when the data period began. To test the effect repurchases have on the results of the model in section 3.5, the model was re-estimated with repurchases removed i.e. with only the first purchase an investor makes of any particular stock during the data period. The results are presented in 3.12. The proportion of the investor's purchases that have been of lottery stocks so far is still highly significant and has the same sign and similar magnitude as in the original model. However the return of the investor's lottery sales in the past year is no longer significant at the 5% level in this model (p-value = 0.099), although the sign and magnitude of the estimated coefficient are the same. In untabulated results, this variable was also not significant in a model estimated using only repurchases. This suggests that the loss of significance is mainly a result of the reduced sample size rather than repurchases being primarily responsible for the effect. The sample without repurchases still contains 60,000 purchases though, so if it exists the effect is small.

3.6.4 Without demographic covariates and a larger sample

As mentioned in section 3.4, including demographic covariates in the model dramatically reduces the number of purchases that can be used since data for these variables is missing in many cases. As a robustness check, the model was re-estimated without demographic covariates and using the maximum possible number of purchases. This larger sample contained 677,641 purchases compared to 92,975 for the sample used in section 3.5.1.

The results of the large sample model are displayed in table 3.13. As can be seen by comparing these results with those in table 3.6, there is agreement about the sign and magnitude of all effects that are significant in both models. Likewise, the indicator for whether the investor has sold a lottery stock

	Coefficient	SE	p-value
Intercept	-3.17130	0.018	< 0.001
Lottery sale in past year	0.00001	0.013	0.999
Return of lottery sales in past year	0.01235	0.004	0.002
Currently holds lottery stock	0.40012	0.012	< 0.001
Paper return of current lottery stock positions	0.00023	0.004	0.955
Portfolio concentration	0.18853	0.024	< 0.001
Log(Number of current positions)	-0.19965	0.011	< 0.001
Log(Number of days per action so far)	-0.11678	0.010	< 0.001
Sqrt(Lottery buy proportion)	1.04432	0.034	< 0.001
Return of lottery stocks in previous month	0.07578	0.005	< 0.001
Year: '92	0.03686	0.015	0.017
Year: '93	0.20863	0.017	< 0.001
Year: '94	0.31817	0.020	< 0.001
Year: '95	0.50877	0.020	< 0.001
Year: '96	0.59520	0.021	< 0.001

Table 3.13: Results from estimating a logistic regression model with investor-level random effects for lottery stock purchases. No demographic variables were included, and as a result a much larger sample was used.

in the past year and the paper return of the investor's current lottery stock positions are the only two main effects that are not significant amongst those that are in both models. These results provide reassurance that the variables with weaker significance relative to the others in the smaller sample model, namely the return of lottery stock sales the investor made in the past year and their portfolio concentration, do have a real effect. This supports the conclusion in section 3.6.3 that the lack of significance of the return of lottery sales in the past year in a model which excludes repurchases is likely due to the smaller sample size, rather than a major difference amongst repurchases with regards to this variable.

3.7 Summary

Variables capturing information about the investor's behaviour since the start of the data period and the current state of their portfolio had a strong effect on their odds of purchasing a lottery stock. More well diversified investors, in terms of both number of stocks held and portfolio concentration, were less likely to purchases a lottery stock. Likewise for investors who traded less frequently. As expected, a lottery stock purchase was more likely if the investor had displayed a stronger preference for them in the past.

The results provided some evidence that an investor's recent experiences with lottery stocks had an effect on their odds of purchasing one. For investors who had sold a lottery stock in the past year, those earning a greater return were more likely to purchase another lottery stock. However this effect was small and only significant at the 5% level, which is weak given the sample size and the significance of other more important effects. The effect persisted and had a greater level of significance in a model estimated without demographic covariates and a much larger sample, which provides some reassurance that it is genuine. Whilst currently holding a lottery stock greatly increased the odds of an investor purchasing another, the current paper return of their lottery stock positions did not have a significant effect. Related to these variables is the recent return of lottery stocks in the market. Stronger performance in the previous calendar month was found to raise the odds of an investor making a lottery stock purchase. A model containing interactions revealed that better diversified investors were less sensitive to the effect of this variable.

Of the static investor-level covariates, only age and income had a significant effect when the time-varying covariates were included in the model. Evidence was found for a non-linear effect of age, with the coefficient of age itself being positive i.e. odds increasing with age, and the square of age having a negative coefficient. As expected, investors in the top half of the age distribution will tend to have have lower odds of making a lottery stock purchase compared to investors who are in the bottom half. But comparisons in the middle are more complicated. The effect of income was as expected, with the odds of lottery stock purchases decreasing with income. But again this effect was fairly small: an investor at the 75th percentile of the income distribution had only a 7.4% reduction in odds to compare to an investor at the 25th percentile.

The supplementary models in section 3.6 showed that the results are not substantially altered by either using a much larger sample as a result of omitting demographic variables, or by controlling for the known preference of investors for stocks they have held in the past and had positive experiences with. To demonstrate that the mixed-effects model for stock purchases can be applied to any category of stocks, a model was estimated for purchases of non-lottery stocks, the qualitative 'opposite' of lottery stocks. This produced some surprising results that warrant further investigation. For example, stronger recent performance of non-lottery stocks in the investor's portfolio or in the market reduced the odds of a non-lottery stock being purchased.

Appendix

	Correlation	χ^2	p-value	Correlation	χ^2	p-value
Gain * Professional occupation	-0.002	0.230	0.631	-0.027	392.837	< 0.001
Gain * Male	-0.002	0.137	0.711	-0.008	48.200	< 0.001
Gain * Experience: N	-0.004	0.990	0.32	-0.025	539.597	< 0.001
Gain * Experience: L	-0.002	0.255	0.613	-0.001	0.913	0.339
Gain * Experience: G	-0.001	0.069	0.793	-0.002	2.148	0.143
Gain * Income 2	-0.003	0.672	0.413	-0.004	7.140	0.008
Gain * Income 3	0.004	0.735	0.391	0.021	219.909	< 0.001
Gain * Income 4	-0.001	0.079	0.778	0.017	186.775	< 0.001
Gain * December	-0.048	119.957	< 0.001	-0.031	133.792	< 0.001
Gain * 95/96	0.014	10.207	0.001	0.035	321.657	< 0.001
Gain * Cap quintile 1	0.019	19.016	< 0.001	0.037	184.752	< 0.001
Gain * Cap quintile 2	0.017	15.200	< 0.001	0.035	273.665	< 0.001
Gain * Cap quintile 3	0.019	19.737	< 0.001	0.035	388.410	< 0.001
Gain * Cap quintile 4	0.024	30.948	< 0.001	0.036	247.704	< 0.001
Gain * Age	0.002	0.129	0.719	0.022	193.935	< 0.001
Gain * Diversification 2	0.003	0.374	0.541	0.002	2.402	0.121
Gain * Diversification 3	-0.004	0.983	0.322	0.007	22.437	< 0.001
Gain * Diversification 4	-0.005	1.154	0.283	0.004	10.244	0.001
Gain * Log sales made	0.013	8.829	0.003	-0.038	1145.005	< 0.001
Gain * Tenure 2	0.020	21.963	< 0.001	0.018	183.899	< 0.001
Gain * Tenure 3	0.000	0.001	0.972	-0.009	42.030	< 0.001
Gain * Tenure 4	0.004	0.724	0.395	-0.014	121.455	< 0.001

Table 3.14: Correlation coefficient between the scaled Schoenfeld residuals and survival times for each interaction term in the marginal and frailty models. Reported with the χ^2 test statistics for whether this correlation is non-zero and associated p-values.

	HR	SE	p-value
Neutral	1.141	0.018	< 0.001
Gain	4.214	0.067	< 0.001
Professional occupation	1.060	0.034	0.083
Male	1.044	0.062	0.492
Experience: N	0.569	0.106	< 0.001
Experience: L	0.617	0.051	< 0.001
Experience: G	0.751	0.047	< 0.001
Income 2	1.036	0.043	0.414
Income 3	1.120	0.045	0.012
Income 4	1.027	0.056	0.635
December	1.620	0.022	< 0.001
95/96	1.175	0.017	< 0.001
Cap quintile 1	0.776	0.046	< 0.001
Cap quintile 2	0.857	0.032	< 0.001
Cap quintile 3	1.006	0.024	0.815
Cap quintile 4	1.107	0.019	< 0.001
Age	1.004	0.001	0.003
Diversification 2	0.983	0.039	0.662
Diversification 3	0.855	0.052	0.003
Diversification 4	0.806	0.057	< 0.001
Log sales made	1.167	0.007	< 0.001
Tenure 2	0.990	0.045	0.818
Tenure 3	0.785	0.043	< 0.001
Tenure 4	1.000	0.060	0.995
SIC: AFF	1.086	0.228	0.718
SIC: CON	0.996	0.058	0.944
SIC: FIRE	0.775	0.017	< 0.001
SIC: MIN	0.807	0.025	< 0.001
SIC: PUBA	1.188	0.130	0.183
SIC: RETL	0.923	0.017	< 0.001
SIC: SERV	1.034	0.014	0.017
SIC: TRAN	0.686	0.018	< 0.001
SIC: WHOL	1.096	0.028	0.001
Gain * Professional occupation	0.989	0.020	0.568

Table 3.15: Hazard ratios, standard errors and p-values for the main effects in the frailty model with all interaction terms included.

	HR	SE	p-value
Neutral	1.087	0.026	0.001
Gain	3.603	0.194	< 0.001
Professional occupation	1.093	0.041	0.028
Male	1.024	0.092	0.795
Experience: N	0.684	0.135	0.005
Experience: L	0.809	0.063	< 0.001
Experience: G	0.892	0.052	0.03
Income 2	1.009	0.054	0.863
Income 3	0.998	0.055	0.974
Income 4	0.915	0.066	0.183
December	1.594	0.034	< 0.001
95/96	0.854	0.031	< 0.001
Cap quintile 1	0.868	0.056	0.011
Cap quintile 2	0.925	0.040	0.049
Cap quintile 3	1.037	0.033	0.264
Cap quintile 4	1.142	0.024	< 0.001
Age	1.004	0.002	0.005
Diversification 2	0.866	0.047	0.002
Diversification 3	0.671	0.062	< 0.001
Diversification 4	0.556	0.063	< 0.001
Log sales made	1.640	0.016	< 0.001
Tenure 2	1.105	0.052	0.057
Tenure 3	0.907	0.051	0.056
Tenure 4	1.189	0.068	0.011
SIC: AFF	1.405	0.228	0.136
SIC: CON	0.992	0.058	0.893
SIC: FIRE	0.775	0.022	< 0.001
SIC: MIN	0.823	0.033	< 0.001
SIC: PUBA	1.099	0.131	0.472
SIC: RETL	0.949	0.020	0.01
SIC: SERV	1.042	0.019	0.027
SIC: TRAN	0.696	0.023	< 0.001
SIC: WHOL	1.127	0.030	< 0.001

Table 3.16: Hazard ratios, robust standard errors and Wald test p-values for the main effects in the marginal model. Standard errors are robust to correlation at the investor level.

	HR	SE	p-value
Gain * Professional occupation	0.974	0.053	0.623
Gain * Male	1.104	0.122	0.417
Gain * Experience: N	1.568	0.223	0.044
Gain * Experience: L	1.292	0.080	0.001
Gain * Experience: G	1.161	0.069	0.032
Gain * Income 2	0.896	0.071	0.123
Gain * Income 3	0.931	0.071	0.313
Gain * Income 4	0.955	0.091	0.614
Gain * December	0.503	0.042	< 0.001
Gain * 95/96	0.742	0.037	< 0.001
Gain * Cap quintile 1	1.265	0.077	0.002
Gain * Cap quintile 2	1.197	0.056	0.001
Gain * Cap quintile 3	1.090	0.045	0.055
Gain * Cap quintile 4	1.018	0.032	0.574
Gain * Age	0.988	0.002	< 0.001
Gain * Diversification 2	0.983	0.061	0.775
Gain * Diversification 3	0.976	0.074	0.744
Gain * Diversification 4	0.936	0.084	0.432
Gain * Log sales made	0.979	0.022	0.331
Gain * Tenure 2	1.035	0.070	0.623
Gain * Tenure 3	1.192	0.070	0.012
Gain * Tenure 4	1.081	0.090	0.386

Table 3.17: Hazard ratios, robust standard errors and Wald test p-values for the interactions in the marginal model. Standard errors are robust to correlation at the investor level.

Bibliography

Brad M Barber and Terrance Odean. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*, 55(2):773–806, 2000.

Brad M Barber and Terrance Odean. Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292, 2001a.

Brad M Barber and Terrance Odean. The internet and the investor. *The Journal of Economic Perspectives*, 15(1):41–54, 2001b.

Brad M Barber and Terrance Odean. Online investors: do the slow die first? *Review of Financial Studies*, 15(2):455–488, 2002.

Brad M Barber and Terrance Odean. The behavior of individual investors. *Available at SSRN 1872211*, 2011.

Brad M Barber, Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean. Just how much do individual investors lose by trading? *The Review of Financial Studies*, 22(2):609–632, 2008.

Nicholas Barberis and Ming Huang. Stocks as lotteries: The implications of probability weighting for security prices. Technical report, National Bureau of Economic Research, 2007.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67 (1):1–48, 2015. doi: 10.18637/jss.v067.i01.

L Bernstein, J Anderson, and MC Pike. Estimation of the proportional hazard in two-treatment-group clinical trials. *Biometrics*, pages 513–519, 1981.

Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135, 2009.

George M Constantinides. Optimal stock trading with personal taxes: Implications for prices and the abnormal january returns. *Journal of Financial Economics*, 13(1):65–89, 1984.

David R Cox et al. Regression models and life tables. *JR stat soc B*, 34(2): 187–220, 1972.

Ravi Dhar and Ning Zhu. Up close and personal: Investor sophistication and the disposition effect. *Management Science*, 52(5):726–740, 2006.

Peter K Dunn and Gordon K Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

Cindy Feng, Alireza Sadeghpour, and Longhai Li. Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution. *arXiv preprint arXiv:1708.08527*, 2017.

Lei Feng and Mark S Seasholes. Do investor sophistication and trading experience eliminate behavioral biases in financial markets? *Review of Finance*, 9(3):305–351, 2005.

William N Goetzmann and Alok Kumar. Equity portfolio diversification*. *Review of Finance*, 12(3):433–463, 2008.

Patricia M Grambsch and Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

Mark Grinblatt and Matti Keloharju. What makes investors trade? *The Journal of Finance*, 56(2):589–616, 2001.

Campbell R Harvey and Akhtar Siddique. Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3):1263–1295, 2000.

Vicky Henderson, David Hobson, and Alex SL Tse. Probability weighting, stop-loss and the disposition effect. 2017.

David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time to event data*, volume 618. John Wiley & Sons, 2011.

Zoran Ivkovic, James Poterba, and Scott Weisbenner. Tax-motivated trading by individual investors. *American Economic Review*, 95(5):1605–1630, 2005.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993.

Peiran Jiao. Losing from naive reinforcement learning: A survival analysis of individual repurchase decisions (november 18, 2015). 2015. Available at SSRN: https://ssrn.com/abstract=2574141.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, pages 263–291, 1979.

John P Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806, 1992.

George M Korniotis and Alok Kumar. Do portfolio distortions reflect superior information or psychological biases? *Journal of Financial and Quantitative Analysis*, 48(1):1–45, 2013.

Alok Kumar. Who gambles in the stock market? *The Journal of Finance*, 64(4):1889–1933, 2009a.

Alok Kumar. Hard-to-value stocks, behavioral biases, and informed trading. *Journal of Financial and Quantitative Analysis*, 44(06):1375–1401, 2009b.

Danyu Y Lin and Lee-Jen Wei. The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408): 1074–1078, 1989.

Stuart R Lipsitz, Michael Parzen, et al. A jackknife estimator of variance for cox regression for correlated survival data. *Biometrics*, 52(1):291, 1996.

Todd Mitton and Keith Vorkink. Equilibrium underdiversification and the preference for skewness. *Review of Financial Studies*, 20(4):1255–1288, 2007.

Gert G Nielsen, Richard D Gill, Per Kragh Andersen, and Thorkild IA Sørensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, pages 25–43, 1992.

Terrance Odean. Are investors reluctant to realize their losses? *The Journal of finance*, 53(5):1775–1798, 1998.

Terrance Odean. Do investors trade too much? *The American economic review*, 89(5):1279–1298, 1999.

Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207, 1972.

Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 2009.

Elena Ranguelova. Disposition effect and firm size: New evidence on individual investor trading activity. 2001.

Gary G Schlarbaum, Wilbur G Lewellen, and Ronald C Lease. The common-stock-portfolio performance record of individual investors: 1964–70. *The Journal of Finance*, 33(2):429–441, 1978.

David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

Amit Seru, Tyler Shumway, and Noah Stoffman. Learning by trading. *Review of Financial Studies*, 23(2):705–739, 2010.

Zur Shapira and Itzhak Venezia. Patterns of behavior of professionally managed and independent investors. *Journal of Banking & Finance*, 25(8): 1573–1587, 2001.

Hersh Shefrin and Meir Statman. The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of finance*, 40 (3):777–790, 1985.

Hersh Shefrin and Meir Statman. Behavioral portfolio theory. *Journal of financial and quantitative analysis*, 35(02):127–151, 2000.

Michal Ann Strahilevitz, Terrance Odean, and Brad M Barber. Once burned, twice shy: How naïve learning, counterfactuals, and regret affect the repurchase of stocks previously sold. *Journal of Marketing Research*, 48 (SPL):S102–S120, 2011.

Richard Thaler. Mental accounting and consumer choice. *Marketing science*, 4(3):199–214, 1985.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL https://CRAN.R-project.org/package=survival. version 2.38.

Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2000.

Francis Tuerlinckx, Frank Rijmen, Geert Verbeke, and Paul Boeck. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255, 2006.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Andreas Wienke. *Frailty models in survival analysis*. CRC Press, 2010.

Ronghui Xu and John O'Quigley. An r2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12(1): 83–107, 1999.

Daowen Zhang and Xihong Lin. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random effect and latent variable model selection*, pages 19–36. Springer, 2008.