LUISS Guido Carli
LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

**PhD THESIS**

PhD Candidate: Francesco Corea

# Thesis Title:
# Essays on Machine Learning for Economics and Finance

Keywords:
*Sentiment Analysis; Big data; Insurance analytics; Regularization; Stepwise regression; Entrepreneurship.*

PhD in Economics
*XXIX* Cycle
LUISS Guido Carli
Supervisor: Prof. Giuseppe Ragusa
*Month 2017*

Thesis Defense:
*Month Day, 2017*

Thesis Committee:
Prof. Giuseppe Ragusa, LUISS Guido Carli University
Prof. *Name Surname*, *Institution*
Prof. *Name Surname*, *Institution*

**Abstract**

Econometrics and machine learning are quite close and related concepts. Nowadays, it is always more important to extract value from raw data, and distilling actionable insights from quantitative values as well as qualitative features. In order to deal with these topics, the first chapters (Chapter 1 - 4) are going to introduce the new wave called *machine learning* or *big data* and they will explain the most common techniques used in the field, respectively regression, clustering, model selection, and tree-based models (Chapter 2); time series analysis (Chapter 3); and eventually forecasting model with shrinkage methods (Chapter 4). Then, three applications are going to be provided.

In Chapter 5, it is going to be shown an example of big dataset for the insurance vertical. Rothschild and Stiglitz ([30]) argued that people signal their risk profile through their insurance demand, i.e. individuals with a high risk profile would buy insurance as much as they can, while people who are not going to buy any insurance are the ones with a lower risk profile. This issue is commonly known as adverse selection. Even if their prediction seems to work quite well in a lot of different markets, Cutler et al. ([13]) proved that there exist some insurance markets in United States in which the expected result is completely different. In the wake of this study, we provide empirical evidences that there are some European insurance markets in which the low risk profile agents are the ones who buy more insurance.

In Chapter 6, a second application is going to be provided. It has been studies the effect of behavioural biases on entrepreneurial choices to insure their firms against kinds of corporate risks. It has been used a large sample of Italian Small and Medium sized - finding that they under-insure themselves. The dataset allows to link corporate insurance choices with the personal traits of the entrepreneur and his household's financial choices.

In Chapter 7, finally, an application to financial markets is going to be shown. Bollen et al. ([10]) reintroduced the idea of formulating prediction based on the general sentiment of the investors, even if they originally exploited microblogging data. The purpose of this study is to verify whether social data may have a predictive power for the stock prices, returns, and volumes. The analysis has been implemented for different large technology companies, and the robustness has been tested through a ten-days rolling window. The evidence shows that there is some intrinsic value in these new features, and that both the sentiment and the amount of tweets posted online can improve the forecast given by a baseline autoregressive model. Some additional variations have been tested eventually with the same dataset.

# DISCLAIMER - LIBERATORIA

# Acknowledgements

I would like to thank first of all Prof. Ragusa, who gave me the tools to complete this thesis, and the chance to pursue my interests setting my own pace.

I want to also thank my family, who supported me throughout the entire period of my PhD, and Lucia, the only one who knows how much effort and days I have spent for completing this path.

Despite everything, all errors are and remain my own.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

We firmly believe that you cannot understand anything in science well enough until you are make the effort to explain that to someone else who does not know the topic at all. Furthermore, paraphrasing Professor Knuth's words, we really understand something deeply when we are able to teach it to a computer. If we do not, we cannot talk about science but we deal with art.

This thesis is about the both sides, science and art, and the motivation behind it is pretty intuitive: *"The sexiest job in the next 10 years will be statisticians/data scientist. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"*. These were the Hal Varian's words, chief Economist at Google, and they stressed how important are becoming the professional figures able to join statistical/econometric knowledge, practical programming skills and a business/economic overview and way of thinking and tackling problems. Hence, the inner intent for this work is to represent a robust bridge between economics and practical machine learning.

The work is going to be structured as follows: first of all, an introduction on machine learning concepts is going to be provided, stressing a wide range of methods, techniques and definitions useful to understand the topic. Then, we will focus on time series analysis and an empirical section will be shown. A first application related to the issue of time series analysis in big data framework will be dealt with, and finally an example of some estimation techniques applied to the insurance market with a medium-large dataset will be handled as well.

In general, the purpose of a consultant that takes advantage statistical methods is to provide a model able to forecast a certain variable (e.g., sales, stock price, etc.) that will entail more money in the client's pocket. In order to do that, he has to collect data. Until some years ago, he actually was used to collect as much data as he could to usually use only a sample of it since he lacked the methodological toolbox to exploit all of them and the storage to analyse and keep them. Basically, what he does is to identify a function $f$ such that, once feed it with inputs (called also *independent variables* or *predictors*), it is able to give back a prediction for the outputs (*dependent* or *response variable*). Unfortunately, this forecasting is only as much accurate as possible, but it is never able to perfectly predict the exact value each single time. This is reflected in what is called *error term ($\mu$)*, that embeds every hidden factor able to explain the discrepancy between the predicted and the correct output value. Part of this error can be reduced through a better model, but the rest of it is systematic. Theoretically speaking, only if you would be able to create a model with all the possible suitable predictors the error will shrink to zero. It is good to keep that in mind for practical applications, since we would not be able to reach a precise prediction of any output value.

$$Y_i = f(X_i) + \mu_i \tag{1.1}$$

where the subscript $i$ runs over observations, *i= 1,...n*, $Y_i$ is the independent variable, $X_i$ the predictors and $f$ the functional form that points out what kind of relationship exists between inputs and outputs. This may involve one or more inputs, can be linear or non-linear, parametric or non-parametric, whether it actually gets a specific form or not. In the parametric case, the model will be more rigid but also "lighter" to be run. The non-parametric case is instead regarding a more flexible model, and will involve the estimation of several parameters with a huge amount of data. Furthermore, if the complexity increases exponentially, this does not always imply a higher predictive power, but rather sometimes the

model ends up describing the error path instead of the underlying relation (*overfitting*).

Even if so far we only talked about prediction, estimating the $f$ function is also useful for inference purpose, i.e. understanding how the dependent variable changes based on the predictors used. Inference answers to different questions with respect to the prediction, such as which are the correct independent variables for a certain model, what kind of form $f$ has and so on so forth. In other words[1], *given the purchases you've made on Amazon if I ask what else you might want to buy, that's a prediction problem. But if instead I ask how certain I am that you're gonna wanna buy those things, if Iask for a confidence interval, then that's an inference problem.*

Another crucial aspect to deal with form the beginning is the tradeoff between interpretability and accuracy. Indeed, more flexible is a model, the less interpretable it is.

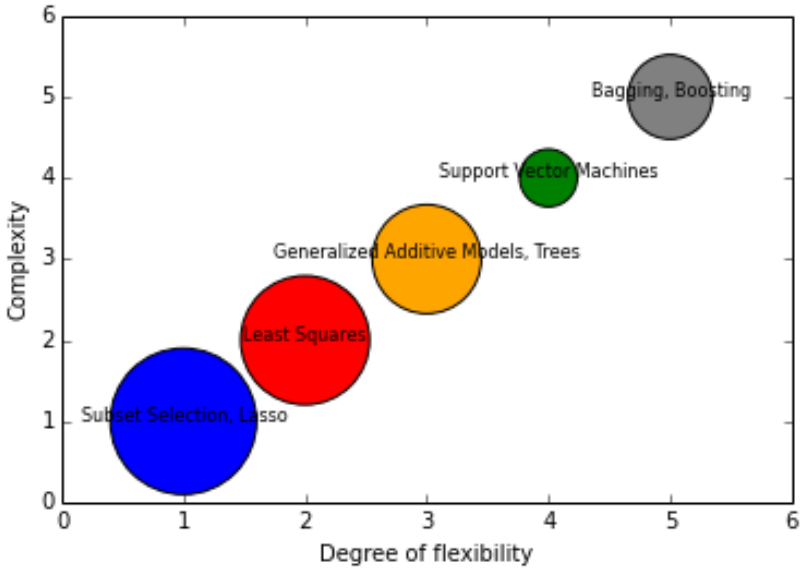



Figure 1.1: Tradeoff between flexibility and interpretability for different methods. The larger is the area of each circle, the easier the interpretation is for that particular model.

We will deal with every model more in details in the following chapters, but as a general rule we can state that if the purpose is to draw some inference simple models perform better, while for pure prediction a complex model is more explanatory and useful.

Before we move on, we should take a while to focus on the last important introductive concept, i.e. how we measure the model performance. The basic idea is to measure the closeness of the prediction to the actual output and in order to do that, the most common measure is the *mean squared error* (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{1.2}$$

Sometimes people make a big mistake measuring the MSE using the *training data*, i.e. the data used in order to estimate the model. In reality, to assess the correct accuracy of the prediction model the MSE for previously unused test data should be computed. We also have to keep in mind a general fundamental rule, i.e. as model flexibility increases, the *training MSE* will decrease while this is not always true for the *test MSE* (in case it does not, we have an *overfitting* issue). Furthermore, it is mathematically provable that the MSE for a given value $x_0$ could always be split into three components, i.e. the variance of the

[1] This is an explanation provided by Professor Daniela Witten in the "Simply Statistics Unconference on the Future of Statistics".

prediction, the squared bias of the prediction and the error variance,

$$MSE_n = E\left(e_{n+1}^2\right) = \sigma_{\hat{y}}^2 + Bias_{\hat{y}} + \sigma_{\hat{\epsilon}}^2. \tag{1.3}$$

Since we cannot intervene on error variance anyway, in order to get the smallest MSE we should try to lower both the estimation variance and the estimation bias, i.e. respectively how much the prediction changes if we change the training set and the error of real-life approximations. Generally speaking, more flexible models have a higher variance as well, but a lower bias (*bias-variance trade-off*).

What we said so far it is also applicable to a classification problem, with some small modifications. The basic method to assess a classification accuracy is called *error rate*, that is the proportion of misclassified observations for the training set used,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(y_i \neq \hat{y}_i\right) \tag{1.4}$$

when $\mathbb{I}$ equals 1 if the estimation and the true value are different and 0 otherwise. Again, using a test dataset we would be able to compute the real rate of classification success.

A simple classifier that minimises the test error rate is the Bayes classifier, which assigns each observation to the most likely class through the conditional probability of y given the observed predictor. Unfortunately, since in reality we do not know the conditional distribution, we usually use the *K-nearest neighbors* (KNN) that estimates the conditional distribution of Y given X and attach the observation to the class with the higher estimated probability. More in details, given a test observation $x_0$ and a positive integer K, the KNN first of all identifies K training observations close to $x_0$, then it estimates the conditional probability for a class $j$ as the percentage of points with value $j$ within the K selected and at the end it applies the Bayes classifier attaching $x_0$ to the most likely class.

We decided to end this chapter explaining two recurrent terms in machine learning: supervised and unsupervised learning. A supervised learning approach involves, broadly speaking, the estimation or prediction of an output given some inputs. We have the output, we have the inputs, and we "only" have to infer the relation between the two. The relevant information to stress here is that the output has to be labeled, in other words it has to be tagged to *feed* the algorithm. The more basics example are the spam filter or a face recognition algorithm. First of all, the user flags what is spam or which is a face, and the algorithm *learns* from the explicit classification made and infers a relation between a certain input and the correspondent output. Hence, the user set the rule, and the algorithm works on it and exploit the user's flags to self-improve. On the other hand, an unsupervised learning approach differs from a supervised one for the output that is, in the unsupervised case, unlabelled or missing. One or more inputs are provided anyway, but no supervised outputs help the algorithm to self-improve. What the algorithm does here is to identify some correlation or some common features to cluster the outputs in a certain way. In other words, the algorithm creates his own rule. For instance, if we do not specify in an image recognition algorithm which is a cat or what is a landscape, the algorithm is not able to understand that by itself, but it can find common features to allocate to some extent each image to the class "cat" or "panorama".

# Chapter 2

# Machine Learning Concepts

## 2.1 Classification

This section talks about model that predicts *qualitative* output, whether for instance the stock price will go up or down tomorrow. We are now going to provide some different techniques to deal with classification problems.

### 2.1.1 Logistic Regression

This regression models the probability of the dependent variable to belong to a certain class and is expressed by the logistic function:

$$p(X) = Pr(Y = y \mid X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{2.1}$$

which gives outputs between 0 and 1 for all the independent variable values. With a bit of manipulation, we come to the *logit* form:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2.2}$$

that interprets the coefficients differently from a linear regression model. Again, even in this case, we should be able to find the coefficients through the least squares analysis, but in practice there exists a method called *maximum likelihood estimation* that has better properties with respect to the OLS:

$$max\ l(\beta) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j)). \tag{2.3}$$

The logit model works well also for response variable with more than two classes, but in practice the following model is more used.

### 2.1.2 Linear and Quadratic Discriminant Analysis

Alternatively to the logistic regression, here we are going to first compute the distribution of the predictors for each of the response classes and then we estimate the conditional one through the Bayes's theorem. Since Linear Discriminant Analysis (LDA) has both less instability problem for well-separated classes and more stability for $n$ small, it is becoming one of the favourite tools for machine learning classification.
First of all, let's assume that $\pi_k$ is the probability that a randomly chosen observation belongs to class $k$th (*prior*), and $f_k(X)$ is the *density function*. Thus, the Bayes' theorem states the *posterior* probability as follow:

$$Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)}. \tag{2.4}$$

So, we have to estimate both the prior and the density function. The prior is usually computed as the fraction of training observations belonging to $k$th class. The density function is instead more complicated, but the LDA general method provides a good way to estimate that function. Let's assume that the predictors are drawn from a multivariate Gaussian distribution., i.e. $X \sim N(\mu, \Sigma)$, and it has the following formal density:

$$f(x) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)}. \tag{2.5}$$

Using now the Bayes' theorem, we get that the Bayes classifier assigns a certain observation $X = x$ to a certain class for which is largest the $\delta_k$, where

$$\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k{}'\Sigma^{-1}\mu_k + log\ \pi_k. \tag{2.6}$$

From this it comes out that the Bayes boundaries are the set for which the $\delta$s are equals. We need then to estimate the unknown parameters vector/matrices, such as $\mu, \pi$ and $\Sigma$ and then to plug them in the delta-function in order to obtain the *discriminant function* $\hat{\delta}_k(x)$.

Clearly, the method is not perfect and sometimes it assigns some points to the wrong class. A common graphic analysis usually used that shows all the possible thresholds for understanding the degree of efficiency of the method is called *receiver operating characteristics* (ROC) curve.

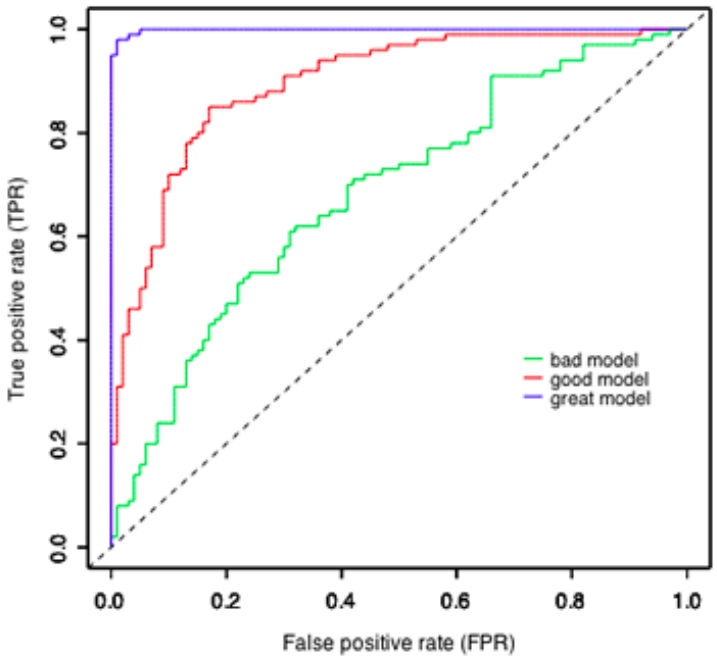



Figure 2.1: ROC curves.

Another variation for the LDA method is the so-called *Quadratic Discriminant Analysis* (QDA), that assumes that each class has its own covariance matrix, so

$$\delta_k(x) = x'\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k{}'\Sigma_k^{-1}\mu_k + log\ \pi_k - \frac{1}{2}log\,|\Sigma_k| - \frac{1}{2}x_k{}'\Sigma_k^{-1}x_k. \tag{2.7}$$

The choice between the LDA and the QDA is taken based on the bias-variance tradeoff. Indeed, LDA is of course less flexible, easier to compute and with a lower variance. On the other hand, the QDA attenuates the bias and thus it performs better in case of many training sets.

To finish the section, we are going to present a graphical comparison between the different models performances ([49]).

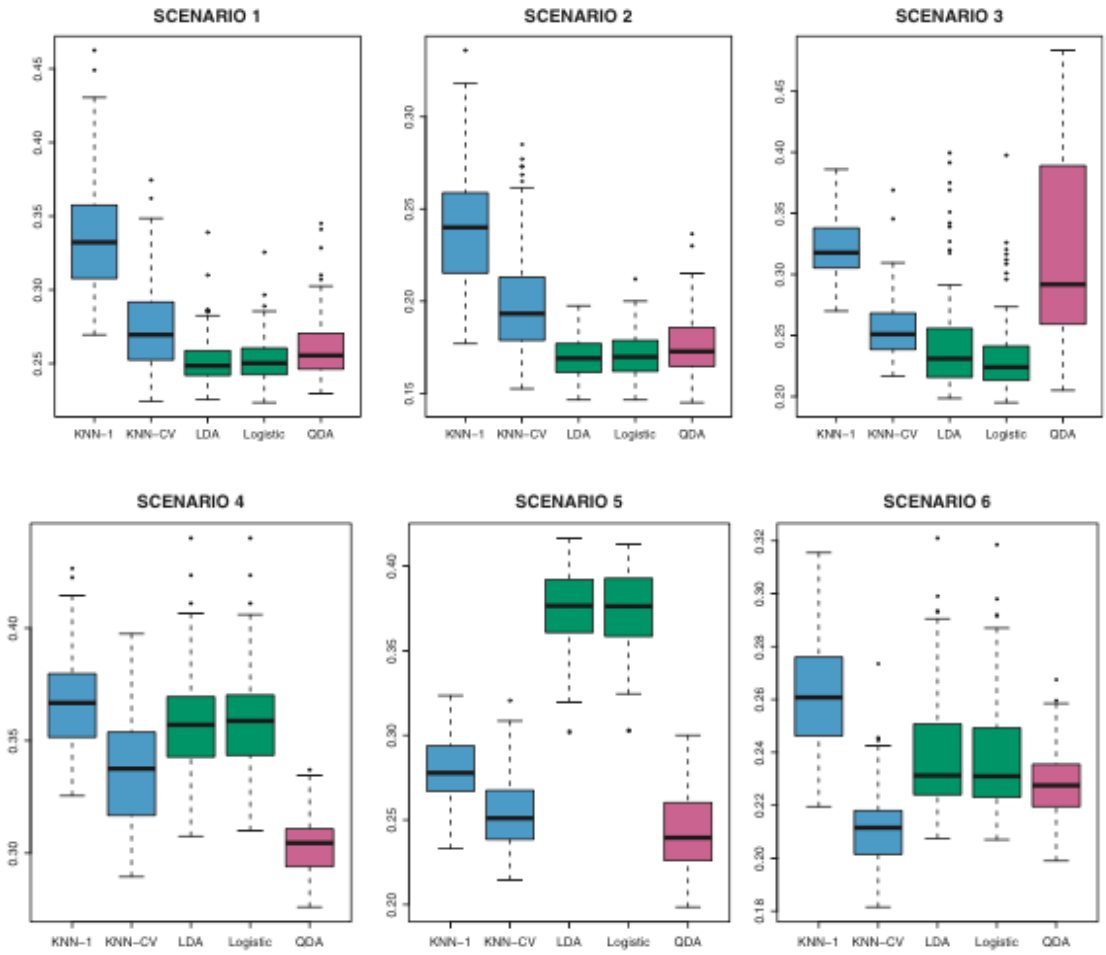



Figure 2.2: Methods comparison in linear (first three) and non-linear (the others) scenarios.

## 2.2 Linear Regression

The regression models deal with forecasting of *quantitative* output, e.g. how much more I earn if improve my education. In a general form, this is represented by:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ...\beta_k X_{ki} + \epsilon_i \tag{2.8}$$

Of course, since we known the predictors $X_i$ and the outputs $Y_i$, here the problem is how to estimate the coefficients to able able to make predictions. The most common approach used to estimate the betas is called *least squares*, that minimises the distance between each point and the average. For regression with more than one predictor the solution is provided below:

$$\beta = (\mathbf{X'X})^{-1}\mathbf{X'y} \tag{2.9}$$

where the bold character means that are vectors (betas and y) or matrix (the predictors X). In the case of single predictor, there are shortcuts and specific simplified formulas to compute the coefficient and the constant term, but we prefer to show a more general approach.

To underline the power of the least squares estimator, we should also stress the it is an *unbiased* estimator in every sample, that means that the estimation never over or underestimate systematically the real parameter and if we average the estimations over the possible values of the predictors we got that the unconditional mean is $\beta$ as well.

Then, two kinds of model accuracy measures have to be computed. First of all, we should verify whether the sample estimate is indeed close to the real value. Hence, we usually compute the standard errors

$$SE(\hat{\beta}_j) = s\sqrt{(X'X)^{-1}_{jj}} \tag{2.10}$$

These errors could be used in order to compute the *confidence intervals* of the coefficients, i.e. the range of values with a certain probability to contain the true value of the parameter. It takes the form

$$\hat{\beta} + z_{\alpha/2} \cdot SER(\hat{\beta}). \tag{2.11}$$

Furthermore, we can also exploit the standard errors to perform some hypothesis tests on the coefficients, i.e. to verify whether for instance there exists a relation between the predictors and the dependent variable:

$$H_0 : \beta_{\mathbf{1}} = \beta_{\mathbf{2}} = ...\beta_{\mathbf{p}} = 0 \qquad VS \qquad H_1 : \text{at least one coefficient} \neq 0. \tag{2.12}$$

What we do in practice in estimating the *F-statistic*

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{2.13}$$

and verify whether the probability of a value to be equal to $|t|$ is small enough to let us reject the null hypothesis (*p-value*).

Once we estimate the regression's coefficients, we may wonder whether we are accurately describing the data or not. There exist different tools to understand what percentage of the model is explained by the regressors selected:

- $R^2$: it is the part of sample variance explained by the regressors, i.e.

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon_i}^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \tag{2.14}$$

  that is the explained sum of squares on the total sum of squares or the complement to one of the sum of residuals over the total sum of squares. Thus, the $R^2$ varies between 0 and 1, depending on whether it has zero or a perfect explanatory power.

- *Standard error of the regression* (SER): is the standard deviation of the error terms and it measures the dispersion of the observations around the regression line:

$$SER = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon_i}^2. \tag{2.15}$$

Finally, it is also very important to understand, through the p-values, whether an independent variable is actually necessary in order to improve the explanatory power of the regression. Since sometimes we cannot run all the model with every possible combination of coefficients, we need an efficient way to select which variables to use in the regression:

- *Forward selection*: first of all, we set a regression with no coefficient but intercept and then we run $n$ regressions. Hence, we add to the *null model* the variable that entails the regression with the lowest RSS and we iterate the procedure until some stopping rule is met.
- *Backward selection*: on the contrary, here we start with a full model including all the possible predictors and then we remove the variable with the largest p-value associated. Then the model is fit and another variable and the procedure is reiterate until some stopping rule is met. This approach cannot be used if the coefficient to estimates are more than the observations used in order to estimate them.
- *Mixed selection*: it works as the forward selection, but every time a variable reach a too high p-value is then removed. The procedure goes on until all the variables remained have a low p-value.

There may be also several potential issues, such non-linearity, correlation between or non-constant variance of the error terms, presence of outliers, or multicollinearity, that will be omit right now since they are not the purpose of this work, but they may be dealt with during further applications.

## 2.3 Resampling

Resampling basically means to drawn more than once a sample from the training set and thus refitting the model based on the new sample, in order to get additional information. Two of the most common resampling method are cross-validation and bootstrap.
When there is not a designated test set that would be able to compute the test error, we have to use some methods able to estimate that. There exist several techniques, and we are going to explain four of them. However, even if they will be applied to a regression context, they can be easily adapted to a classification environment through with the precautions explained in the section 1.

- *The Validation Set Approach*: it involves to divide the available set into two almost-equal part, and use one subset as training set and the other as validation set, i.e. fitting on the first one and predicting with the second one. The resulting MSE is a good proxy for the test error rate. The two issues related to this method are I) the validation MSE is highly variable (it depends on what is included in the sample used) and II) it may overestimate the real error rate, since there are less observations than what it could be.
- *Leave-One-Out Cross Validation* (LOOCV): it works exactly as the previous one, except that for the validation set only one observation is attached. The procedure is repeated for every observation and the amount of all the MSE associated are then averaged. In this way, both the problems related to the validation approach are solved, but unfortunately the several steps make the method quite computationally expensive. Only in case of least squares or polynomial regression, we can reduce this inefficiency dividing each MSE for $(1-h_i)^2$, where $h_i$ is the leverage statistics and measures how much an observation influences his own fit:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}. \tag{2.16}$$

It represents those points that have unusual values for $x_i$ (that is different from the *outliers*, for which the response in unusual).

- *k-fold Cross Validation*: again, it is exactly like the previous method, except that the initial set is divided into $k$ equal-size subsets, and then model is validated on the first one of these. As before, the estimate error rate is given by averaging all the MSEs that come out from using every different fold as a validation set. Of course, with respect to the previous one the k-fold method requires less computational effort and gives also more accurate estimates. Indeed, even if the LOOCV gives almost unbiased estimator, it also has a higher variance, since every model is fitted on almost-equal training set and thus they are strongly correlated.
- *The Bootstrap*: it is a very flexible instrument that samples repeatedly *with replacement* from the original dataset.

## 2.4 Model Selection

We have seen so far that for every model we choose to implement, we always have to take care of both accuracy and interpretability. From the accuracy point of view, sometimes we find overfitting phenomenon, or simply a too high variance, or still an infinite variance in case the predictors are more than the observations. In this case it would be required to reduce the number of predictors through some techniques. On the other hand, we should also leave out irrelevant variables, in order to be able to better interpret the model.
The big data era involved further studies on high-dimensional problems. In this setting, the issue regards the fact that the predictors are more than the observations. When it happens, even if it exists some relation between X and Y, the OLS for instance will come out with coefficients that perfectly fit the data and we now from theory that this is impossible (overfitting). Another problem is called in literature *curse of dimensionality*, meaning that the test error increases as dimensionality grows (unless the predictors are truly related to the responses). Hence, we are going to talk about several methods that allow us to select the best variables for a certain model and to reduce the dimensionality.

### 2.4.1 Subset Selection

Here what we do is exactly select the predictors we think are related to the dependent variables. There exist several different subset methods we may use to select the best variable to fit the regression and we are going to show some of them here.

A first example is the *Best Subset Selection*. We first of all fit all the $p$ models that contain only one predictor, then all the $\binom{p}{2}$ models which contain two predictors, and so on so forth, and for each step we select the model with the lowest RSS. Then, we will come up with a series of "best" models, from the one with no predictors up to the one with all the $p$ predictors used, and we select the most performing one with one the following criteria that indirectly estimate the test error adjusting the training one:

- $\boldsymbol{C_p}$: for a model containing $d$ predictors,

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{2.17}$$

- **Akaike informaiton criteria** (AIC): for a model containing $d$ predictors,

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2) \tag{2.18}$$

- **Bayesian information criteria** (BIC): for a model containing $d$ predictors,

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2) \tag{2.19}$$

- **Adjusted** $R^2$: for a model containing $d$ predictors,

$$Adj.R^2 = 1 - \frac{RSS/(n-d-1)}{TSS(n-1)} \tag{2.20}$$

Since they add somehow penalties to the training RSS, they have to take a low value for models with low test errors, except the last one, which works on the other way round.

Alternatively, we can try to directly estimate the test error using a validation or cross validation approach.

A second class of subset selection methods concerns the *stepwise algorithms*, that are less computationally demanding and have less problems when the number of predictors is quite large:

- Forward stepwise: starting from the null model, it adds one-at-a-time predictors until every predictor is in the model. The crucial thing is that each step it is added the variables which entails the most intensive marginal contribution. Hence, if for instance we have four predictors, we start from a model with no predictors at all and we run four different models with each one a different predictor. Then, we investigates which one has the lowest RSS. We repeat the procedure starting from the model selected in the previous step and finally we identify the best model through the ones so obtained using one of the criteria shown above. This clearly means that we will eventually have not as many combinations as the subset selection had.

- Backward stepwise: it works exactly like the forward model, but instead of starting from the null model it begins with a fully-completed model and then each step it removes one single predictor.

### 2.4.2 Shrinkage

This approach estimates the model with all the predictors, but then some of them are shrunken (also to zero) in order to reduce the variance. The most known techniques of this kind are the *ridge regression* and the lasso.

The ridge regression differs from the OLS because of the $\lambda$ shrinkage penalty applied:

$$min \quad \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.21}$$

It will clearly produce a different set of parameters for each $\lambda$ used, and contrarily to the OLS it is not scale invariant, that is why it usually needs to be standardised before applying ridge regression. As we have seen before, the ridge method has of course less flexibility with respect to an OLS estimation (and lower variance and higher bias as well), but it actually performs better since for $\lambda$ increasing, it reduces the number of independent variables such that to a variance drop is associated a less proportional increase in the bias. As final note, the estimation of the tuning parameter $\lambda$ can be easily approached through cross-validation techniques.

The second method is instead the *lasso* and tries to compensate one drawback of the ridge regression, i.e. the presence of all the predictors (they are actually shrunken toward zero, but never exactly equal to):

$$min \quad \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{2.22}$$

The penalty term here has the effect of forcing some coefficient to be exactly zero, performing also a sort of variable selection.

Between the two methods provided in this section, there is not unequivocally one that is the best, but it depends on different applications and circumstances. Broadly speaking, the lasso works better in fewer predictors environment, while the ridge regression in case of several predictors roughly equal to zero.

### 2.4.3 Dimension Reduction

The predictors are combined in some of the possible ways and these combinations are then used as predictors. Indeed, we use the $M$ linear combinations of original predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{2.23}$$

in order to fit the model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i. \tag{2.24}$$

However, the issue is how to find the right parameters $\theta$s, and this could be obtained through one of the following two methods:

- **Principal Component Analysis** (PCA): it is one of the most powerful unsupervised learning tools and it is used to reduce the dimensionality of the predictors matrix. The *first principal component* is the direction in which the observations vary the most. The underlying assumption here is that the directions in which the variability is the highest are the ones associated with the response variables. Furthermore, what PCA does is finding a low-dimensional representation of a highly variable data. In particular it exploits only that interesting variables, where this means that along that dimension the variability is quite high. Hence, the first principal component is given by the normalised combination of variables with the highest variation:

  $$Z_1 = \phi_{11} X_1 + ... \phi_{p1} Xp. \tag{2.25}$$

  So what we practically do is solving the optimisation problem

  $$\underset{\phi_{11},\ldots,\phi_{p1}}{max} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \tag{2.26}$$

  $$subject\ to \sum_{j=1}^{p} \phi_{j1}^2 = 1. \tag{2.27}$$

  Once we found the first principal components, the second ones are given by the linear combination of predictors uncorrelated with the first principal components. In addition, for the PCA is relevant to notice that the results strictly depend on different scaling or not each variable and that the each principal component vector is unique (regardless who performs the analysis or what software uses). A last important thing to stress is that only a part of the variability is explained by the PCA, i.e.

  $$Proportion\ of\ Variance\ Explained\ (PVE) = \frac{\sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} \tag{2.28}$$

- **Partial Least Squares** (PLS): it is the supervised learning counterpart of the PCA. It again reduces the independent variable dimensionality, it uses liner combinations of predictors to fit the model but it does it exploiting the response variable, i.e. in a supervised way. Hence, it identifies new variables as good proxy of the old ones, but also selects the ones that are (more) related to the responses. First of all, it standardised the predictors as usual because it is not scale invariant, and matches the coefficients in (2.23) with the $\beta$s as the first direction $Z_1$. Finally, it assigns the highest weight to the $X$ most tightly associated to the response and the $\mathbf{z}$ are then regressed on $\mathbf{y}$. The last step is to orthogonalised $\mathbf{x}$ with respect to $\mathbf{z}$. We now repeat the process M times in order to obtain M directions, computing the second direction and the following ones adjusting the variables for the residuals (the information not used yet by the first direction) of regression on Z. Finally, the model is fitted with an OLS, with Y and $\mathbf{Z}$ as dependent/independent variables.

## 2.5 Nonlinear Models

Of course in reality not everything could be explained through a linear relation, and this section handles exactly the idea of non-linearity. Setting a non-linear model is somehow an additional degree of flexibility and involves an improvement sometimes over the linear model seen so far. We are going to provide now several extensions:

- **Polynomial Regression**: it adds additional predictors raising some of the original ones to the power (usually no more than the third of fourth one):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i. \tag{2.29}$$

- **Step Functions**: clusters the variables into $d$ groups (*bins*) and creates a qualitative variable:

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + ... + \beta_d C_d(x_i) + \epsilon_i \tag{2.30}$$

where

$$\begin{gathered} C_0(X) = \mathbb{I}(X < c_1) \\ C_1(X) = \mathbb{I}(c_1 < X < c_2) \\ \vdots \\ C_{d-1}(X) = \mathbb{I}(c_{d-1} < X < c_d) \\ C_d(X) = \mathbb{I}(c_d < X). \end{gathered} \tag{2.31}$$

- **Regression Splines**: it runs a polynomial regression on a step function, constraining the polynomial to smoothly join at the boundaries (*knots*):

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + ... + \beta_{d1} x_i^d + \epsilon_i & \text{if } x < c_1 \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + ... + \beta_{d2} x_i^d + \epsilon_i & \text{if } c_1 < x < c_2 \\ \vdots & \\ \beta_{0d} + \beta_{1d} x_i + \beta_{2d} x_i^2 + ... + \beta_{dd} x_i^d + \epsilon_i. & \text{if } x > c_{d-1} \end{cases} \tag{2.32}$$

Using more knots implies a higher flexibility, and in order to get a better behaviour, we should also constrain the function to be smooth enough (first and second derivatives continuous). The problem is in the number of knots we should select and where we should insert a knot. Even if it should make sense to put more knots in areas in which we believe the function varies the most, in practice they are usually put uniformly. Concerning the number, it is a trial-and-error process; otherwise, it could be obtained through cross-validation. However, the splines regression provides more stable estimates with respect to a simple basis function such as step or polynomial, and that is why is preferred in practice.

- **Smoothing Splines**: similar to the previous one, but here the intent is to derive a function $g(x)$ (*smoothing spline*) that fits the data very well (thus reducing a lot the RSS) but that is also smooth enough:

$$min \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt. \tag{2.33}$$

where $\lambda$ is once again the tuning parameter as in lasso/ridge regression. Hence, this is again a function that incorporates a loss term and a penalty for the variability. The function $g(x)$ that minimises the (2.33) can be proved to be a piecewise continuous cubic polynomial with knots at $x_1, ..., x_n$, but it is more shrunken with respect to a basis function because of $\lambda$. Moreover, this parameter is chosen making the cross-validated RSS as small as possible.

- **Local Regression**: it concerns the computation of a target point using only *close* training observations. So, we should take the fraction $s$ of points that are closest to our target $x_0$. We assign to these points a weight inversely proportion to the distance from $x_0$ and every point out of the area chosen gets a zero weight. Then, we implement a weighted least squares regression using the weights above obtained. It is pretty clear that from the choice of the fraction $s$ depends the flexibility of the model, so one way to select the optimal $s$ is again through cross-validation.

- **Generalised Additive Model** (GAM): it includes non-linear functions although it keeps the additivity property of linear models:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i. \tag{2.34}$$

This approach allows us to model automatically non linear relations, which could result in a better prediction for the responses. Furthermore, since it is still additive, it lets us studying the individual impact on every independent variable on Y. On the other hand, this assumption may let us miss some crucial interactions.

## 2.6 Tree-based Methods

In this section we are going to provide some insights on this powerful decision tools called *tree method*, that involves a stratification of the predictor space into several areas and we formulate predictions based on (usually) the mean of every single region. Unfortunately, if from one hand these models are really easy to interpret, from the other side they are not so accurate as other form of supervised learning.

The algorithm to obtain a *tree* is to divide the feature space into $J$ distinct non-overlapping regions such that the RSS is minimised:

$$min \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - y_{\hat{R}_j}\right)^2. \tag{2.35}$$

It is not possible to explore every combinations of this kind, but we can use a top-down greedy approach called *recursive binary splitting*, i.e. starting from the top it splits the tree at each as best as possible at each nodes without taking care of the future. Using this useful tool, we would no more minimise the RSS as expressed before, but we have to do that separating the RSS for each branch. So, for each predictor we consider all the values of the cutpoint $s$ and we choose the predictor-cutpoint pair which result in the lowest RSS:

$$min \sum_{i:\ x_i \in R_1(j,s)} (y_i - y_{\hat{R}_1})^2 + \sum_{i:\ x_i \in R_2(j,s)} (y_i - y_{\hat{R}_2})^2. \tag{2.36}$$

The process is then repeated for each region identified previously and the responses are predicted using the training mean for each region. The real issue is that many times this procedure overfits the data because of the excessive complexity of the data. However, there exists a technique called *pruning* that may help in building a large tree avoiding informational losses. Indeed, by imposing a cost for complexity, we would be able to prune the tree backward. A common measure of complexity is the number of the *terminal nodes (leaves)* and the related cost is measured by a tuning parameter that provides the best out-of-sample predictions. In fact,

$$\sum_{m=1}^{|T|} \sum_{i:\ x_i \in R_m} (y_i - y_{\hat{R}_m})^2 + \alpha |T| \tag{2.37}$$

where $T$ indicates the number of terminal nodes and $\alpha$ is the tuning parameter. This function has to at minimum. We thus obtain a sequence of subtrees as a function of the tuning parameter, and then we can use K-fold cross validation to pick the right $\alpha$ that minimises the average error.

The trees may also be build for classification purposes, if we think in term of how each observation belongs to the most commonly occurring class. The other difference with respect to the trees for regression is that the criteria for making the binary splits are others, such as the *Gini index* (measuring the total variance around the classes)

$$G = \sum_{k=1}^{K} \hat{\rho}_{mk}(1 - \hat{\rho}_{mk}) \tag{2.38}$$

where $\hat{\rho}_{mk}$ is the proportion of training observation in the $m$th region from $k$th class, and the *cross-entropy*

$$D = -\sum_{k=1}^{K} \hat{\rho}_{mk}\ log\hat{\rho}_{mk}. \tag{2.39}$$

Some final general remarks on trees methods. This class of models are not univocally better than the linear models, but there are some circumstances in which they outperform the linear regressions. The strong aspect of the trees is their simplified explainability and their graphical representation. Besides, they can easily deal with qualitative variable without creating dummies and are intuitively more close to the human thinking and decision making process than the regression analysis. Their prediction accuracy is not as high as the one from the regression, but many models have been created to try to improve this weakness.

### 2.6.1 Bagging

As said before, one problem of the previous tree may be the high variance due to the splits of the training data. This issue could be tackle with a bootstrapping technique called bagging and sacrificing a bit of interpretability. What we should do is to take many training sets and estimate different prediction models for each one. The results would then be averaged since this reduces the total variance:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{\,b}(x). \tag{2.40}$$

In practice we do not have many training sets, so we bootstrap many times from a single sample and that is what the bagging does. Furthermore, the test error is pretty easy to compute, since we can use the *out-of-bag* observations not used to fit the model.

### 2.6.2 Random Forests

A further improvement is provided by the random forests, that gives a way to decorrelate the trees. It bootstraps as the bagging method, but at each split it takes into account a random sample of $m$ predictors to be used as split candidates (typically we use $\sqrt{p}$ of the $p$ total predictors). In this way, we avoid that if there is a single stronger predictor, it instills the same shapes to all the trees created and thus a very high correlation.

### 2.6.3 Boosting

The final model presented in this section is the boosting, that is a general approach similar to the bagging, but the trees are grown sequentially. Indeed, each tree is going to be built exploiting the set of information contained in the previous ones. There is not any bootstrapping here, but only a series of modifications over the original dataset. It actually learns slower but it will avoid overfitting situations. So, we fit a decision tree to the residuals and then we add this fitted function in order to update the residuals.

## 2.7 Support Vector Machines

This is the last section of the introductive chapter, and it will deal with Support Vector Machine (SVM), a generalisation of the so-called maximal margin classifier, that represents a limited tool since can only be applied to classes with linear boundaries.

Let's define an hyperplane as a flat affine (not needed to pass through the origin) subspace of dimension $p-1$ in a $p$-space, or also

$$\beta_0 + \beta_1 X_{1i} + ...\beta_k X_{ki} = 0. \tag{2.41}$$

It is very natural and easy to build a classifier based on that hyperplane, that measures whether an observation stands from one side or the other to the hyperplane and with what *magnitude* (the closeness

of the point to the hyperplane).

In practice, if there exists a hyperplane that separates perfectly two classes fro instance, this implies that there are an infinite number of those. It would be definitely better to establish an optimal hyperplane that has the farthest distance from the training set (the so-called *margin*). In order to mathematically construct this classifier, we should solve the following system:

$$\underset{\beta_0,\beta_1...\beta_p}{max} M \tag{2.42}$$

$$subject\ to\ \sum_{j=1}^{p} \beta_j^2 = 1 \tag{2.43}$$

$$y_i(\beta_0 + \beta_1 x_{1i} + ...\beta_p x_{pi}) \geq M \tag{2.44}$$

where $M$ is the margin of the hyperplane. The two constraints make sure that each observation in one the correct side of the plane and that the distance between them and the plane is at least $M$.

The only problem arises in case in which a hyperplane able to separate the observations does not exist at all, or it is only able to almost classify the observations. In this situations, we switch to a different classifier called *Support Vector Classifier*. This classifier does not exactly separate the observations, but it has a greater robustness and a better classification of most of the training observations. The mathematical solution for this new classifier is provided by the system

$$\underset{\beta_0,\beta_1...\beta_p,\epsilon_1,...,\epsilon_n}{max} M \tag{2.45}$$

$$subject\ to\ \sum_{j=1}^{p} \beta_j^2 = 1 \tag{2.46}$$

$$y_i(\beta_0 + \beta_1 x_{1i} + ...\beta_p x_{pi}) \geq M(1 - \epsilon_i) \tag{2.47}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \tag{2.48}$$

where $C$ is a tuning parameter (nonnegative, a sort of budget constraint), and $\epsilon_1, ..., \epsilon_n$ are slack variables that allows a margin/hyperplane misclassification.[1] $C$ is in practice chosen via cross-validation, and there is not a way ex-ante to define that.

The interesting difference of this classifier is that, if the previous one was only affected by the the closest observations to the hyperplane, the support classifier is affected only by observations on the margin or that lie over the margin. In any case, the observations able to affect the hyperplane slope are called *support vectors*.

The final (and the most general) extension of this class of model is the Support Vector Machine. It enlarges the feature space through the *kernels* to take into account non-linear boundaries. Basically, the solution of the previous system could be expressed by the $\binom{n}{2}$ inner products of the observations, i.e.

$$\langle x_i, x_i' \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}. \tag{2.49}$$

Instead, in the SVM case, let's assume that this inner product is going to be generalised by the kernel $K$ that quantifies the similarities between two observations,

$$K\ (x_{ij} x_{i'j}). \tag{2.50}$$

[1] If $\epsilon_i$ is greater than 0, the observation will be on the other side of the margin; it is greater than 1, it will be on the other side of the hyperplane.

As soon as the kernel selected is non-linear, in that case we have a SVM classifier (if it was linear, we came back to the support vector classifier), with a functional form of

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \; K(x_{ij} x_{i'j}). \tag{2.51}$$

Of course, the main advantage from using the kernel is computational, since it uses only $\binom{n}{2}$ pairs combinations. However, the SVM could be easily extended to more than two classes and it is interesting to notice that it is closely related to the logistic regression above mentioned. More in detail, the *hinge loss* is strongly connected to the loss function presented in the logistic regression

$$L(\mathbf{x}, \mathbf{y}, \beta) = \sum_{i=1}^{n} max \; [0, 1 - y_i(\beta_0 + \beta_1 x_{1i} + ...\beta_p x_{pi})]. \tag{2.52}$$

## 2.8 Clustering methods

The clustering methods are usually used to identify homogeneous subgroups in a heterogeneous dataset. The most common approach is called *k-means*, and it classifies observations into $K$ non-overlapping clusters. Each observations have to belong to at least one cluster by definition, and the basic idea of the method is to give back the lowest within cluster variance $W(C_k)$ for the clusters $C_1, .., C_k$:

$$\underset{C_1,...C_k}{min} \sum_{k=1}^{K} W(C_k) \tag{2.53}$$

The minimisation, set the functional form of $W(C_k)$ to be a squared Euclidean distance, is

$$\underset{C_1,...C_k}{min} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \tag{2.54}$$

Broadly speaking, what we should do to implement the k-means algorithm is assigning *randomly* from 1 to $K$ a label to each observation. Then, compute the centroid[2] for each cluster, and finally attach each observation to the cluster whose the centroid is the lowest. We should repeat the procedure until a local optimum is reached, i.e. until the situation does not change further.

A second very well-known clustering procedure is the *hierarchical clustering* (HC). The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant. Moreover, the HC is very used because unlike k-means, it has no parameter $K$ to be selected, and it has a quite nice graphical tree representation as well (*dendrogram*). The basic idea behind the HC algorithm is to attribute create one cluster for each observation and then try to fuse them through Euclidean dissimilarity measure. The procedure is then reiterated, until the observations belong to a single cluster (and so the dendrogram is completed with a bottom-up approach). An important problem may arise in how we define the dissimilarity measure for clusters with multiple observations. That is why we need to extend the Euclidean measure used for observations to clusters. This new operator is called *linkage*. We provide the following table to summarise information about a categorisation of linkages provided by Hastie et al. ([49]).

[2] The centroid is defined as the vector containing the means for the observations in a certain cluster.

| *Linkage* | *Description* |
|---|---|
| Complete | Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities. |
| Single | Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time. |
| Average | Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities. |
| Centroid | Dissimilarity between the centroid for cluster A (a mean vector of length $p$) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*. |

Table 2.1: Summary of different most common linkage definitions.

# Chapter 3

# Time Series Analysis

The machine learning techniques describes so far are usually applied to cross-sectional panel datasets. Time series analysis is more cumbersome and sometimes represents a problem. We are going to then focus on this topic and we decided to structure the following chapter as follows: first of all, some notions of time series are presented. Later on, an overview about machine learning for prediction will be wrapped up using concepts from the first chapter as well, and finally we will see the local learning, the forecasting and a list of models, to end up with a comparison between those methods.

## 3.1 Time Series Review

Beginning with a basic definition, a time series is a sequence of observations ordered in time. Since in practice the recording of the data is discrete, we are going to focus only on discrete series. As proposed in Bontempi ([15]), a general model for them is

$$s_t = g(t) + \varphi_t \qquad for\ t = 1, ..., T \tag{3.1}$$

where the first part is called systematic component names *signal* and the second is a stochastic error term usually called *noise*. What we usually do is trying to decompose this model into three pieces: a trend, a seasonal effect and a noisy fluctuation. Only once we deseasonalized and detrended we are able to provide accurate relation between past and future, that is formally described by probability notation. In this sense, a time series is seen a realisation of a stochastic process, with a joint density of the form

$$p\ (\varphi_1, ..\varphi_T). \tag{3.2}$$

Making a prediction about a time series is only possible if the underlying relation observed in the past persists in future. This is expressed by the concept of *strictly stationarity*, i.e. regardless the time span (for instance, from 1 to $T$ or from $1 + i$ to $T+$) the joint distribution is the same. In other words, none shifting has any effect of the joint distribution that depends only on the interval. From this definition it derived that the first two moments of the distribution are finite and that the autocovariance function is only related to the lag,

$$E[\varphi_t] = \mu_t = \mu \tag{3.3}$$

$$Var[\varphi_t] = \sigma_t = \sigma \tag{3.4}$$

$$\gamma(k) = Cov[\varphi_t, \varphi_{t+k}] = E[(\varphi_t - \mu)(\varphi_{t+k} - \mu)] \tag{3.5}$$

with an autocorrelation function like

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}. \tag{3.6}$$

On the other hand, a less strict definition is the one of *weakly stationary*, that means having a constant mean and an autocorrelation function depending only on the lag. Of course, strictly stationary implies weakly but not viceversa, unless the process is *normal*.

There exist several stochastic processes used in finance and economics, and we present a non-exhaustive list of the most common ones:

- *Purely Random Process* (**white noise**): a series of mutually independent and identically distributed variables, with constant mean and variance and zero autocorrelation

$$p(\varphi_{t+k} \mid \varphi_t) = p(\varphi_{t+k}). \tag{3.7}$$

When a process of this kind is a discrete one, it may be called *random walk*, i.e.

$$\varphi_t = \varphi_{t-1} + \omega_t \qquad \Rightarrow \qquad \varphi_t = \sum_{i=1}^{t} \omega_i \tag{3.8}$$

and, since its mean and variance depend on $t$, it is clearly non-stationary. In case, you can be made stationary if we take the first difference of the random walk (i.e., the difference between the realisation at time $t$ and $t-1$).

- *Autoregressive Process (n)* (**AR**): a process in which the value at $t$ is given by a weighted sum of previous realisations and a random shock

$$\varphi_t = \alpha_1 \varphi_{t-1} + \alpha_2 \varphi_{t-2} + ... + \alpha_n \varphi_{t-n} + \omega_t \tag{3.9}$$

The stationary here depends on the parameters $\alpha_i$. Moreover, the process has zero mean and, for $|\alpha| < 1$, we have a second moment and an autocorrelation such as

$$\begin{aligned} Var[\varphi_t] &= \tfrac{\sigma_\omega^2}{(1-\alpha^2)} \\ \rho(k) &= \alpha^k. \end{aligned} \tag{3.10}$$

The unknown variables for this kind of process are the order $n$ and the parameter $\alpha$. The first one is usually found through the analysis of the autocorrelation function, while the second is estimated by least squares minimizing the difference between the value at $t$ and the sum of the past values.

- *Non-linear Autoregressive Process* (**NAR**):

$$\varphi_t = f(\varphi_{t-1}, \varphi_{t-2}, ..., \varphi_{t-n}) + \omega(t). \tag{3.11}$$

Of course, the NAR is more difficult to be computed with respect to a linear AR process and it has statistical properties not easily identifiable, but it fits better the reality.

- *Moving Average Process (q)* (**MA**):

$$y_t = \omega_t + \phi_1 \omega_{t-1} + ... + \phi_q \omega_{t-q} \tag{3.12}$$

where the first moment is constant, and the auto covariance is given by

$$\gamma_\tau = cov(X_t, X_{t+\tau}) = \begin{cases} \sigma^2 \sum_{s=0}^{q-|k|} \phi_s \phi_{s+k} & \text{if } |\tau| < q \\ 0 & \text{if } |\tau| > q \end{cases} \tag{3.13}$$

- *Autoregressive Moving Average Process (p,q)* (**ARMA**):

$$X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \sum_{s=0}^{q} \theta_s \epsilon_{t-s}. \tag{3.14}$$

If the original process is non-stationary, we can analyse the $d$-order difference (*ARIMA*).

- *Generalised AutoRegressive Conditional Heteroskedasticity* (**GARCH**):

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2. \tag{3.15}$$

## 3.2 Machine Learning for Time-Series Prediction

So, as we have seen before, we would like to investigate how machine learning could help us in forecasting the value at $t$ given the series of the past values through a supervised learning approach. To start, we can assume a general process having zero mean and independent distributed errors (from the predictors).

When we try to tackle a problem of this kind, we should be able to pick a hypothesis that approximates quite well the real function (*structural estimation*), but also on the base on the function chosen and of the training set used, to estimate a set of parameters that approximate the function as best as possible (*parametric identification*). So, we had implement a model selection algorithm that, at the meantime, runs an inner parametric identification loop ([15]). This loop has the goal of minimising the training error through a multivariate optimisation approach, while the model assessment could be set to be

- *Complexity based with a penalty criterium*: we have already discussed some model of this kind before, but broadly speaking this class embeds models such that they have a mean squared error term and a complexity term that grows up with an increasing number of parameters. Different examples are, in addition to the AIC and $C_p$ seen before, the

$$Final\ Predictor\ Error : \hat{MSE}_{empirical}\ (\alpha_n)\ \frac{1+(n+1)/N}{1-(n+1)/N} \tag{3.16}$$

$$Generalised\ Cross\ Validation : \hat{MSE}_{empirical}\ (\alpha_n)\ \frac{1}{(1-p/N)^2} \tag{3.17}$$

$$Predicted\ Squared\ Error : \hat{MSE}_{empirical}\ (\alpha_n) + 2\hat{\sigma}_{\omega}^2 \frac{n+1}{N}. \tag{3.18}$$

- *Data-driven (with validation techniques)*: as we have previously seen, we can test our model on an additional new dataset (rare in practice), otherwise we should split our dataset in two or $k$ to estimate a good approximation of the error rate.

Once the parametrisation is complete, we should take care of the best model to use. Again, we have in the first chapter seen some standard techniques, and we are going to stress again a couple of those for time series analysis:

- *Winner-Takes-All Algorithm*: what it does for $S$ models is running through a leave-one-out validation a parametric identification $N$ times and compute the MSE associated to each realisation. Then, the best model (the one with the lowest MSE) is selected and a final parametric identification (and output prediction) is obtained.

- *Combination of Estimators*: in this class there are procedure like bagging or boosting for instance.

Finally, we have discussed before that we should be able to select ex-ante the best feature for our prediction model, and for time series this is usually done through *filter method*, that assesses goodness of a feature from the data itself (PCA, clustering), a *wrapper method*, that exploits subset of variables (stepwise regression) or *embedded method*, that selects the variables in the learning process (CART, random forests, lasso, etc.). Now, we are going to provide some known methods to implement time series forecasting using machine learning.

## 3.3 Local Learning

A local learning (LL) procedure is a powerful tool since it does not assume anything *a-priori*, it does not need time-consuming retraining as soon as new data can be used and can be used in non-stationarity case ([18]). It involves computing a forecast using a subset of past values. For example, we can briefly talk about two LL methods, such as the *Nearest Neighbour* (NN) and the *Lazy Learning*. The NN case basically assumes that a certain point (the current state) will behave as his neighbours. Hence, we compute the distance between a certain point $x_i$ and the training samples, then we rank the neighbours with respect to the distance they have form that point and selecting a subset using a *bandwidth* for the

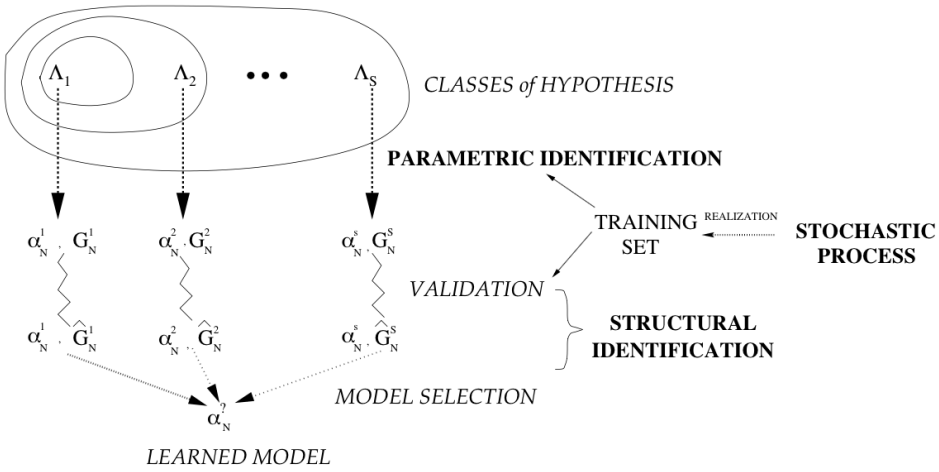



Figure 3.1: This is a schematic representation of the model selection process as provided in [15], [18].

size of the neighbourhood. Finally, we need to fit a local model.[1] The lazy approach instead adapts the size of the neighbourhood on a cross-validation basis. We can cross-validate the process repeating $k$ times the steps so far pointed out. Alternatively, there is an easiest and less expensive way for linear model to do that, i.e. through the *Prediction Sum of Squares* (PRESS). We first of all estimate the linear regression coefficients, and return the outcomes as by-product the Hat matrix

$$H = X(X'X)^{-1}X'. \tag{3.19}$$

It is easy now to compute the residuals, and thus the PRESS statistics will be

$$(y_j - x_j{}'\hat{\alpha})_{l-o-o} = \frac{y_j - x_j{}'\hat{\alpha}}{1 - H_{jj}} \tag{3.20}$$

and the correspondent MSE is given by

$$\hat{MSE}_{l-o-o} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{y_i - \hat{y}_i}{1 - H_{ii}}\right\}^2. \tag{3.21}$$

## 3.4 Forecasting

Once we have a mode, we can try to get a forecast out of our data and in case we can then fed back that value for the further prediction (*iterated prediction*).

In general, the one step-ahead forecasting may be obtained by iterating the one step predictor with the minimum training error on either the one step-ahead forecast or the $h$-steps-ahead forecasts. Otherwise it could be directly estimated, making a prediction at $t + 1$ or extending the input at each step with the forecast outcomes. The last alternative is to use a multi-input multi-output (MIMO) combination to return the forecasts.

In the first case, the result obtained is given back to the algorithm as a new input (as opposed to actual observation), but from the literature ([15]), we already know that in the long run this structure fails since the errors accumulate at each step. That is why we use sometimes the second version, that takes into account a multi period strategy (e.g., Recurrent Neural Networks as in [16] and [63]).

The case of direct estimation instead ([23], [99]), envisages $H$ different models from which it infers a multi-step forecast. There exist several variations of it, as for instance neural networks ([55]), nearest neighbours ([80]) and decision trees ([93]). Unfortunately this class of model has the problem to be

[1] A very narrow band will involve an overfitting issue and so a high variance, while on the other hand a too large one an underfitting problem and thus a high bias.

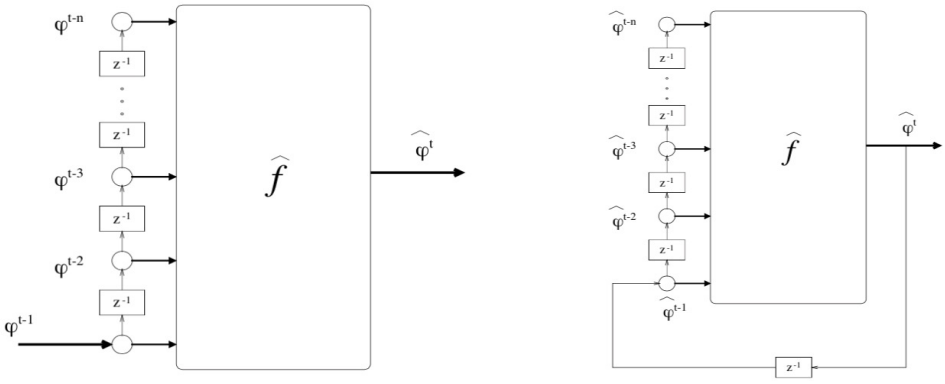



Figure 3.2: Bontempi ([15]) provides this graphical interpretation of the one step-ahead forecasting procedure (on the left), and about the iterated prediction (on the right).

highly complex and requires both computational power/time and a deeper analysis since no dependencies is guaranteed ([14], [17], [40]). However, one modification of the basic class models has been given by Sorjamaa and Lendasse ([79]): it merges the rationales of direct strategy and recursive method, so that it computes for different models each horizon forecasting but at each time it adds previously forecasted variables.

The last strategy analysed is the MIMO one (still proposed by Bontempi in [14] and [17]), which avoids to assume conditional dependence between future values and model the dependency between the predicted values ([90], [91]).

The list of strategy provided has to be integrated with a non-exhaustive one about the model to use ([2]). According to Ahmed et al. ([2]), the best model that outperforms the others (excluded the BSTS) are the following multilayer perceptron and the Gaussian process regression:

- *Multilayer Perceptron* (MP, [13]):

$$\hat{y} = \upsilon_0 + \sum_{j=1}^{NH} \upsilon_j g(w_j{}' x^{(1)}) \tag{3.22}$$

  where $x^{(1)}$ is the input vector augmented by 1, $w$ and $\upsilon$ are the weights for the hidden and the output nodes. Even if is strongly parameters-dependent, the model has the *"universal approximation property"*, i.e. any continuous function can be represented by a neural network (given certain conditions).

- *Bayesian Neural Network* (BNN, [61], [62], [35]): it sees the network parameters as random variables, constrained to some prior distributions. The typical prior used is, for $L$ parameters (weights) and $E_w$ the sum of the squares of the parameters,

$$p(w) = \left(\frac{1-\alpha}{\pi}\right)^{\frac{L}{2}} e^{-(1-\alpha)E_w} \tag{3.23}$$

  so that the posterior, for $M$ training points and $E_D$ as sum of square errors, is given by

$$p(w \mid D, \alpha) = ce^{-\alpha E_D - (1-\alpha)E_w}. \tag{3.24}$$

- *Radial Basis Function Neural Network* (RBF, [70]): the only difference with the first method proposed is that t has a localised activation function.

- *Generalized Regression Neural Network* (GRNN, [66]): is a nonparametric model in which the forecast is provided by the average of the targets in the neighbourhood of a certain point.

- *KNN*.
- *CART*.
- *SVM* and *SVR*.
- *Gaussian Processes* (GP, [72]): it treats the responses as coming from a multivariate normal random distribution. The smoothness is guaranteed by the prior distribution, and the correlation is going to be high if the input vectors are quite close in Euclidean sense. Hence, the prediction for a covariance matrix $V(X, X)$ is given by

$$\hat{f}_* = E(f_* \mid X, y, x_*) = V(x_*, X)[V(X, X) + \sigma_n^2 I]^{-1} y. \tag{3.25}$$

- *Bayesian Structural Time Series* (BSTS, [96],[76] [78]): it is a structural model that takes into account a trend, seasonal and time series components and try to reduce the number of potential predictors through spike-and-slab method. It uses a Monte Carlo Markov Chain (MCMC) in order to simulate the posterior distribution and finally it smooths the prediction with a Bayesian averaging. In contrast with the dynamic factor model ([32], [33]), here the distribution of the regressors is not directly modelled. Thus, the model proposed can be shown in state-space form

$$\begin{cases} y_t = Z_t{}'\alpha_t + \epsilon_t & \epsilon_t \sim N(0, H_t) \quad \textit{Observation Equation} \\ \alpha_t + 1 = T_t\alpha_t + R_t\eta_t & \eta_t \sim N(0, Q_t) \quad \textit{Transition Equation} \end{cases} \tag{3.26}$$

and, in particular, the BSTS model is given by

$$\begin{aligned} y_t &= \mu_t + \tau_t + \beta' \mathbf{x}_t + \epsilon_t \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + u_t \\ \delta_t &= \delta_{t-1} + v_t \\ \tau_t &= -\textstyle\sum_{s=1}^{S-1} \tau_{t-s} + w_t. \end{aligned} \tag{3.27}$$

Once the time series is decomposed, what the BSTS method does next is using the Kalman Filter ([42]) to estimate both the trend and the seasonal components. Then, a spike-and-slab prior has been set, i.e.

$$p(\beta, \gamma, \sigma_\epsilon^2) = p(\beta_\gamma \mid \gamma, \sigma_\epsilon^2)\; p(\sigma_\epsilon^2 \mid \gamma)\; p(\gamma) \tag{3.28}$$

where $p(\gamma)$ is the spike, while the slab is given by the conditional conjugate pair

$$\beta_\gamma \mid \gamma, \sigma_\epsilon^2 \sim N(b_\gamma, \sigma_\epsilon^2(\Omega_\gamma^{-1})^{-1}) \qquad \frac{1}{\sigma_\epsilon^2} \mid \gamma \sim Ga\left(\frac{\nu}{2}, \frac{ss}{2}\right). \tag{3.29}$$

Then, a MCMC simulates the latent $\alpha, \beta, \sigma_\epsilon^2$ and the Bayesian tool averages everything.

# Chapter 4

# Forecasting with Shrinkage

Since, as we have shown before, machine learning applied to time series analysis represents a hot topic to deal with, we are now going to provide our personal approach to tackle this class of problems. In particular, machine learning usually deals with cross-sectional dataset from one hand or univariate time series from the other, but unfortunately there is not a strong literature on merging the two concepts (neither [29] or [26]), i.e. how to implement for instance forecasts for a panel dataset (except from [7], that it is going to be used extensively in this chapter).

A second issue is that in a panel data world, things blows up very quickly. Indeed, as in the application we are going to show later, it is very easy to embed in the model many more predictors than observations in our hand. Of course, no every variable included has to be important and with a strong explanatory power. Hence, our goal is to present now a generic model for panel data forecasting and try to merge it with best shrinkage technique able to identify only the variables really crucial for forecasting our outputs.

## 4.1 Panel Data Forecasting

As above mentioned, in this section we are going to show the forecasting model we will use from now on. In order to choose or build one ad hoc, we can start from the basic panel data model, according to [7]:

$$y_{it} = \alpha + X_{it}'\beta + u_{it} \tag{4.1}$$

and now we can assume different error specifications, such as

$$one - way\ component \qquad u_{it} = \mu_i + \upsilon_{it} \tag{4.2}$$

$$two - way\ component \qquad u_{it} = \mu_i + \lambda_t + \upsilon_{it} \tag{4.3}$$

In other words, the error term has an unobservable individual effect or both and individual and a time effect, beyond the classic noise, that are not included in the regression by default.[1] As a further assumption, we claim that each error component is identically and independently distributed with zero mean and sigma variance both between each other and with the $X_{it}$ matrix.

The one-way component error model can also be written as

$$\underset{NTx1}{y} = \alpha \underset{NTx1}{\iota} + \underset{NTxK}{X} \underset{Kx1}{\beta} + \underset{NTx1}{u} \tag{4.4}$$

$$= Z\delta + Z_\mu \mu + \upsilon \tag{4.5}$$

where $Z = [\iota_{NT}, X]$ and $\delta' = [\alpha', \beta']$, with $\iota_{NT}$ as vector of ones and $Z_\mu = I_N \otimes \iota_T$ as dummies matrices. From this it derives that the matrix that averages the observation across time will be $P = Z_\mu(Z_\mu{}'Z_\mu)^{-1}Z_\mu{}' = I_N \otimes J_T/T$, while $Q = I_{NT} - P$ is the deviations from the individual means.

We can now distinguish between the following two cases:

[1] There exist several functional forms in literature that have been used during the years, e.g., [75].

- *Fixed effect case*: we premultiply each term for $Q$ and then we run an OLS regression knowing that $QZ_\mu = Q\iota_{nT} = 0$

$$\tilde{\beta}_{FE} = (X'QX)^{-1}X'Qy \tag{4.6}$$

- *Random effect case*: given the variance-covariance ($\Omega$) spectral decomposition, we have that

$$\Omega^{-1/2} = \frac{1}{\sqrt{(T\sigma_\mu^2 + \sigma_v^2)}}P + \frac{1}{\sigma_v}Q \tag{4.7}$$

from which we can easily infer the best estimators of the variances

$$\begin{aligned} \hat{\sqrt{(T\sigma_\mu^2 + \sigma_v^2)}} = \hat{\sigma}_1^2 = \tfrac{u'Pu}{tr(P)} = T\textstyle\sum_{i=1}^N \tfrac{\bar{u}_{i.}^2}{N} \\ \hat{\sigma}_v^2 = \tfrac{u'Qu}{tr(Q)} = \tfrac{\sum_{i=1}^N\sum_{t=1}^T(u_{it}-\bar{u}_{i.}^2)^2}{N(T-1)}. \end{aligned} \tag{4.8}$$

We also know from [36] that the best linear unbiased predictor would be given by the GLS estimate corrected for a fraction of the GLS residuals related to each individual $i$:

$$\hat{y}_{i,T+S} = Z'_{i,T+S}\hat{\delta}_{GLS} + w'\Omega^{-1}\hat{u}_{GLS} \tag{4.9}$$

where $w = E(u_{i,T+S}\ u)$ and

$$w'\Omega^{-1} = \frac{\sigma_\mu^2}{\sigma_1^2}(l_i' \otimes \iota_T') \tag{4.10}$$

being $l_i'$ a vector with 1 in $i$th position and zero otherwise. Clearly, this predictor is the optimal one but only given the true variance components ([4], [57]).

On the other hand, for the two-way component model, we would have a variance-covariance matrix of the form

$$\Omega = E(uu') = \sigma_\mu^2(I_N \otimes J_T) + \sigma_\lambda^2(J_N \otimes I_T) + \sigma_v^2(I_N \otimes I_T) \tag{4.11}$$

so that

$$cov(u_{it}, u_{js}) = \begin{cases} \sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2 & \text{if } i = j, t = s \\ \sigma_\mu^2 & \text{if } i = j, t \neq s \\ \sigma_\lambda^2 & \text{if } i \neq j, t = s \\ 0 & \text{if } otherwise \end{cases} \tag{4.12}$$

and so, the only thing that is going to change from the one-way model is the functional form of $\Omega^{-1}$

$$\hat{y}_{i,T+S} = Z'_{i,T+S}\hat{\delta}_{GLS} + \left(\frac{T\sigma_\mu^2}{T\sigma_\mu^2 + \sigma_v^2}\right)\frac{1}{T}\sum_{t=1}^T \hat{u}_{it,GLS}. \tag{4.13}$$

This is true in case the model has a constant, otherwise

$$\begin{aligned} \hat{y}_{i,T+S} = Z'_{i,T+S}\hat{\delta}_{GLS} &+ \left(\frac{T\sigma_\mu^2}{T\sigma_\mu^2 + \sigma_v^2}\right)\frac{1}{T}\sum_{t=1}^T \hat{u}_{it,GLS} \\ &- \left(\frac{T\sigma_\mu^2}{T\sigma_\mu^2 + \sigma_v^2}\right)\frac{1}{T}\sum_{t=1}^T \hat{u}_{it,GLS} \\ &\times \left(\frac{N\sigma_\lambda^2}{T\sigma_\mu^2 + \sigma_v^2 + N\sigma_\lambda^2}\right)\frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N \hat{u}_{it,GLS}. \end{aligned}$$

For the sake of clarity and completeness, we point out that the optimal estimator could also be obtained in the case of:

- *Serial correlation*: the results can be extended assuming the erro component to be an autoregressive process or a moving average one;
- *Spatial correlation*: observation from different locations are not independent between each others;
- *Heterogeneous panel*: the coefficient varies with respect to each different individual. Usually, homogeneous estimators seems to forecast on average better than heterogeneous ones, due to their simplicity.

From now on, we are going to use our own chosen model with two-way error components, such as

$$y_{it} = \alpha + X_{it}{}'\beta_t + \mu_i + \lambda_t + \upsilon_{it} \tag{4.14}$$

where we would assume that the time-effect is following an AR (1) process, as proposed in literature ([34], [51], [52]):

$$\lambda_t = \rho_j \lambda_{t-1} + \epsilon_t \tag{4.15}$$

where $j$ represents different groups. Furthermore, we decide to set also the $\beta$ coefficient to vary over time according to an AR (1) process:

$$\beta_t = \gamma_g \beta_{t-1} + \eta_t \tag{4.16}$$

where again $g$ represents different groups.

Hence, we can rewrite our model in following state-space form:

$$\begin{pmatrix} \begin{bmatrix} y_{11}-\mu_1-\alpha \\ y_{12}-\mu_1-\alpha \\ \vdots \\ y_{1T}-\mu_1-\alpha \end{bmatrix} \\ \begin{bmatrix} y_{21}-\mu_2-\alpha \\ \vdots \\ y_{2T}-\mu_2-\alpha \end{bmatrix} \\ \vdots \\ \begin{bmatrix} y_{N1}-\mu_N-\alpha \\ \vdots \\ y_{NT}-\mu_N-\alpha \end{bmatrix} \end{pmatrix}_{NT\times 1} = \begin{pmatrix} \begin{bmatrix} x_{11} & 0 & \cdots & 0 \\ 0 & x_{12} & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & x_{1T} \end{bmatrix} & \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\ \begin{bmatrix} x_{21} & 0 & \cdots & 0 \\ 0 & x_{22} & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & x_{2T} \end{bmatrix} & \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\ \vdots & \vdots \\ \begin{bmatrix} x_{N1} & 0 & \cdots & 0 \\ 0 & x_{N2} & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & x_{NT} \end{bmatrix} & \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{pmatrix}_{NT\times 2T} \begin{pmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} \\ \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_T \end{bmatrix} \end{pmatrix}_{2T\times 1} + \begin{pmatrix} \begin{bmatrix} \upsilon_{11} \\ \upsilon_{12} \\ \vdots \\ \upsilon_{1T} \end{bmatrix} \\ \begin{bmatrix} \upsilon_{21} \\ \vdots \\ \upsilon_{2T} \end{bmatrix} \\ \begin{bmatrix} \vdots \\ \upsilon_{NT} \end{bmatrix} \end{pmatrix}_{NT\times 1} \tag{4.17}$$

This equation represents what is called *measurement* or *observation* equation, while the following one is the *state* equation:

$$\begin{pmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{T+1} \end{bmatrix} \\ \begin{bmatrix} \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_{T+1} \end{bmatrix} \end{pmatrix}_{2T\times 1} = \begin{pmatrix} \begin{bmatrix} \gamma_g & 0 & \cdots & 0 \\ 0 & \gamma_g & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & \gamma_g \end{bmatrix} & \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} \rho_j & 0 & \cdots & 0 \\ 0 & \rho_j & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \cdots & \rho_j \end{bmatrix} \end{pmatrix}_{2T\times 2T} \begin{pmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} \\ \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_T \end{bmatrix} \end{pmatrix}_{2T\times 1} + \begin{pmatrix} \begin{bmatrix} \eta_2 \\ \eta_3 \\ \vdots \\ \eta_{T+1} \end{bmatrix} \\ \begin{bmatrix} \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{T+1} \end{bmatrix} \end{pmatrix}_{2T\times 1} \tag{4.18}$$

We neglect any further specification, because we are seriously interested in only forecasting and not in

estimating the parameters or making inference.

Indeed, expressing our system through a state-space form is not only a mental and technical exercise, although it is an useful way to run a Kalman Filter ([50]) on it in order to be able forecast the value one period ahead. According to [42], [41] or [46], the filter needs a system to be expressed as follows:

$$\xi_{t+1} = F\xi_t + v_{t+1} \tag{4.19}$$

$$y_t = A'x_t + H'\xi_t + w_t \tag{4.20}$$

For what we have written above, we can see that we have no $A$ or $x_t$, while the $\xi$ vector is made by $\beta$ and $\lambda$, and the $H$ and $F$ matrices are clearly identifiable.

## 4.2 Shrinkage Method

Of course, one issue that distinctly emerges from the matrix notation is that we would have a lot of individual regressions to implement, or in other words not all the variables and coefficients have to be relevant in order to forecast our dependent variable. We should then find a way to shrink our system to let it take into account only what has an explanatory power for our outcomes. There are several ways to do that, and we are going to study and provide the most valuable alternative for this purpose.

The method we believed to be more accurate in this circumstance is the *Least Absolute Shrinkage and Selection Operator*, or shortly the LASSO

$$\min_{\beta} \ (\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.21}$$

It indeed works quite well both as a shrinkage method and in selecting the variables, but it has a several disadvantages as well. If from one hand it is clearly better than ridge regression because smalls coefficients are set to zero as fast as possible and large coefficients are in the meantime shrunk, on the other hand it does not provide standard errors at all. A more detailed overview was already proposed in section 2.4.2, but in this part of the work we are going to provide a new interpretation for the Lasso model. Indeed, according to [69], it is very easy to think about the lasso in Bayesian terms. We can then interpret the Lasso as a Bayesian posterior mode estimate if we assumer the prior on $\beta$ to be of the form[2]:

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \tag{4.22}$$

Thus, the posterior can be expressed as

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid \tilde{\boldsymbol{y}}) \propto \pi(\sigma^2)(\sigma^2)^{\frac{-(n-1)}{2}} e^{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|}. \tag{4.23}$$

The maximising $\beta$ will be a Lasso estimate, and so the posterior will. Furthermore, the Bayesian framework seems to be a pretty good alternative to standard forecasting methods (such as principal component analysis or similar) in high dimensional contexts ([27]).

In practice, since the maximisation is not properly a common Bayesian way to find the estimate, usually the mean/median is used. Hence, the functional form becomes slightly different, i.e.

$$\pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda \frac{|\beta_j|}{\sigma}} \tag{4.24}$$

with eventually a scale invariant prior for the variance $\pi(\sigma^2) = \frac{1}{\sigma^2}$.

[2] This prior form is referred in literature ([92], [44]) as *double exponential Laplace* prior.

The system could be eventually expressed through a hierarchical representation, i.e.

$$\begin{aligned}
\boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \\
\boldsymbol{\beta} \mid \tau_1^2, \tau_2^2, ..., \tau_p^2, \sigma^2 &\sim N_p(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau), \quad \boldsymbol{D}_\tau = diag(\tau_1^2, \tau_2^2, ..., \tau_p^2) \\
\tau_1^2, \tau_2^2, ..., \tau_p^2 &\sim \textstyle\prod_{j=1}^{p} \frac{\lambda^2}{2} e^{\frac{-\lambda^2 \tau_j^2}{2}} d\tau_j^2, \qquad \tau_1^2, \tau_2^2, ..., \tau_p^2 > 0 \\
\sigma^2 &\sim \pi(\sigma^2) d\sigma^2.
\end{aligned} \tag{4.25}$$

# Chapter 5

# First Application: Insurance Big Data Analytics

## Advantageous Selection in European Insurance Markets

### Abstract[1]

Rothschild and Stiglitz ([30]) argued that people signal their risk profile through their insurance demand, i.e. individuals with a high risk profile would buy insurance as much as they can, while people who are not going to buy any insurance are the ones with a lower risk profile. This issue is commonly known as adverse selection. Even if their prediction seems to work quite well in a lot of different markets, Cutler et al. ([13]) proved that there exist some insurance markets in United States in which the expected result is completely different. In the wake of this study, we provide empirical evidences that there are some European insurance markets in which the low risk profile agents are the ones who buy more insurance.

### 5.0.1 Introduction and literature review

The insurance market is usually the most common example used in textbooks trying to explain the impact of the information on any economic activity. Indeed, the model proposed ([30]) is usually quite straightforward: an insurance company should be suspicious concerning people who want to buy some coverage because only individuals with a high expected claims are willing to pay a premium for being compensated in case an accident occurs. Therefore, asking for an insurance is thus a signal that a person will need to be reimbursed at some point in future. Since the insurance company makes profit on the probability that not every client will need to be paid more than the premium deposited, it is also not going to sell any insurance if it is certain that every client will need to be paid in the contract lifetime. On the other hand, people that are not expected to have a high claim in future are not willing to pay any premium for being insured. This is an asymmetric informational issue called in literature *adverse selection* ([1]).

Hence, the insight behind this concept is that the correlation between the individual's demand for insurance and the risk of losses has to be positive. In the health sector, many works tested this positive correlation idea, such as Mitchell et al. ([29]) for the American annuities market, while McCarthy and Mitchell ([28]) focused on the Japanese annuities market and Finkelstein and Poterba on the English one in different works ([23], [24], [25]). A more extensive review of the verification of the positive correlation between insurance coverage and risk occurrence can be found in Cutler and Zeckhauser ([12]).

The framework has also been extended in several different ways, but the prediction is again confirmed,

[1] This paper uses data from SHARE Waves 1 and 2 release 2.6.0, as of November 29th 2013 (DOI: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N. 211909, SHARE-LEAP, N. 227822 and SHARE M4, N. 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged.

as for instance proved in Chiappori and Salanie ([9]) and Chiappori et al. ([10]).

On the other hand, even if the classic and intuitive adverse selection hypothesis has been validated and proved to be robust in many circumstances, some influential exceptions exist. Indeed, Einav et al. ([17]), Einav and Finkelstein ([18]), Cardon and Hendel ([7]), as well as Cutler et al. ([13]), and Finkelstein and McGarry ([21]) among all, showed that the prediction of positive correlation fails in some countries and markets, even in sectors other than health ([15], [11]). In particular, Medigap insurance demand seems to be negative correlated with the risk occurrence ([19], [20], [27]), as well as life insurance ([8]) and long-term care ([22]). This seems to be due to a wider spectrum of private information owned by the individuals that would entail a preference heterogeneity and an unexpected irrational coverage. The result of these analysis has been named *advantageous* or *propitious* selection ([14], [26]).

Some explanations are identified in the variable individual's risk tolerance. In fact, preference heterogeneity for both risk tolerance and risk type may let the sign between insurance demand and accident occurrence to be anyone ([16]), since I) individuals with lower (higher) risk tolerance can either buy more (less) insurance or invest in instruments or activities that lower (higher) the expected claims; and II) individual's behaviour may vary across different markets, i.e. the correlation may be positive for some markets and negative for others.

On the wave of the works above mentioned, the purpose of our analysis was to verify (or disprove) whether the correlation between insurance coverage and risk occurrence was indeed positive, or on the other hand negative or absent within European countries. Five different health insurance markets have been considered: term life, annuity, long-term care, acute health and eventually Medicare supplemental insurance (Medigap), as already proposed in Cutler et al. ([13]). The demand of each one of this insurance type has been studied with respect to both risky behaviours (i.e., behaviours suitable to proxy risk tolerance) and risk occurrence (i.e., the event that should trigger the payment from the insurance company).

### 5.0.2 Data and empirical framework

As already mentioned in Section 5.0.1, the purpose of the analysis is to see what kind of relationship exists between five insurance market demands, risk tolerance and risk occurrence within different European countries. The analysis implemented used micro panel data on health from the Survey of Health, Ageing and Retirement in Europe (SHARE) project. We used a sample of people aged more than 51 in 2004-2005, for eleven European countries (Austria, Germany, Sweden, Netherlands, Spain, Italy, France, Denmark, Greece, Switzerland, Belgium) plus Israel.[2] The Appendix presents key summary statistics for each country. Figure 8.1 - 8.2 show the average age to the population and the average medigap expenses during the period considered. As it can be seen, the average age is pretty stable across countries, with the highest pick corresponding to Spain, followed then by Austria, Sweden and Switzerland. On the other hand, Sweden is the country in which people spend the most in additional medicines and/or cure, i.e. where the people buy a supplementary insurance more likely. Another Scandinavian country, the Denmark, is ranked second, followed directly by Israel and Italy. If we instead have a look to the Figure 8.3, we can observe that for almost each country the population is on average slightly overweighted. There are more obese than underweighted, and these two measures seem to be at a glance inversely correlated. The Figure 8.4 claims instead that, on the total population considered, only a small amount of persons undertake preventive health actions, and this happens in particular in Germany, Greece and Spain (Italy just following). Within the group of persons who take actions of the kind described above, it is very common to do the minimum possible, i.e. undertaking only one preventive measure (this is particularly true in Greece and in Switzerland, Germany and Austria). There is a consistent amount of people who go further and implement a second preventive action as well, but above that threshold the number shrinks toward very low levels (Greece is emblematic from this point of view, since it has the highest percentage of people undertaking one single action and the lowest of who undertakes more than two preventive actions). The best examples here are The Netherlands, Spain and Belgium. Finally, Figure 8.5 exhibits a wider spectrum of variable summary statistics, expressed in percentage terms, for the groups of insurance coverages, the remaining risky behaviour and risk occurrence variables, and finally for the controls as well. Instead of focusing on a single variable, what we infer from this last figure is the high heterogeneity within the population. Already since this figure, we observe how this sparsity may be reflected in heterogeneous preferences, a fundamental concept which may help us in enlighten the

[2] The panel nature of the dataset was essential, for instance, to track mortality and nursing home.

advantageous selection phenomenon.

From the SHARE survey, we indeed extracted several answers to construct the variables used in our regressions. In particular, as insurance and risk occurrence, we measured:

- Life insurance as whether the individual has a term life insurance at the time of the survey (or both term and whole life policies), and the correspondent occurrence is whether the individual dies between 2004 and 2006/7. According to Cutler et al. ([13]) we use the term life insurance since it represents a pure investment compared to a whole life insurance, where we should take care also about the saving component;

- Acute health as whether the individual has a hospital care with unrestricted choice of hospitals/clinics and/or hospital care with limited choice of hospitals and clinics. The risk occurrence is whether the individual has been in a hospital in the last twelve months;

- Annuity as whether the individual has a personal and private annuity insurance, with the corresponding risk occurrence of whether the interviewed is alive at the time of the second survey (2006/7);

- Medigap as whether the individual has a supplementary insurance.[3] The risk occurrence here is the amount the individual incurred as extra medical expenses;[4]

- Long-term care as whether the individual has at the time of the survey a long term care in nursing home insurance and/or a nursing care at home in case of chronic disease or disability. The corresponding risk occurrence is whether the individual has been into a nursing home between 2004 and 2006/7.

Instead, as proxy for risk tolerance, we decided to use the following measures able to capture the risk preferences:

- Smoking, i.e. whether the individual currently smokes;

- Drinking problems, that is whether the individual drinks two or more glasses of alcohol each day or 5/6 days a week;

- Body mass index (BMI), considered as an indicator of incorrect actions about individual's diet, is computed as individual's weight divided the square of the height, times 10,000. In this way, it has been possible to classify the individual under the following four categories: Underweight (BMI below 18.5), Normal (18.5 - 24.9), Overweight (25 - 29.9) and Obese (30 or higher). Finally we assigned 0 to the variable if the weight was in the normality range, 1 otherwise;

- Level of physical inactivity, defined as never or almost never engaging in neither moderate nor vigorous physical activity;

- A variable reflecting preventive health actions followed out by the interviewed.[5]

[3] In particular, a person has a supplementary insurance if he has at least one of the following: Medical care with direct access to specialists; Medical care with an extended choice of doctors; Dental care; A larger choice of drugs and/or full drugs expenses (no participation); An extended choice of hospitals and clinics for hospital care; (Extended) Long term care in a nursing home; (Extended) Nursing care at home in case of chronic disease or disability; (Extended) Home help for activities of daily living (household, etc.); Full coverage of costs for doctor visits (no participation); Full coverage of costs for hospital care (no participation).

[4] It has been computed as the total sum in euros of paid out-of-pocket expenses for inpatient care, paid out-of-pocket expenses for outpatient care, paid out-of-pocket expenses for prescribed drugs and paid out-of-pocket expenses for day care, nursing home and home-based care.

[5] This variable has been constructed as an indicator of whether the individual has consulted a specialist for regular controls, whether he had a flu vaccination in the last year, whether he had a sigmoidoscopy or colonoscopy less than 10 years ago, whether he had a mammogram (x-ray of the breast) and if he had another test to detect hidden blood in his stool in the last 10 years. From each action undertaken, he got one point and the final indicator is expressed as the sum of all the point obtained, i.e. if an individual has the preventive variable equal to two it means that he did only two preventive actions out of five.

Therefore, we run the following two different regressions:

$$Pr(Y_i|PRT_i) = \alpha_0 + \alpha_1 * PRT_i + X_i\Gamma + \epsilon_i \tag{5.1}$$

$$Z_i = \beta_0 + \beta_1 * PRT_i + X_i\Pi + \eta_j \tag{5.2}$$

where $Y_i$ represents the fact that an individual has or not the particular kind of insurance under analysis, *PRT* stands for *Proxy of Risk Tolerance*, that is the behavioural variables discussed above, while $Z_i$ is the risk occurrence for the insurance studied, and $X_i$ are the covariates (gender, age, education and marital status).[6] We then run both the unconditional regression and the one controlled for the covariates. The control variables are used according to the usual insurance practices and are applied differently with respect to the insurance markets. Indeed, about the term life/long term insurance we will control for education, age and gender; then we will check the Medigap for education and age, the annuity for age, gender, education and marital status and the acute health only for education.[7] We decided after careful consideration to use the probit in the model 5.1 because, although does not differ almost at all from a standard least squares regression model, it provides a better probabilistic interpretation. The model 5.2 is instead a classic least square estimation.

Since we should also embed somehow the differences due to being analysing different countries, we decide to follow the Bryan and Jenkins' approach ([6]) on hierarchical (multilevel) datasets. According to them, to prove the robustness of our analysis we are going to run a simple pooled regression, a separate regression for each country and a country fixed effect model. This multiple choice could prove the results to be not related to the technique used and will improve the understanding of the phenomenon we are trying to capture providing different interpretations of the data.

First of all, a pooled regression with clustered-robust errors is going to be run. This would ignore that different countries have different unobserved features and will underestimate the standard errors of $\beta$, but it could be easily corrected using countries-robust standard errors that allow for a more general correlation within countries.

The second analysis implemented concerns instead a separate regression for each country. The country effect is in this way internalised and it is merged with the intercept of each regression model. It is a bit computationally more demanding, but it allows to put no restrictions on the variances of country-specific errors and to let $\beta$ to vary across countries.

The final approach used is the fixed effect estimation, and it is set as a middle way between the two models explained above. It indeed pooled all the data but allows the intercept to differ across countries to be able to capture individual-specific effects. The other greatest difference with the single-country regression is that the residuals are here constrained to be the same across countries. Besides, it is useless to include further country-level variables, since the intercept embeds country differences. Every regression will then be corrected for cluster-robust errors and cross-sectionally weighted by the weights system provided by SHARE.[8]

### 5.0.3 Results

The first two regressions presented in the Appendix are the pooled regressions. At a first glance, it seems that at an aggregate level the effects are not so weak, although very sparse. Indeed, even if some of the results are generally either not significant or confirming the classic adverse selection theory, some relationships between insurance coverage and risky behaviours proved to be robust, meaningful and able to confirm our initial hypothesis of advantageous selection in European markets. Furthermore, the

[6] The education variabile has been set as a binary variable on whether the individual has pursued or not a higher level of studies, such as university, college, nursery school, etc. In addition, the marital status variable has been created as well as a binary variable, on whether the individual is married/in a registered partnership or not married/divorced/widowed.

[7] For a more detailed definition of risk classification controls, see Cutler et al. ([13]).

[8] For a more detailed explanation of the weights system, look at SHARE release guide for wave 1, pag. 39-46.

control variables seem to not affect considerably the estimation results. For instance, according to the classic theory individuals who currently smoke or drink should buy more insurance, but in reality they are more likely to buy less insurance. This is particularly true for long term care and term life/acute health respectively for smoking and drinking, and the same it is also verified for annuity markets and long term care for people physically inactive and for who implements more preventive health care actions. In addition, people not in the normality weight range are actually going to buy few insurance in three different markets, i.e. annuities, medigap and acute health.

In addition, the Table 8.2 shows that both smoking and physical inactivity increase the likely to die (and to not live long). While drinking seems to not be statistically significant in any circumstances, physical inactivity will also involve a higher level of medigap expenses as well as a higher likely to be hospitalised, as expected. On the other hand, preventive health actions reduce this risk and the smoking does not increase the chance to get hospitalised. This may seem counterintuitive, but since we considered a short time hospitalisation period and since the smoking effect are quite long term, it may be reasonable that the two variables are not positively correlated. Surprisingly, some anomalies characterise the BMI variable, meaning that the BMI seems to not reduce the life expectation. Further studies may be necessary in order to understand the reason why these kind of anomalies happen, but in general we may think of some psychological disease, misperception of the illness or simply the stress as possible causes of those strange phenomena, since it seems reasonable that people who, for example, are hypochondriac (or that somatizing a lot) are the ones who implement more prevention, who then spend more in extra medicines and cure and the ones who go to the hospital more likely as well. One general interpretation of the deviations presented is that maybe more risk averse individuals have less risky behaviours, and are the ones who value the insurance the most.

As above mentioned, the results are not verified for all the insurance markets and with respect to each dependent variable, but already in the comprehensive overall regression they provide robust insights about the advantageous selection issue.

After that, we run instead the Linear Probability Model analysis at a country-level. A regression for each country has been run and the results are visualised in Appendix as Figures 8.6 - 8.9. There are five subgraphs corresponding to each insurance market and each coefficient for every independent variable is drawn by a smaller circle and a line that represents the confidence interval for that coefficient estimates at a level of 95%. For the sake of completeness, even if the results are not extremely different, the following figures have also the coefficient estimates taking into account the control variables. The results are clearly not so distant from the ones observed at an aggregate level, but they are again really mixed within each country and insurance market. What it should be noticed from these graphs are the numbers of point under/above the zero line, since as before we are more interested in the sign of the relations more than in the magnitude. In particular for the term life, the annuity and the medigap insurance markets, having riskier behaviours or taking less care about own health does not directly entail a higher demand of insurance. Again, the relation between the risk occurrence and the risky behaviours is instead generally confirmed, in particular regarding physical inactivity or the smoking addiction.

The final regressions showed in the Appendix regards the country fixed effect model (with cluster-robust errors), that is usually used in this situation because, with respect to for instance a random effect model, it underlines the unique features of each country. In the regressions run here, the control variables looked still to not have a crucial role.

The Table 8.3 points out again that, as expected, people who smoke or drink/with weight problems, are more likely to buy a term life or a long term care insurance, respectively. The opposite is instead verified still for smoking, drinking and BMI with respect to long term care, term life and acute health markets. The prevention is still ambiguous, since if from one hand shows an expected result such as the negative correlation with the annuity insurance purchase, on the other hand involves a positive relation with the acute health market, that is to some extent counterintuitive. Finally, physical inactivity proved again to provide the most robust results, i.e. it is negatively correlated with annuities, long term care and medigap as well. All our consideration may still make sense, behaviourally speaking, if we think again about people affected by apprehension or hypochondria, or physical inactivity reflected also in disregarding for personal care.

On the risk occurrence side instead, smoking is as expected associated to a higher chance to die (and

to not live long), as well as physical inactivity, that proved also to be positively correlated with medigap expenses and hospitalisation. Prevention may require, as above mentioned, a higher possibility to get hospitalised, while counterintuitively the BMI is positively correlated with a higher life expectation and the smokers are less likely to go to the hospital (in one year time).

Even in the country-fixed effect framework, although the results are less strong than in the pooled regression case, some anomaly seems to persist, and we believe the reasons behind this deviation could be interesting to be investigated in future works. We cannot conclude univocally in favour of our initial hypothesis neither in the fixed-effect scenario, but we can claim that the standard adverse selection theory seems to not hold strongly as the theory stated.

### 5.0.4 Conclusions

Our analysis aimed to investigate whether an advantageous selection phenomenon was proved to be robust in different insurance markets, as in Cutler et. al ([13]). We focused on five insurance markets for eleven European countries plus Israel, specifically on term life, annuity, long term care, Medigap and acute health insurances. Our main finding has been that it looks like that riskier behaviours are not always associated with higher mortality, but above all they are not unconditionally associated with higher insurance demand as the classic theory would predict. This result does not hold for each country and each market with respect to each risky behaviour, but the outcomes are mixed, suggesting that further analysis may shine a light on this puzzle. In particular, in the most robust analysis, no systematic relation between risky behaviours and any of the insurance market, although some risky behaviours are not coherent (while others are) with Rothschild and Stiglitz ([30]). In any case, it is interesting to notice that the adverse selection proposed in the '70s does not hold anymore so strongly and extensively, but also to consider that maybe preferences heterogeneity for insurance could explain the different behaviours of the participants. A different risk tolerance may indeed explain the insurance puzzle, but of course further investigations will be required in order to test this hypothesis.

# Chapter 6

# Second Application: Behavioral Economics

## The Effect of Entrepreneurs' Behavioral Biases on the Choice of Insuring Their Companies


### Abstract

We study the effect of behavioural biases on entrepreneurial choices to insure their firms against kinds of corporate risks. We use a large sample of Italian Small and Medium sized - finding that they under-insure themselves. The dataset allows us to link corporate insurance choices with the personal traits of the entrepreneur and his household's financial choices.


### 6.0.1 Introduction

We study the effect of entrepreneurs' behavioral biases on the choice of under-insuring their companies against different kind of risks.
Following traditional financial theory, bigger firms should purchase less insurance compared to smaller ones, because they may self-insure themselves diversifying their businesses. In addition, shareholders of big companies should be less willing to pay for insurance because they can typically hedge risk by investing in a diversified portfolio.
Instead, empirical evidence shows exactly the opposite: while bigger companies do buy insurance - sometimes even resulting over-insured, thus imposing an undesired cost to their shareholders - smaller ones tend to be under-insured. Typically, the form value constitutes a large portion of the owner's wealth, thereby exposing him to a large, uninsured risk.
Corporate finance theory and some empirical studies show that underinsurance lead firms to invest less than optimally, thereby foregoing profit opportunities and ending up with a lower return on equity. Moreover, underinsurance reduces their chances to get credit from banks (Guiso and Schivardi, 2010a; 2010b), with a negative impact on their growth opportunities. This behaviour appears irrational especially if we consider that smaller firms typically get credit at worst conditions compared to bigger ones (Hubbard, 1998). Therefore, they should consider insurance as a way of obtaining better conditions from banks, as an alternative, for example, to collaterals. We claim that this irrational choice of under-insuring their companies is due, at least in part, to entrepreneurs' behavioural biases. In particular, we find a significant relationship with overconfidence, the illusion of control, and over-optimism.
We analyse a detailed survey conducted in 2009-10 on a sample of 2,295 Italian Small and Medium Enterprises (SMEs). The survey is composed by questionnaire addressed to the owner or the person in charge of taking decisions about insurance, who was also interviewed directly. The survey includes detailed information about the types of insurance contracts related to different kind of corporate risks, information on damages suffered in the past, and on the entrepreneurs' personal and household characteristics.
The database combines both entrepreneurs' personal information as well as data on their companies. This unique feature is important to link their insurance-related decisions both at a personal and corporate

level, but also to discern the entrepreneurs' choices on how much to invest in their company with respect to their total wealth, and the degree to which they underestimate the riskiness associated with their business. Because, on average, Italian entrepreneurs invest about 40% (Guiso, 2010b) of their household wealth in their company, under insuring it leads them to bear too much risk. Eventually, this may be transferred to their household, and affect their wealth.
As a matter of fact, SMEs are exposed to several types of risk, but the companies in our sample on average insure only three out of ten types of distinct risks. Of course, the choice on how much insurance to buy depends on the entrepreneurs' risk aversion, but also on their perception of the risk of suffering a loss, and the probability of provoking it to others. The capacity of bearing the regret associated with a loss also seems to play an important role in these kinds of decisions. Among others aspects, trust in the insurance companies plays a major role in the entrepreneurs' decision to get insurance.
We use the detailed information in our database to account for these biases, because part of the survey has been conducted through direct interviews. More in details, we own data on risk attitude, ambiguity aversion, overconfidence, over optimism, regret aversion, trust in insurance companies, banks and, in general, in other institutions or people. Furthermore, we have information on the entrepreneurs' psychological traits and demographic characteristics, but also with regard to their families, household wealth, degree of diversification, personal insurance contracts, and etc. This unique dataset allows us to link information on companies, entrepreneurs and their families, analysing the relationships between risks borne by these three distinct entities, but also of potential spillover effects between them.

### 6.0.2 Literature Review

Most of the theoretical literature has focused on the role of insurance in mitigating the principal-agent problems arising when managers do not fully own the company. Therefore, this literature is not completely relevant for our analysis of SMEs where the owner (and in some cases still the founder) has a large or full control of her firm. Yet, some insights, especially those related to the relationship between insurance cover and leverage, are pertinent.
Turning to the empirical studies, the breadth of the analysis is somehow limited by the availability and quality of data on corporate insurance purchases. Still, they can shed some light on the empirical evidence related to the theoretical results. In what follows, we review the key theoretical insights and discuss their relevance for our analysis. Then, we turn to the empirical literature, summarizing the main results.[1]

#### Theoretical studies

Mayers and Smith (1982) claimed *"The corporate form provides an effective hedge since stockholders can eliminate insurable risk through diversification"*.
Most of the theoretical literature studying the incentives corporations has for buying insurance aims at analyzing in which cases the above claim is true.
This amounts to say that, from the point of view of an investor holding a diversified portfolio, the value of an insured corporation is the same of an insured one, and therefore purchasing insurance is not necessary as a risk management tool. Such a result is established also by Mayers (1982) and MacMinn (1987) and holds in a model with stocks, insurance and risky debt, where default costs are nil.

However, the introduction of a conflict of interests between managers - acting in the interest of stockholders - and bondholders dramatically alters the results. Two main agency problems arise, which require the purchase of insurance: underinvestment and asset substitution (or risk shifting).
Underinvestment originates when the manager of a firm has no interest in undertaking investments above a certain threshold, as mostly bondholders will enjoy the additional returns. As a consequence, the corporation may forego positive net present value projects if their profits accrue just to bondholders (Jensen and Smith, 1985). Purchasing insurance can alleviate this problem (Mayers and Smith, 1987; Garven and Mac Minn, 1993). In a nutshell, given a positive probability of insolvency, the optimal investments schedule for a leveraged firm is non-decreasing with respect to insurance coverage. Insurance reduces the probability of insolvency due to non-market risks. Therefore, it protects (at least partially) stockholders from the extra risk involved in additional investment problems and bondholders. The conflict between bondholders and managers also emerges when the firm has to choose mutually exclusive projects, and

[1] MacMinn and Garven (2013) provide a detailed survey of the most recent results. See also Gollier (2010).

between sources of financing. If a firm substitutes high-risk projects with low-risk ones, the value of equity increases at the expenses of that one of loans (MacMinn, 1993; Jensen and Smith, 1985), and thereby value shifts from bondholders to shareholders. As far as the choice between mutually exclusive projects is concerned, purchasing insurance increases the value of the safest one, and therefore managers acting on behalf of shareholders will prefer it (MacMinn and Garven, 2014). Moreover, if the purchase of insurance in the financing decision is made before deciding on the scale of productions and the investment choices, even an insurance purchase with zero risk-adjusted net present value would increase current shareholders value (MacMinn and Garven, 2014). The purchase of insurance is also related to the firms' preference to use internal funds to finance projects (the so-called "pecking-order hypothesis", Myers and Majiluf, 1984). An uninsured adverse shock cuts into liquidity buffers, reducing the number of projects that can be financed and therefore the overall value of the firm (Froot, Scharfestein and Stein, 1993). Thus, insurance helps preserve internal funding. The compensation scheme adopted for managers also influences insurance choices. Han and MacMinn (2006) show that a manager paid in stock options and using low-risk debt to finance investments has an incentive to buy a cover. They show that insurance increases the value of stock options by transferring money between states of nature.

Taxation offers another incentive to buy insurance: normally earnings up to a certain amount are not taxed and some expenses are deductible from taxable income. This implies that, with a proportional tax rate, after-tax earnings are a concave function of gross earnings, making firms averse to shock to total earnings, even though they could be diversified by portfolio choices (Eeckhoudt, Gollier et Schlesinger, 1997). If the insurance premium is fully tax deductible, while accident-related losses are not, buying a fairly priced insurance reduces the expected tax payment.

Finally, industrial firms are better equipped at managing risks coming from their core business (e.g., related to the launch of a product, or the control of costs, etc.) and prefer to delegate to insurers or financial intermediaries the other risks for which they do not have a comparative advantage (Doherty, 2000). Thus, insurance can be thought as a tool for externalizing some functions, especially in smaller companies.

**Empirical studies**

The above-mentioned theoretical studies underline the importance of the financing structure of the firm and the tension between managers and shareholders as key drivers of the decision of purchasing insurance. This tension is at the heart of the empirical studies of the corporate demand for insurance, whose number is however severely limited by the lack of data. Firms are not legally required to declare their expenditures on insurance and, moreover, information on the premium paid is a rather crude proxy for coverage and ideally should be complemented by further information on the contracts, i.e., deductibles and limits to coverage, which are difficult to obtain.

The seminal paper by Hoyt and Khang (1999) - that inspired most of the recent contributions - uses a large sample of Chinese firms, assessing what drives the decision of how much coverage to purchase against risks to property, proxied by the ratio between premiums paid and the value of firm's insured assets. Their findings corroborate many conclusions of the theoretical literature. First of all, firms with more debt as a ratio to equity and with higher growth opportunities have, other things being equal, purchase more property coverage, consistent with the underinvestment hypothesis. Moreover, the share of company owned by managers has a negative correlation with insurance, reflecting the role of insurance in aligning incentives, although the effect is higher for larger firms. Larger firms tend to buy less insurance, in line with the real services, comparative advantage hypothesis, as well as those with higher tax shield (share of tax credits and carry forward losses to assets), consistent with the tax incentive for insurance purchase.

Aunon-Nerin and Ehling (2008) use very detailed data on over 10,000 insurance contracts written by US corporations and analyse, using a simultaneous equation model, the choice of deductible and limit coverage (i.e., the ceiling for compensation). Accounting for the possible endogeneity of the financing structure, they find that the deductibles and limits have different drivers and size has a negative impact on the limit, but no impact on the deductible. The results on leverage are in line with the underinvestment hypothesis: the share of long-term debt is positively correlated with limits and negatively with deductibles, moreover, its interaction with size is positive and statistically significant, indicating that the

bankruptcy costs that insurance helps mitigate are proportionally higher for smaller firms. Moreover, insurance cover is negatively related with the pay-out ratio, as cash in excess of investment needs helps self-insurance.

A simultaneous equation model is also employed by Zou and Adams (2008) to model insurance purchases, debt capacity and the cost of debt, under the implicit assumption that the amount of insurance coverage and the financing structure are determined simultaneously. This holds true if the banks providing credit and bondholders are informed of the insurance purchase. They analyse a sample of Chinese listed companies, from 1997 to 2003, finding a negative relationship between debt and insurance purchase, which they rationalize with the implicit bailout distressed firms obtain by the State. However, a higher cost of debt leads to more insurance coverage. The usual negative relationship with size is found as well. Turning to the effect of insurance, they show that higher cover helps to expand leverage and reducing debt costs.
All the studies surveyed below, instead, focus on the amount of insurance purchased, implicitly ruling out uninsured firms.
Zou and Adams (2006) consider both the choice of whether buy insurance and the degree of cover and in this sense is more akin to our paper. They use data on Chinese corporation spanning the period 1997-1999, and their results are mostly in line with those of the theoretical literature. The decision to purchase insurance is positively related to the incidence of physical assets on total assets and the leverage. Moreover, for a given level of managerial ownership (which in itself is not statistically significant), the propensity to buy insurance increases with leverage. This result runs against the management alignment hypothesis, but it is consistent with managers of leveraged firms being concerned with the security of their job and the value of their stock options. The probability of being insured is negatively correlated to the tax rate and the extent of tax-loss carry forwards. Additionally, the amount of purchase cover (defined as the ratio between premiums and insurable assets) is, conditional to the decision to get covered, positively related to growth opportunities, corroborating the underinvestment hypothesis, but negatively, and somehow counter-intuitively, to the intensity of physical assets. The extent of management shareholding is again positively correlated with the size of insurance cover, while the expected negative relationship with size is statistically significant.

### 6.0.3 Dataset Description

We use a dataset obtained through the survey run in 2008-2009 by the Italian National Association of Insurance Firms (aka, ANIA). The database is composed of face to face interviews and a questionnaire covering 2,295 Italian companies. The answers are matched with individual balance sheet data.[2]
The survey was addressed to the person in charge of taking insurance-related decisions, often the entrepreneur herself.[3] First, it was asked to fill in a questionnaire containing data on the insurance coverage and other information related to the company. Subsequently, entrepreneurs' personal information has been collected through face-to-face interviews. Even though the initial survey was meant to address all the Italian SMEs (i.e., firms with less than 250 employees), because the survey was conducted in the middle of the financial crisis, it was not possible to interview the entrepreneurs of the largest companies, given their intense commitments in that period. Hence, even though the dataset does not entirely capture the Italian SMEs landscape - and it is thus biased towards smallest companies - it provides extremely valuable information on small firms and their entrepreneurs.

**Companies characteristics**
The figures that follow refer to 2007, the last year before the survey for which we have balance sheet data. Figure 6.1 shows the distribution based on the firms' age.

The left-skewed distribution clearly indicates that few very old companies took part to the survey (71- 99 years, 3.1%; 100-199 years, 1.8%; $>$ 200 years, 0.1%), while the average age of the companies

[2] They are supplied by CERVED, the largest Italian information and rating provider.
[3] Other times the person in charge of taking insurance related decisions was the CEO of the firm or a director. See below for details. However, for the sake of simplicity, and since many times the founder is also the CEO or has a managerial position, from now on we will refer to this person as the "entrepreneur".

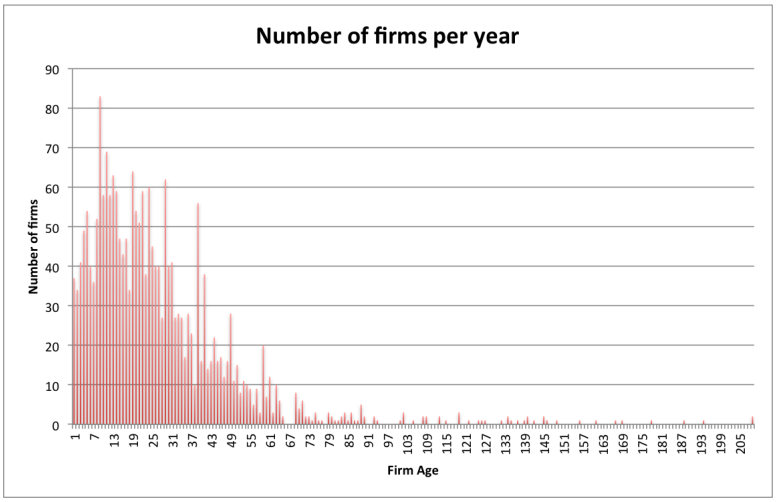



Figure 6.1: Number of firms, per year of incorporation.

is around 34 years (median 29 years). Figure 6.2 shows the distribution of the companies in terms of revenues, underlining the small dimension of firms analysed, i.e., between 1 and 10 million euros.

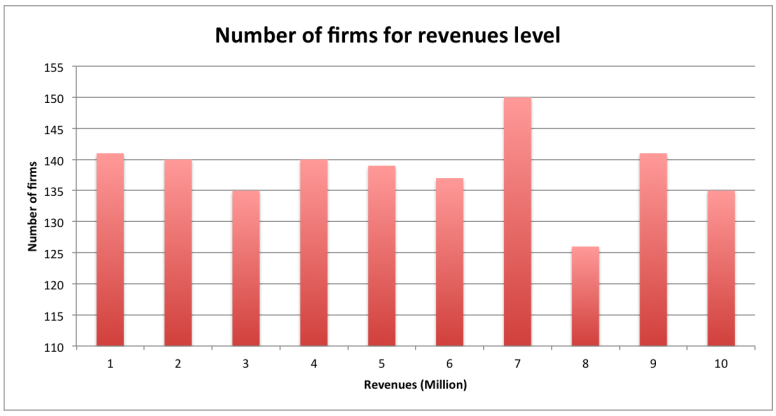



Figure 6.2: Number of firms per revenues achieved.

Figure 6.3 displays the distribution with respect to the number of employees, which conveys another dimensional measure of the population considered for the study.

Even though on average those companies can count on 31 employees, most of them have no more than 20 people working at the company (both mode and median are equal to 20). In addition, we categorize the companies with respect to industry (Figure 6.4) and business name (Figure 6.5).

As Figure 6.4 shows, the vast majority of firms belong either to manufacturing industry, hospitality, or other services. The industry breakdown is interesting because sectors that are usually more prone to buy insurance because of intrinsic rationales (e.g., manufacturing or mining) are the smallest part of our sample. Thus, we claim that the entrepreneurs in our sample have a major role in deciding either to insure their firms or not and what kind of risks to insure, not being forced by the peculiarities of their sector. Of course, we will consider potential industry effects in our empirical analysis that we present in section 6.0.4.

Without digging too much in legal details - that are not essential for the sake of this study - it is important to notice that the majority of the companies in our sample have limited liability structures (i.e., Spa and Srl), as shown in Figure 5. However, about 14% of the firms have an unlimited liability (i.e., Snc). Therefore, while in general SMEs should insure themselves more than bigger companies, we expect entrepreneurs in the subsample of firms with unlimited liability to properly insure their companies to avoid that a loss at firm level may impair their private or household wealth.

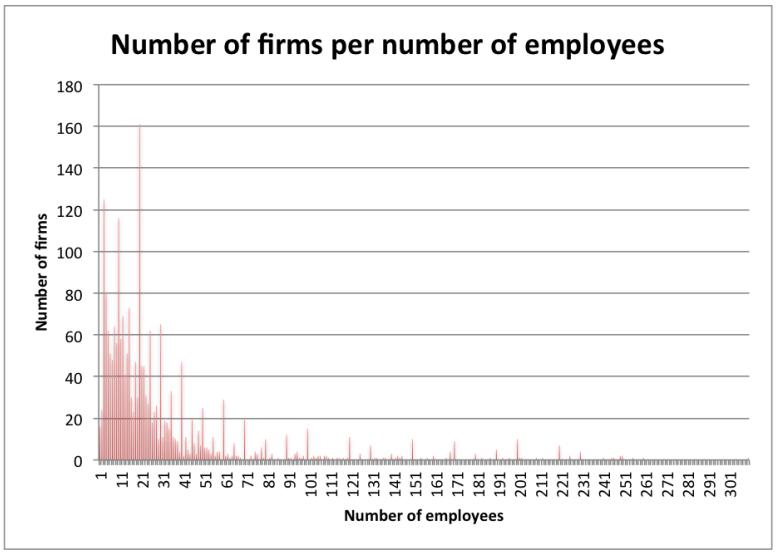



Figure 6.3: Number of firms, per number of employees.

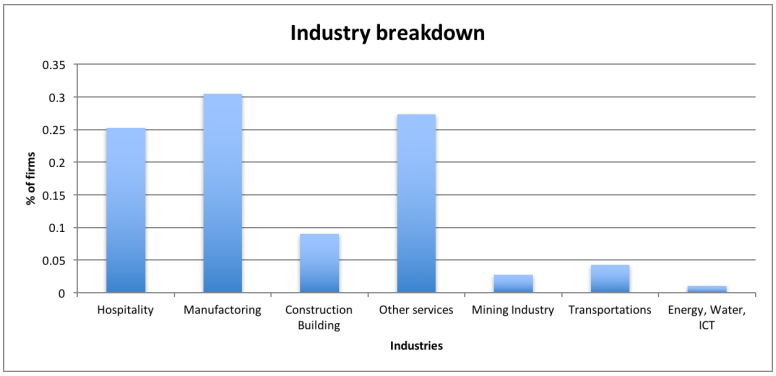



Figure 6.4: Industry breakdown.

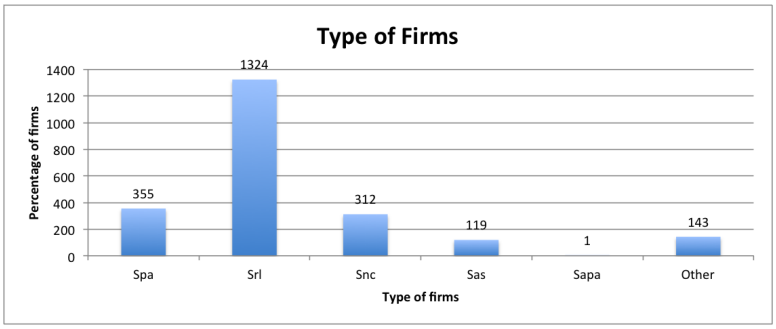



Figure 6.5: Companies classified by business name.

Figure 6.6 classifies the respondents to the survey, a very important information to understand how personal characteristics may affect insurance-related decisions.

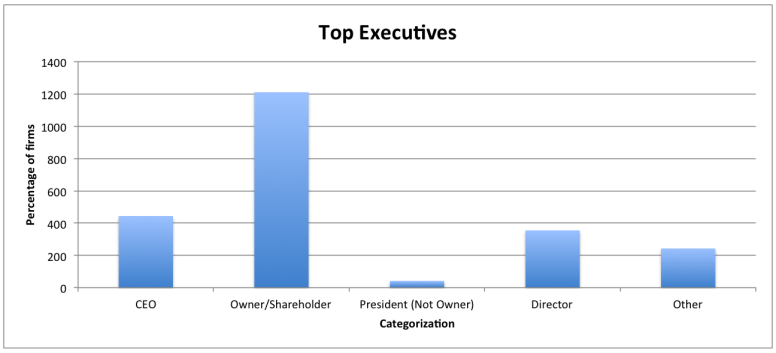



Figure 6.6: Executive who answered the questionnaire.

Entrepreneurs (owners or major shareholders of the firm) represent more than 50% of the respondents to the survey (i.e., the person in charge of insurance-related decisions), followed by CEOs (almost 20%), and Director (around 15%).

**Data on Insurance Policies underwritten**

The survey considers 11 different types of risk against which the firm can buy insurance: Fire; Technological risk; Theft; Goods transported; Credit risk; Foreign investments and exports; Business interruption; Third parties' damages; Product liabilities; Environmental risk; and employees' insurance. Figure 6.7 shows the distribution of firms with respect to the number of insurance policies underwritten.

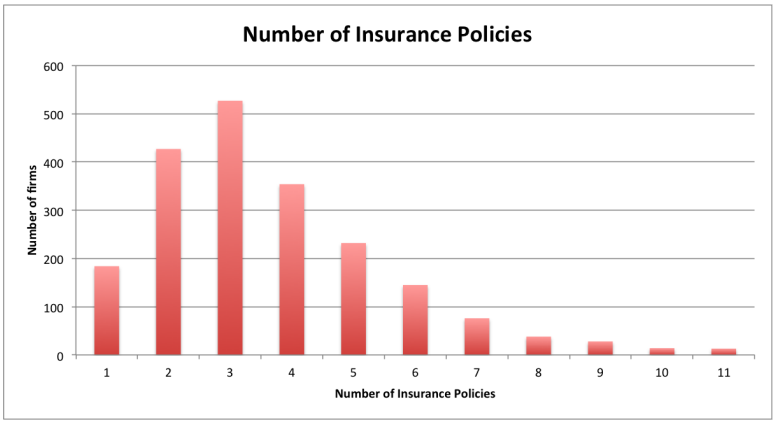



Figure 6.7: Breakdown of companies, by number of insurance policies underwritten.

The average number of insurance policies underwritten is three. Out of 11 possible risks to cover, this means that companies are under-insured.
Figure 6.8, instead, shows the number of insurance companies agreed with to cover the above-mentioned risks.

The majority of the companies have a single insurance coverage provider. Almost all the companies signed insurance policies with two or at most three different insurance companies, while really few have more than three. Given then the low number of insurance policies and insurance providers for every company, it is natural to check for the degree of satisfaction of the service offered. One of the questions in the survey captured exactly this variable, asking to rate from 1 (very bad) to 10 (excellent) the degree of satisfaction with respect to every single policy. On average, as shown in Figure 6.9, the majority of the interviewed are highly satisfied with the service received (more than 62% assigned a score of 7 or more), while very few did not like at all their current state of service. This might be an indication of

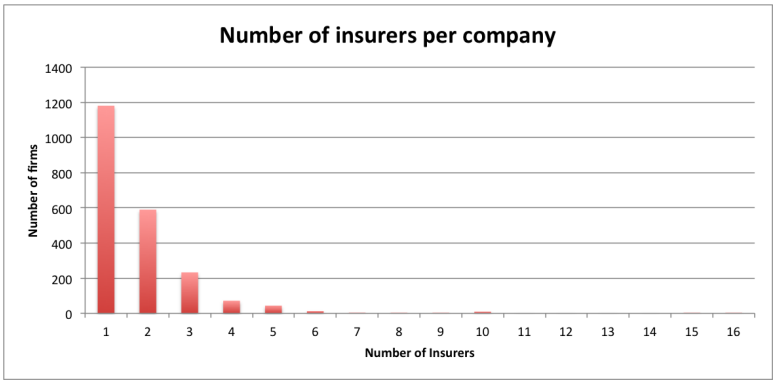



Figure 6.8: Number of insurance providers per company.

the competition intensity in the Italian market that pushes insurance companies to provide high-quality services in order to retain their clients. Since the price does not represent a barrier for customers to change insurance providers, the service quality represents an important factor to be accounted for.

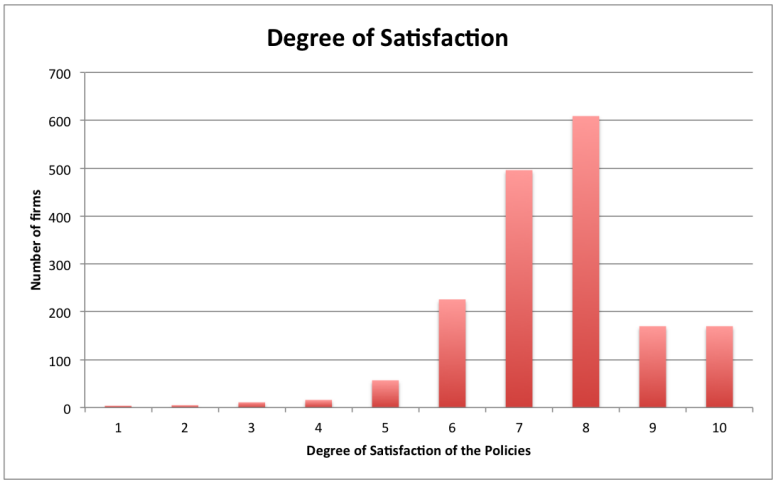



Figure 6.9: Degree of insurance policies satisfaction, per company.

Finally, often the decision of whether or not buying an insurance policy may depend on the person (or office) in charge of that decision. In small companies, the risk attitude of the company itself is somehow a reflection of who actually decides. The questionnaire also asked who took insurance decisions. Figure 6.10 summarizes the results.

Interestingly enough, the finance/credit office does not take the majority of insurance related decisions, but it is instead the administrative office or the owner herself that decide how much investing and what risks should be covered by the insurance policies. In addition, typically the person (or office) deciding on insurances is also the one dealing with banks.

### Entrepreneurs' Personal Characteristics and Behaviours

The survey asked a set of different questions that are useful to infer some behavioural aspects and firm-specific features, which are summarized below. A question of the survey asks what the entrepreneurs think it might be the (subjective) probability of their firm to be damaged by others or to damage third parties. Figure 6.11 shows those subjective probabilities, divided into 7 different intervals.
On average, many respondents estimate the likelihood of both suffering and causing damage as no higher than 5%, while a lower part (20%) has a less optimistic point of view.

Figure 6.12, instead, represents the breakdown of the above-mentioned subjective probabilities of suffering or causing damage in the following year, conditioned on having suffered or caused damage in

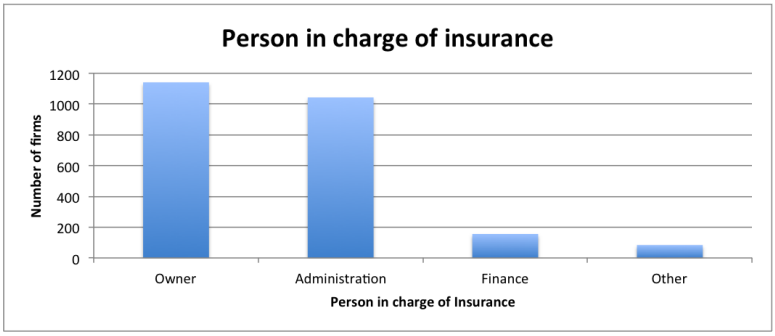



Figure 6.10: Person in charge of insurance decisions.

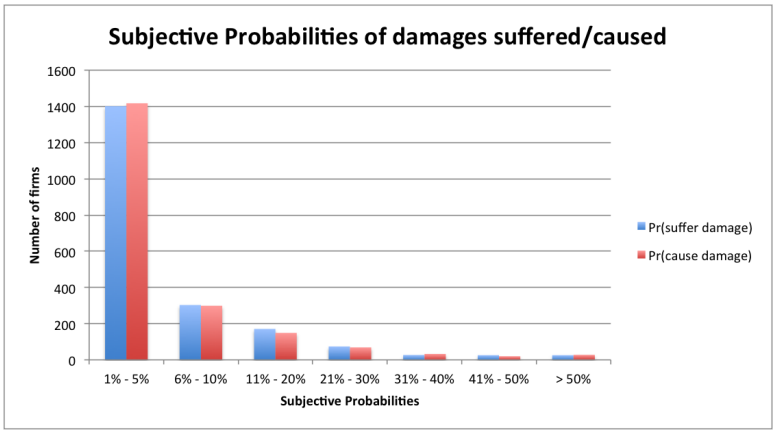



Figure 6.11: Subjective probabilities of suffering/causing damage in the following year.

the current and previous year.

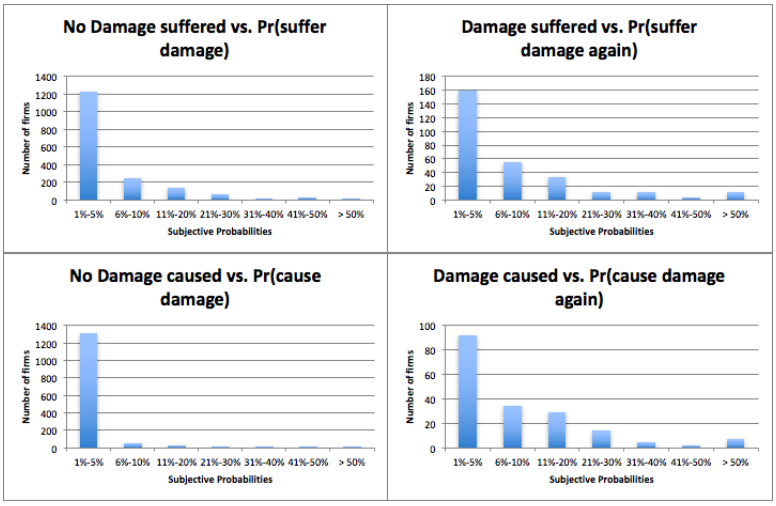



Figure 6.12: Subjective probabilities of suffering or causing damage in the following year, conditioned on having suffered or caused damage in the current and previous year.

The matrix chart in Figure 6.12 shows that i) if a firm has suffered damages in the past, many entrepreneurs expect a higher likelihood of suffering some damages again; ii) if a firm has caused damages to third parties, many entrepreneurs perceive the probability to damage others again as fairly small. However, the relative proportion of firms that think that the probability of causing (suffering) damage, having caused (suffered) it in the past is still low (1%-5%) is pretty high compared to other categories. This evidence may signal a certain degree of under-estimation of risk, typically linked to overconfidence or illusion of control. It is also relevant to understand whether the probability of causing and suffering damages are anyhow correlated. In this respect, Figure 6.13 shows that the perceived probability of causing or suffering damage affects the perception of the probability of suffering or causing damage. Entrepreneurs show a behavioural bias for which they estimate different probabilities in the same way. If the entrepreneur estimates the probability of damaging as 1%-5% (first class), and the probability of causing damage as 11%-20% (third class), the difference is two and this is the intensity shown on the x-axis. In other words, entrepreneurs are not able to accurately estimate what is their probability of causing/suffering damages, and then they simply assume that is the same probability they have to suffer/cause damages.

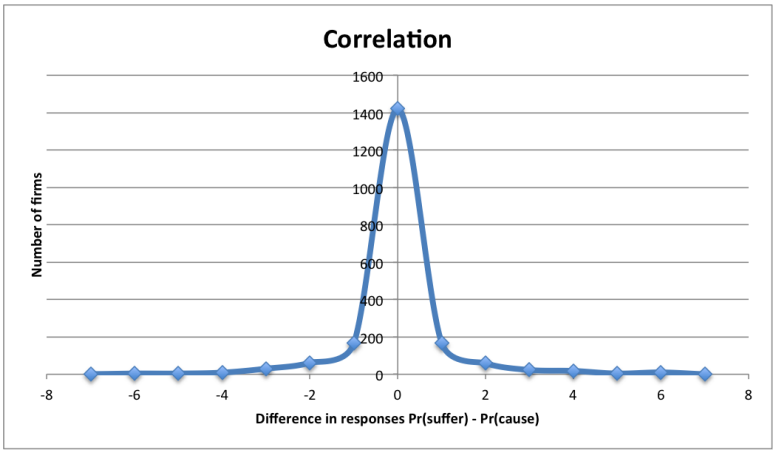



Figure 6.13: Correlation between subjective probability of suffering and causing damages.

The questionnaire provides information relative to the behavioural aptitudes of the person in charge of insurance decisions. In particular, there are questions trying to capture entrepreneur's optimism, risk

aversion, overconfidence, and attitude toward ambiguity. The survey asked whether the entrepreneurs expected more good things than bad things to occur in their business. Figure 6.14 shows that the majority of entrepreneurs are optimists with respect to their future (7 or more in the scale).

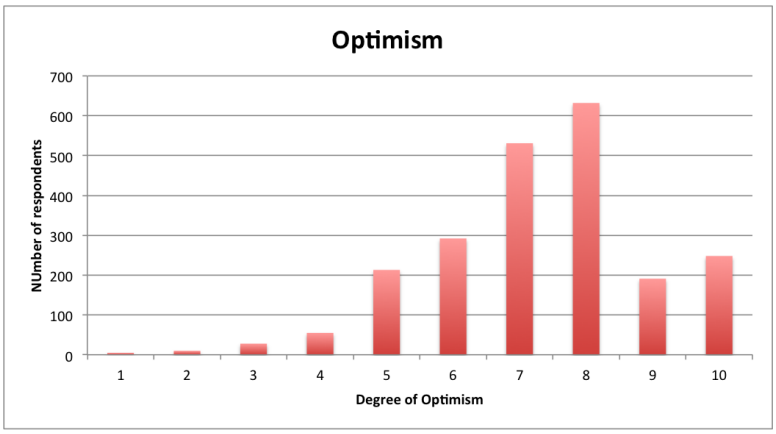



Figure 6.14: Degree of optimism in the respondents.

Optimism is a positive personality trait, while over-optimism is dangerous, particularly in business. In the section devoted to the empirical analysis, we try to detect if the entrepreneurs in our sample are "just" optimists or over optimists. Another question asked entrepreneurs if they thought to be worse/better than their peers, asking to grade themselves as on, above or below average. This question was aimed at measuring the so-called "Better-Than-Average" (BTA) effect, a type of overconfidence. Figure 6.15 shows the results.

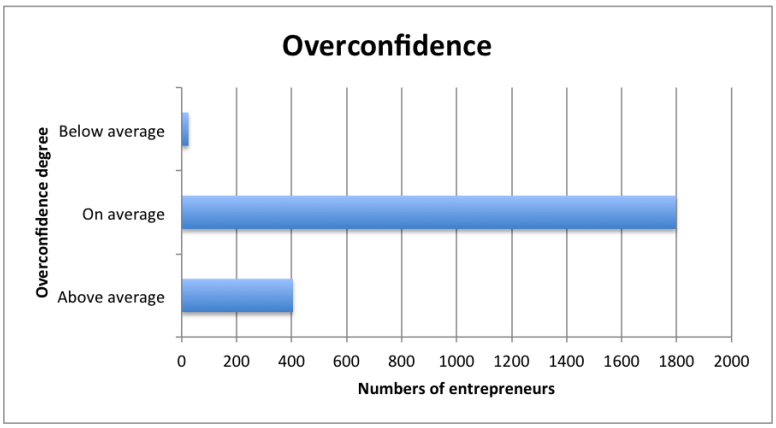



Figure 6.15: Degree of overconfidence in the respondents.

The great majority of respondents (almost 80%) rated themselves as "On average", while only a small portion (1%) believes to be below average, and the remaining 18% as above average. While this evidence may initially lead to think that entrepreneurs in our sample are not overconfident, we underline that this question was asked in the face-to-face interview. Thus, we claim, the results are probably biased. Even an overconfident person, if asked directly, may rate herself as "on average", to avoid to "show off". It thus might be the case that at least some of the respondents that answered "on average" actually perceived themselves as "above average". In the empirical analysis, we will combine the results of this answer with other proxies of overconfidence.

A different question asked entrepreneurs what was their attitude when things get harder to manage, if they prefer to quit or if they keep working, no matter what. The idea was to try to detect "stubbornness". Figure 6.16 shows the results.

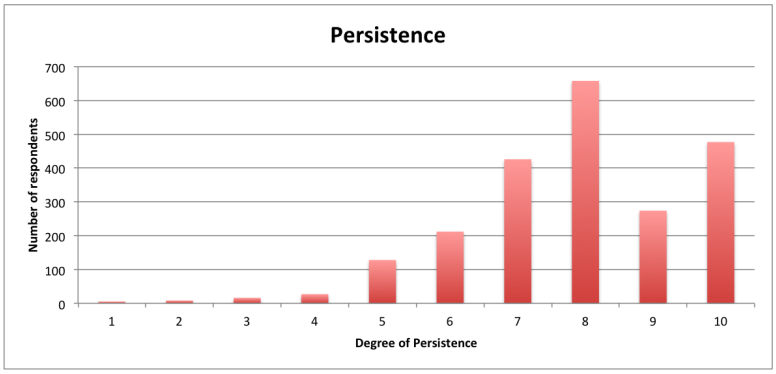



Figure 6.16: Degree of persistence (stubbornness) of the entrepreneurs.

Figure 6.16 therefore concerns the stubbornness of the entrepreneurs, or in other words their persistence and "not-giving-up" attitude. The answer ranged from 1 "I immediately give up" to 10 "I never give up". It clearly emerges that stubbornness is a strong common feature among entrepreneurs, regardless of industry sector, geographic regions, or other variables. To detect entrepreneurs' risk aversion, the survey asked two questions. The first questions asked to choose between two projects with the same cost, where the first one returned a certain amount of 1 million euros, while the second one returned either a 10 million euros with a given probability or 0 otherwise. The very same question has been asked changing the probability assigned to the risky project, as Figure 6.17 shows.

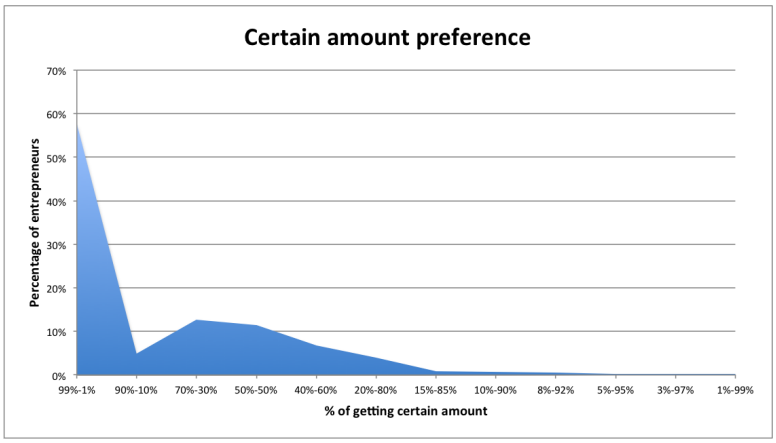



Figure 6.17: Choice trade-off between a risk- free and a risky alternative project.

The chart shows a high intrinsic degree of risk-aversion: almost 60% of the individuals prefer a certain amount of 1M over a risky bet of 10M or 0, even in the case of a likely positive scenario such as 99% of getting 10M and 0 otherwise. We also know that the switching regime point (i.e., the ratio between people who prefer the uncertain amount over the ones who prefer the certain sum) for the majority of the people to choose the bet over a certain amount is 20%-80%. This is somehow the threshold value that, according to the traditional psychology of risk, it is perceived as high to determine the strict dominance of the risky option.
What instead does not sound to be framed in the right way is that, even if they prefer a certain amount over the uncertain for the level 99%-1%, they prefer the bet if the probability trade-off is 90%-10%, but not anymore if it is 70%-30%. This is really counterintuitive, and it might mean that entrepreneurial brain works and perceive risk in a different (and maybe irrational) way. In order to cross-check, we use another question to assess the degree of entrepreneurs' risk aversion. Figure 6.18 considers the answers to the question about the investment strategy and its goal, and it confirms the high risk-aversion attitude of entrepreneurs in our sample.

In general, one may expect entrepreneurs to be less risk averse that what this evidence shows. However,

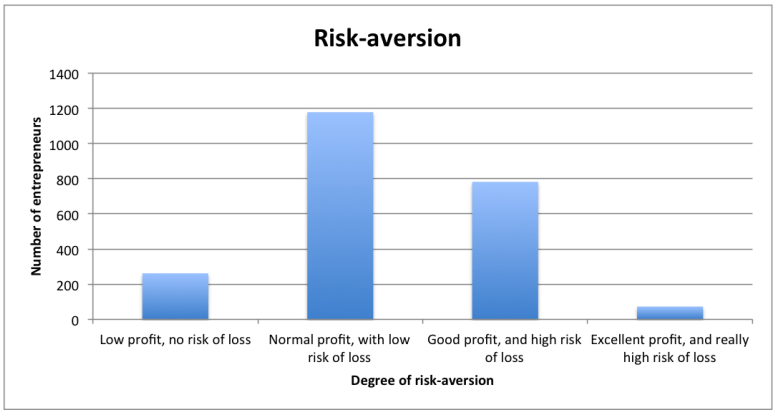



Figure 6.18: Degree of risk-aversion.

we should point out that the survey was given in the middle of the recent financial crisis. Thus, the typical risk-seeking entrepreneurial aptitude may have been replaced by a higher degree of risk aversion. Nonetheless, we would like to underline that more than 50% of respondents (almost 1,200 out of 2,295) preferred "Normal profit, with low risk of loss", and more than a third (almost 800 respondents) went for "Good profit, and high risk of loss", while only about 10% of entrepreneurs admitted to go for "Low profit, no risk of loss".
We also check if companies had installed (not compulsory) risk prevention devices (Figure 6.19, Panel A) or if they set aside some "emergency funds" to use in case of an accident (Figure 6.19, Panel B).

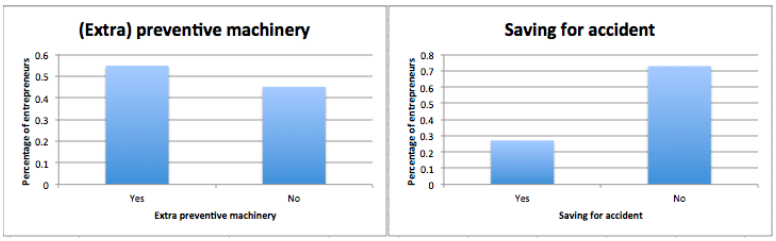



Figure 6.19: Panel A (left): Companies that have installed a risk prevention device. Panel B (right): Companies who put aside some emergency funds.

About 55% of the companies in our sample had installed risk prevention devices that were not compulsory required by law. While this may suggest that entrepreneurs correctly estimate risk and try to prevent it - and in part this is, of course, the case - we should underline that some of these devices, even if not legally required, may have to be installed in order for the insurance to underwrite a particular contract, or to lower the premium paid. For example, the insurance company may ask the company to install an alarm to prevent thefts, or proposing to lower the insurance premium in case of installation. On the other side, we also point out that 45% of the companies did not install any risk prevention device. Even more interestingly, more of 70% of the respondents did not set aside any emergency funds to use in case of accidents. In the empirical analysis that follows we will use these two just mentioned issues as possible proxies for entrepreneurial overconfidence.
Finally, two further aspects are worth analysing, i.e., ambiguity aversion and regret. Regarding the former, a standard set of questions that elicit ambiguity aversion through the choice preferences have been used, and the results are shown in Figure 6.20.

Hence, not only the respondents are risk averse on average, but also averse to ambiguity. Figure 6.21, instead, exhibits how people reacted to missed gain or unexpected loss.

The vast majority of people interviewed showed a greater regret in the case of choices that brought to a loss rather than a missed gain. In general, however, the overall the degree of regret is quite moderate.

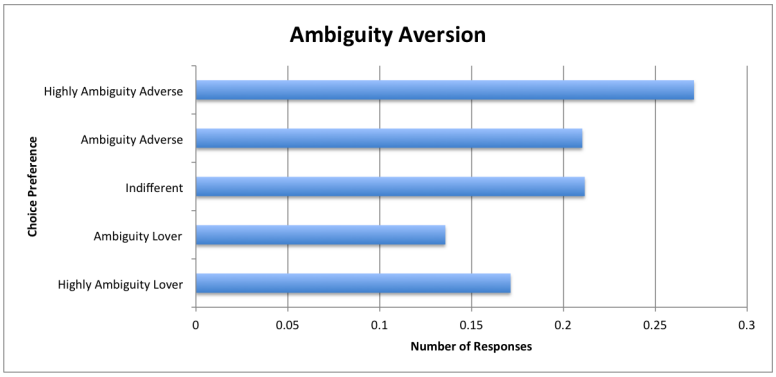



Figure 6.20: Ambiguity aversion.

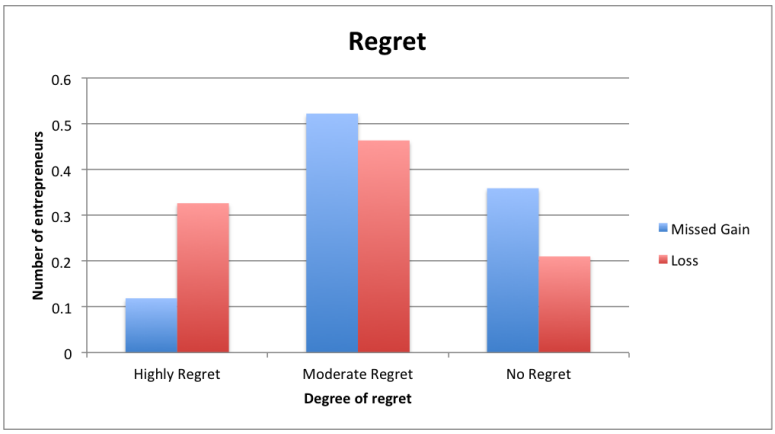



Figure 6.21: Regret reaction to missed gain or unexpected losses.

### 6.0.4 Regression Analysis

As a first step, we consider the relationship between the number of insurance policies. Table 8.5 summarises the results of our first regression model in which we consider the potential determinants of the number of insurance policies underwritten by the firms.

We performed a stepwise regression to propose the model that best fits our data. We consider the first two variables, *Employees* and *Age*, as proxies of the firm size. The higher the number of employees, the bigger it is the firm size. We also claim that the older the firm, the greater - again, on average - the size is. Both variables present a positive coefficient, suggesting that the bigger the firm, the higher the number of signed insurance contracts. This result is in line with our intuition as the complexity of a firm may increase by its size, as well as the number of risks to cover. However, this is in contrast with traditional theories affirming that bigger firms should insure themselves less, given the alternative ways of neutralizing risks. The number of insurance policies underwritten also increases when the perceived probability of being damaged rises, as it was reasonable to expect, and decreases when the likelihood of going bankrupt is high. Furthermore, if the entrepreneur asks for an advice on the insurance budgeting to a professional consultant, it seems she could dissuade the entrepreneur to purchase a policy. The second interpretation of this negative coefficient might be the trade-off between the costs of hiring this class of consultants and the residual purchasing capacity, as well as the fact that a consultant is usually hired to optimize and cutting costs (which in turn means fewer insurance policies).

*Business Name* is a variable that takes the following values: 1=Snc, 2=Srl, 3=Spa, 4=Sapa. Snc are firms with unlimited liability of its shareholders, while all the other firms have a limited liability provision. However, passing from Srl to Spa and Sapa the firm size and the complexity of the firm increase. The negative coefficient associated with this variable seems to suggest that passing from a legal form with unlimited liability (Snc) to the limited liability legal forms (Srl, Spa and Sapa), the number of insurance contracts underwritten by the company tends to decrease. A potential explanation for this result is that, given that shareholders of a firm with limited liability do risk only the money invested in the firm (and not also their own), they may afford to insure their companies less compared to shareholders of firms with unlimited liability that, instead, are also risk their personal (or household) wealth. Interestingly, having suffered damages in the previous five years leads to underwrite a lower number of insurance policies. In the same vein, having caused damages in the past is associated with a lower number of insurance contracts. The last two results are then counter-intuitive, but they could be justified with a snake-effect bias - i.e., "it already happened, and it cannot happen again to me".

As expected, instead, as the level of trust in the insurance company increases so does the average number of insurance contracts.
*Owner office* is a dummy variable indicating if insurance-related decisions are taken by the firm owner. The positive coefficient suggests that when the owner takes the insurance decisions, on average, the firm has a higher number of insurance contracts.
*Personal Life Insurance* (PersonalLD) is a dummy variable equal to one in case the respondent has a life insurance (against the case of death). The negative coefficient may suggest that the respondent considers her personal life insurance as a buffer in case something negative happens to the firm.
*Export* is a dummy that takes value one in case the company exports abroad. The associated positive coefficient may suggest that also this variable may be considered as a proxy for firm size. Typically, bigger companies do exports. We offer the same intuition for the variable *Factories*, capturing the number of branches of the firm in Italy, which it works the other way round in the specific case of the construction industry (*Building*).
Interestingly, if the company has a foreign manager, the number of insurance contracts increases. This may suggest that foreign managers have a higher sensitivity to risk management.
*Overconfidence* measure the BTA effect. In contrast with our intuition and former results in the literature, overconfident respondents tend to buy more insurance. There might be anyway an intrinsic bias in how the question has been asked in a first place, so we do not believe this conclusion around overconfidence to be universally valid.
As *Optimism* increases, also the number of insurance policies increases, in contrast with our intuition and former results in the literature. Interestingly, as *Stubbornness* increases, also the number of insurance policies increases. A behavioural explanation for this could be that entrepreneurs are aware that their

perseverance might entail extra risks, and for this reason, they would need more insurance policies to be in place.
*Loan* is a dummy variable that equals one when the firm has obtained at least a loan from a bank. The positive coefficient is in line with our intuition that banks tend to require companies to which they lend money to be insured, even if this is not compulsory by law. For example, it is typical to require a Fire insurance.
*Personal Damage* is a dummy variable that equals one when the respondent has a personal insurance against personal damages. In this case, the negative coefficient suggests that when the respondent is personally insured, her firm has a lower number of insurance contracts. This might mean that the mental accounts of business risks (and insurance) and personal risks are highly correlated, if not completely overlapped.
*Savings* is a dummy variable that equals one when the firm set aside some funds to cover damages that it can suffer in case of accidents. In this case, the negative coefficient suggests that firms treat emergency funds as a substitute for insurance. Unfortunately, the survey did not contain information on the amount of funds set aside by firms. It would definitely be interesting to know this data, because entrepreneurs may underestimate the amount of money needed to cover potential damages from accidents and incorrectly treating emergency funds as an alternative to insurance.
*AdmOffice* is a dummy variable that equals one when it is the administrative office that takes insurance-related decisions. The positive coefficient suggests that in these cases the firm tends to underwrite more insurance contracts.
As a second test, we analyse the determinants of the choice to purchase specific cover.

In addition to what previously showed in Table 8.5, we see other significant variables affecting the choice of whether purchasing or not an insurance policy against a specific event. From a behavioural perspective, *Passion* (proxied by the number of hours worked per day) and *Ambiguity Aversion* have an impact on a few policies, while the trust in other entrepreneurs, in the market, and more in general in other people play a crucial role as well. Demographics (gender, marital status, height, age, education) and the possibility of owning personal policies - death (PersonalLD), health (PersonalHealth), damages (PersonalDamage), life (PersonalLI), and insurance for social security (PersonalLP) - have controversial and different impacts on the different types of risk insured. The effect is quite different also depending on sectoral dummies (Transportation, Mining, Energy/Water/Telco, mass Production, Building, Trade), whether the company is listed or not, and if the entrepreneur still owns the majority of the shares in the company. Finally, it is also really relevant the percentage of the company value on the total personal wealth of the entrepreneur (*EV/assets*).

As it emerges from the analysis, there is not a unique formula to be applied to every insurance policy, and every determinant is different and has a different impact on the likelihood of buying a certain type of insurance. However, the list of characteristics does not explain the purchasing rationale of the entrepreneurs, but Figure 6.22 can help with that. It has been explicitly asked why certain types of insurance were not underwritten to every entrepreneur, and it seems that they overall think the risk is almost absent. There are cases in which the high cost or the fact that the insurance policy was never proposed to the entrepreneur are further reasons to not buy insurances. However, since the data confirms that the risk perception is the first rationale for not purchasing the insurance policies, we might wonder whether this is a direct consequence of an overconfidence bias, or simply a wrong calculation of the risk.

A final interest aspect we decided to try to understand was then why entrepreneurs actually buy insurance policies. The intuitive answer is because they either expect to cause damages to someone (and they want to avoid to pay out-of-pocket expenses), or they want to prevent someone else to make consistent damages that could impair their company. Hence, we analysed the determinants that affect the perceived likelihood of suffering or causing damages that exceed the mean of the sector (Table 8.7).

The results we obtained are really useful and intuitive: the probability of suffering damages is positively linked to operations in the specific mining industry, to the increasing chances to go bankrupt and if you are listed (you have a greater exposure to the market in both the cases), and minimally affected by the percentage of personal wealth the entrepreneur uses for her company. On the contrary, it is negatively impacted by the firm savings (if you save more, you are more prepared to whatever negative event might come), the fact that you might have been already damaged in the last five years ("snake-bite effect"),

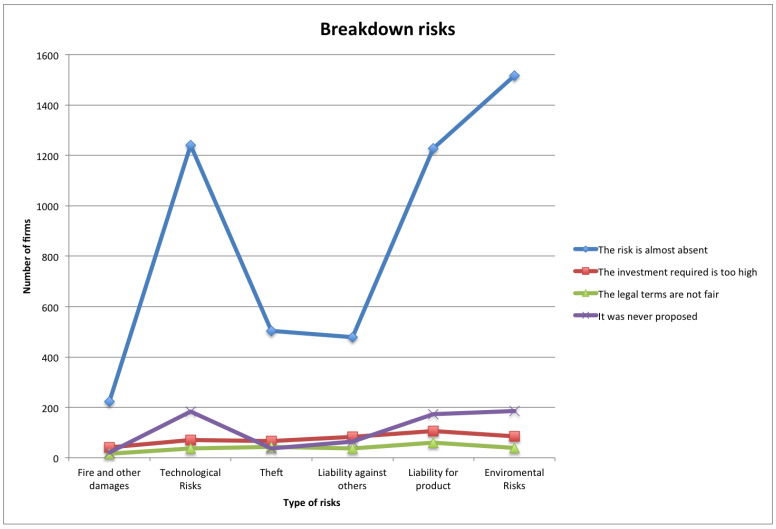



Figure 6.22: Reasons for not buying insurance.

but above all is negatively correlated with behavioural traits such as overconfidence and stubbornness. In fact, if the entrepreneur is particularly overconfident and stubborn could not consider the worst-case scenario in which her activity is damaged by external sources. The likelihood of damaging others is correlated in the same way as the other with a few variables (savings; personal wealth invested in firm; stubbornness; bankruptcy) but it is also related to the number of employees (the bigger the company more possibilities there are to damage third parties); age of the company (the older the company, the more experienced it gets and the lower the probability becomes); if you have already been damaged in the last five years (snake-bite effect again); high passion of the entrepreneur is usually associated with higher attention to details and ethical behaviour; specific activities (negatively with trade and mass production, and positively with the transportation sector); if the entrepreneur has already been personally damaged, she tends to be more careful in damaging others as well; and finally, and maybe counterintuitively, the increasing trust in others and the higher numbers of competitors going bankrupt might increase the perceived likelihood of damaging others.

### 6.0.5 Conclusions

Entrepreneurs should be rational decision-makers when it comes to their businesses, and they should always act in the best interest of their companies. Unfortunately, this seems to be disproved by empirical analysis. More in details, we analysed 2,295 small and medium companies in the Italian market, and we studied the effect of behavioural biases on entrepreneurial choices related to different kind of corporate risks. We concluded that Italian entrepreneurs under-insure themselves and that this decision is conditioned by many variables and events: for example, entrepreneurs who suffered or caused damages in the past 5 years are likely to buy less insurance, and the same is true even in the case of likely bankruptcy. On the other side, trust leads to buying more insurance, and we found personal and business insurance to be substitutes.

# Chapter 7

# Third Application: Sentiment Analysis

## 7.1 The power of micro-blogging: how to use Twitter for predicting the stock market


### Abstract

The availability of new data and techniques enriched the existing extensive literature on the importance of investors' sentiment and on his impact of the stock price oscillations. The purpose of this paper is to exploit micro-blogging data in order to construct a new index-tracking variable that may be used to earn some insights on the Nasdaq-100's future movements. The results are promising: the models augmented with the newly created variable show an incremented explanatory power with respect to the benchmark.


### 7.1.1 Introduction and literature review

Nowadays, financial markets are increasingly volatile and difficult to understand. In order to reach a greater comprehension of the mechanisms that regulate the markets, it is thus necessary to take into account new and unexpected source of data in the creation of pricing and forecasting models. New technologies and the capacity to gather and store a huge amount of data (i.e., big data) allow us to take the cue from data which were not available ten years ago, but that are now extremely relevant and with strong explanatory power. The most probably innovative kind of data we can exploit is the one coming from the social networks, e.g., Facebook, Twitter, Instagram. They all are able to give us some extra insights and further information in order to build more efficient trading models (Asur and Huberman, 2010). Recently, the use of social networks and micro-blogging platforms is becoming extremely popular. A huge spectrum of distinct applications may be found in completely different fields, e.g., predictions of presidential elections (Tumasjan et al., 2010), music albums release (Dhar and Chang, 2009), epidemics and disease spread (Culotta, 2010), movie revenues (Mishne and Glance, 2006) or commercial sales (Choi and Varian, 2012) forecasting. Although the use of these new sources of data may be extremely clear for some of the above-mentioned applications, it might not be so well defined for financial markets. The fundamentals may in fact not be able to explain everything, but we could gain some insights from news and information (Nosfinger, 2005; Peterson, 2007). We should then investigate how we could use the socials to build new efficient forecasting trading models and what kind of information they provide us with. Even though a wide knowledge could be inferred from them, what we focus on is how to extrapolate the investors' sentiment from their opinions on the web.

The literature on how embedding the investors' sentiment into trading, pricing or forecasting models is quite varied: Da et al. (2012) propose first a direct measure of investor demand for attention using search frequency in Google for a five-year sample of Russell 3000 stocks, while in a second work (2015) they use daily Internet search volume from millions of households to reveal market-level sentiment and to build a new index as a new measure of investor sentiment. Furthermore, Tetlock (2006), and Tetlock et al. (2008) showed how different financial languages in the financial news affect the stock returns. More in details, it seems that i) the impact of the negative news is larger for the stories focused on the fundamentals, ii) negative words in firm-specific news predicts better low firm earnings, and iii) the

prices shortly underreact to this information. Fisher and Statman (2000) instead dealt with the investors' sentiment in relation with tactical allocation strategies, while Baker and Wurgler (2006, 2007) studied incorporate to some extent some behavioral biases into the stocks selection process.
Regarding instead the social networks and micro-blogging uses for financial markets, a robust literature is arising in the last few years. Agarwal et al. (2011) and Ruiz et al. (2012) showed how to use micro-blogging data in the stock market and how they are correlated with financial time series. However, one of the probably most popular works in this field is the one from Bollen et al. (2011). Several other works (Bollen and Mao, 2011; Mao et al., 2011), and Mittal and Goel in another study (2012), used Twitter to forecast the stock prices for general index such as Dow Jones Industrial Average based on different investors' mood. A very recent paper (Mao et al., 2015) analyzed instead the tweets predicting power for international financial markets, obtaining very good results for countries such as United States, United Kingdom, and Canada. Zhang (2013) and Brown (2012) followed the same trend - with slight variations - while Oliveira et al. (2013) found a positive effect of the posting volume on robust forecasting. Instead, Sprenger et al. (2010) proved how the sentiment of the tweets is indeed associated with abnormal stock returns and message volume. Finally, Oh and Sheng (2011) provided a model for irrational investor sentiment. Nevertheless, there also exists a vast literature on alternative sources of data relevant for financial modeling purposes, such as blogs (De Choudhury et al., 2008), security analyst recommendations (Barber et al., 2001), web search queries (Bordino et al., 2012), stock message boards (Antweiler and Frank, 2004; Koski et al., 2008), or simply financial news (Schumaker and Chen, 2009; Lavrenko et al., 2000). It then seems clear that it may be valuable to try to integrate this extra information into our forecasting models, and this is indeed the purpose of this paper. The structure of the work will be as follows: in section 7.1.2 we present the data and the methodology used, in the section 7.1.3 we show some empirical results, and in section 7.1.4 we conclude.

### 7.1.2 Data and methodology

The micro-blogging data are nowadays widely used and available to the research community. They can be obtained through several means, and they usually only report basic information. In our analysis we use Twitter data, for which the basic information that may be exploited are the text of the tweet itself, the username, the hour and location from which it has been posted, and the gender of the user. We decided to obtain these data from the DataSift provider, mainly because in addition to the usual information it also assigns a sentiment score to each tweet. This scoring system is built in such a way that an algorithm can consistently rate the positivity/negativity of a particular text. Within DataSift, this score may typically swing between -20 and +20, even if particular topics/texts require sometimes a higher/lower evaluation. In order to build our estimator for the Nasdaq-100 Index, we collected tweets for a two-months period, for three major technology stocks belonging to the index, i.e., Apple, Google, and Facebook. We decide to select only three out of 100 stocks comprised in the index in order to perform a sort of an ex ante feature selection: not every stock in the bucket is useful for prediction purposes, but on the principal component analysis design fashion, some stocks have a greater explanatory power with respect to others, and are the only worthy to be used. We therefore skimmed the data to take into account only the English-speaker users, and we took into account only the tweets that showed a pre-existing user's financial knowledge. We only selected the tweets containing the company's ticker, respectively AAPL, GOOG, and FB.[1] Hence, for the period that goes from September 24th to November 21st 2014, we were able to gather about 88,000 tweets for Apple, 43,600 for Facebook, and almost 32,000 for Google, as illustrated in Figure 7.1.[2]

We construct two variables: a sentiment mean variable for each stock, averaging the sentiment score on a daily basis, and a tweets volume, a simple counting variable of the number of tweets for a certain stock on a certain date. Hence, we test the predictive power of individual stocks on the Nasdaq price and volume, with respect to a benchmark autoregressive process:

[1] This filter was easily constructed through the CSDL coding environment - a specific language provided by DataSift for this purpose The following codes have been used for filtering the data:

- twitter.symbols contains "AAPL" ;
- twitter.symbols in "GOOG, GOOGL" ;
- twitter.symbols contains "FB".

[2] The data for the price and volume of the Nasdaq-100 have been instead obtained through Yahoo!Finance.

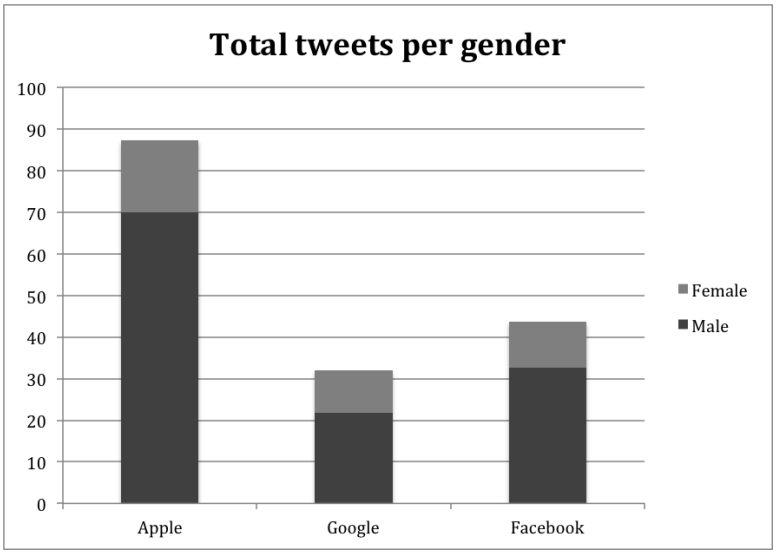



Figure 7.1: Number of total tweets per gender for each stock

$$M1 : P_t = \alpha + \phi_1 P_{t-1} + \epsilon_t \tag{7.1}$$

where $P_t$ is the Nasdaq price, $P_{t-1}$ is the lagged Nasdaq Price, and $\epsilon_t$ is the error.

$$M2 : P_t = \alpha + \phi_1 P_{t-1} + \beta_1 SM_{Apple} + \beta_2 SM_{Facebook} + \beta_3 SM_{Google} + \epsilon_t \tag{7.2}$$

where $SM_i$, for i = Apple, Google, Facebook, represents the sentiment mean related to that specific company. The first model represents the benchmark, while the second one is augmented for the sentiment mean variable for each stock.
Similarly, we implemented the same models for the volume as well:

$$M3 : Volume_t = \alpha + \phi_1 Volume_{t-1} + \epsilon_t \tag{7.3}$$

$$M4 : Volume_t = \alpha + \phi_1 Volume_{t-1} + \beta_1 TV_{Apple} + \beta_2 TV_{Facebook} + \beta_3 TV_{Google} + \epsilon_t \tag{7.4}$$

where *Volume* is the Nasdaq transactional volume, and $TV_i$ the tweets volume for a certain stock in a certain day.
We then construct our *sentiment index-tracking* (SIT) variable, as the average of the sentiment mean for each stock in a certain day weighted for the respective tweets volume:

$$SIT_t = \frac{SM_{Apple} TV_{Apple} + SM_{Facebook} TV_{Facebook} + SM_{Google} TV_{Google}}{3} \tag{7.5}$$

We therefore augment the autoregressive benchmark models for the SIT variable so constructed (M5 for the price, and M6 for the volume):

$$M5 : P_t = \alpha + \phi_1 P_{t-1} + \phi_2 SIT_{T-1} + \epsilon_t \tag{7.6}$$

$$M6 : Volume_t = \alpha + \phi_1 Volume_{t-1} + \phi_2 SIT_{T-1} + \epsilon_t \tag{7.7}$$

### 7.1.3 Empirical results

We perform an ordinary least square regression for every model previously exposed. In particular, the estimations obtained are shown in the Table 8.8.

As it can be noticed, the models where the single sentiment means or volume are taken into account (i.e., M2 and M4), are poorly explicative and non statistically significant. It is thus clear that further specification is needed in order to create an efficient forecasting model able to identify the specific source of price variation. It is instead interesting how the SIT variable well captures some further useful information. Indeed, this variable is statistically significant for both price and volume forecasting (M5 and M6), and above all it is able to improve both the adjusted $R^2$ and the root means square error (RMSE) of the model, as shown in Table 8.9.

The models augmented with the SIT variable show an improvement both in term of $R^2$ and RMSE with respect to their own benchmark.

### 7.1.4 Conclusions

Our results are not fully conclusive, because the model cannot be generalized to any sector, stock or index yet. However, we can infer some important insights on how to incorporate new kinds of information into trading and forecasting model.
We created a dataset for a two-months period using data from Twitter, and we constructed some indicators in order to forecast the Nasdaq-100. We then compared our models to the benchmarks, and it seems that they are able to increase the explanatory power and to provide a better prediction of the Nasdaq price and volume. Both the adjusted $R^2$ and the RMSE improved as a consequence of the index we built, and even if we may improve the model in the future - we can test it for different sector and for a different time frame, as well as for a different frequency - it gives anyway great intuitions and superior advantage for a potential trading strategies built on it. It would be also interesting to assess how the sentiment might better (or worse) capture the stock market oscillations in crisis or boom periods, and finally to see whether other socials may be actually helpful in term of stock market predictions.

## 7.2 Why social media matters: the use of Twitter in high-frequency portfolio strategies

### Abstract

In previous works ([9], [14]), it has already been showed that Twitter and social media in general give an interesting additional predictive power to the models that take them into account. However, the contribution of social media is relatively small on a daily basis, because of the speed and the increasing efficiency of the stock markets. It has been decided then to deal with intraday prices to test whether micro-blogging data may actually be used to implement high-frequency forecasting models. It has been constructed an indicator to earn some insights on the Nasdaq-100's future movements. Once again, the results are very encouraging: the use of social media data increases the predictive power for general stock market index such as the Nasdaq, and becomes thus an essential building block for any pricing model.

### 7.2.1 Introduction and literature review

Nowadays, the two main characteristics for any financial market or stock exchange are the enormous volumes traded, and the increasing speed of the order submitted and/or filled. Big data techniques and models from one hand, and computers and financial algorithms from the other hand, are changing the way we approach the markets and the tools we use. They are also breaking down the barriers regarding both the type and the amount of the data we can use to feed our models, so that interesting insights can be earn by a variety of different sources. The majority of this information is quantitative and easy identifiable, such as prices, volumes, volatilities, and so on so forth. On the other hand though, it is quite cumbersome to find a way to quantify measures such as the market or investors sentiment. An extensive literature exists on both theoretical models and empirical applications that study how to embed this factor into a forecasting model and portfolio strategies: back in the first half of the century, Fisher and Statman ([23]) dealt with the interaction of investors' sentiment and tactical allocation, while only with Baker and Wurgler few years later ([4], [5]) a better comprehension has been achieved regarding how to incorporate behavioral biases and thus the market sentiment into the stocks selection process. Furthermore, Da et al. (2012) proposed in a first place how to quantify the investors demand using search frequency in Google for a five-year sample of Russell 3000 stocks, and afterwards ([18]) daily Internet search volume to construct a new index able to assess the investors' sentiment. Tetlock et al. ([40], [41]) proposed some analysis in which different financial languages in the financial news affect differently the stock returns. The impact of the negative news is thus larger for the stories concerning the fundamentals, and negative firm-specific words have a greater forecasting power on low firm earnings - to which the prices shortly underreact. Many other different works have been written about market sentiment and the investors' perceptions, but only recently the field has experienced a turning point, i.e., when social media became part of the equation in formulating a more efficient trading model ([3]). The applications for social networks have been indeed manifold: epidemics and disease spread ([16]), movie revenues ([31]) or commercial sales forecasting ([13]), presidential elections ([42]), music albums release ([22]), and financial decision making ([33], [36]). Furthermore, it is also true that different kind of sources have been analyzed for this last purpose, e.g. financial news ([27], [38]), blogs ([21]), web search queries ([11]), stock message boards ([2], [26]), and security analyst recommendations ([6]). Bollen et al. (2011) - and in a series of other works ([10], [28]) - have first deepened how Twitter could be used to forecast the Dow-Jones Index using a spectrum of different human emotions, and similar applications have been analyzed then by Mittal and Goel ([32]), Zhang ([43]), and Brown ([12]). In a most recent paper, Mao et al. ([29]), exploited tweets predicting power in order to understand the international financial markets trend for several countries, including the United States, United Kingdom, and Canada. At the same time, Agarwal et al. (2011) and Ruiz et al. (2012) studied the existing correlation between micro-blogging data and financial time series. Then, Oh and Sheng ([34]) created a model for irrational investor sentiment, Oliveira et al. ([35]) assessed a positive effect of the Twitter volume on robust forecasting, and finally Sprenger et al. ([39]) proved how abnormal stock returns and message volume are associated to an increment in the posting volumes. Hence, differently from any other work before, this one is going to focus on the use of tweets about few stocks in order to predict the trend of a market index. The structure of the work will be as follows: first, the data will be presented and described. Then, some new indicator-tracking variables will be built and then different forecasting models will be tested, to finally assess the differences from the

autoregressive benchmark model. In the section 7.2.3 some empirical results will be showed, to conclude then in section 7.2.4.

### 7.2.2 Data and Methodology

Since the aim was to analyze the Nasdaq-100's behavior, three of the major technology stocks belonging to the index have been taken into account, i.e., Apple, Google, and Facebook. The reason why is quite intuitive: hundred companies compose the index, but clearly not every company has the same weight on the bundle. Hence, selecting ex-ante the biggest ones, it has been reduced the model complexity and the number of features to be taken care of. In other words, this prior could be considered to have the same function of a "qualitative" principal component analysis, which allows us to shrink the model and allocate a greater explanatory power to firms that are more meaningful to the index. First of all, the data for the intraday prices for the Nasdaq-100 have been obtained through Bloomberg. It has been collected therefore a dataset for the period that goes from September 24th to November 21st 2014, and it was possible to gather almost 88,000 tweets for Apple, 43,600 for Facebook, and slightly less than 32,000 for Google. There are indeed many ways to gather this kind of dataset, e.g. through APIs or similar, but it has been used instead a data provider called DataSift because it was able to supply a scoring algorithm for the tweets' content. Figures 7.2 - 7.4 display indeed the overall tweets volume for each stock (blue bars), and the sentiment means for the daily tweets (black lines). The black lines are indeed built so that the lowest value represents the average for the negative tweets, while the highest extreme is the mean for the positive ones, and the small dashes the overall daily averages.

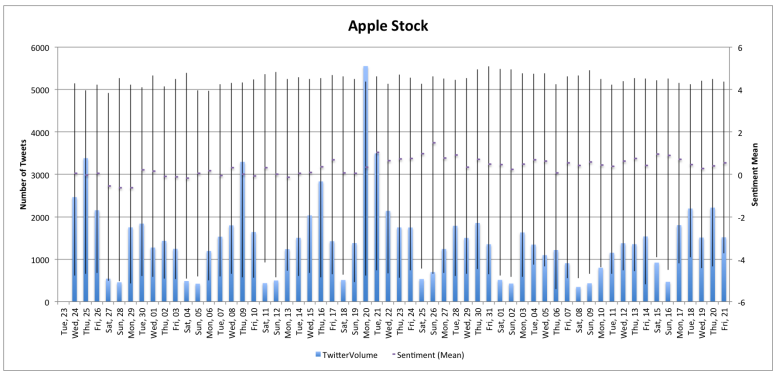



Figure 7.2: Twitter Volume, Mean of positive sentiment, Mean Negative sentiment and Daily sentiment mean for Apple.

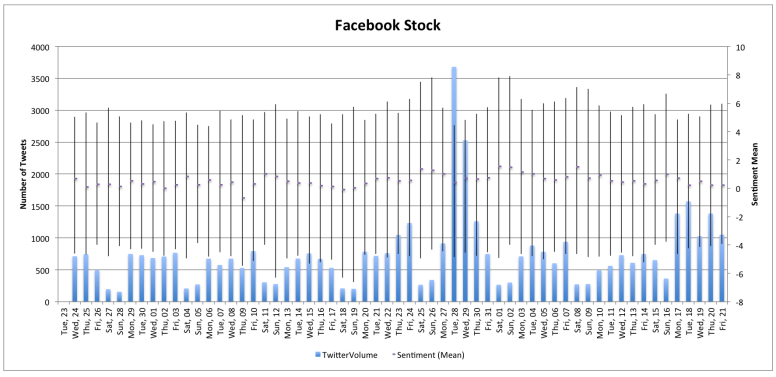



Figure 7.3: Twitter Volume, Mean of positive sentiment, Mean Negative sentiment and Daily sentiment mean for Facebook.

In addition, DataSift provides a wide spectrum of information, such as the gender, location, time, username, and much more. For the preliminary analysis, many of them were actually meaningless, but additional studies may be implemented using this information. Further noise comes directly from the

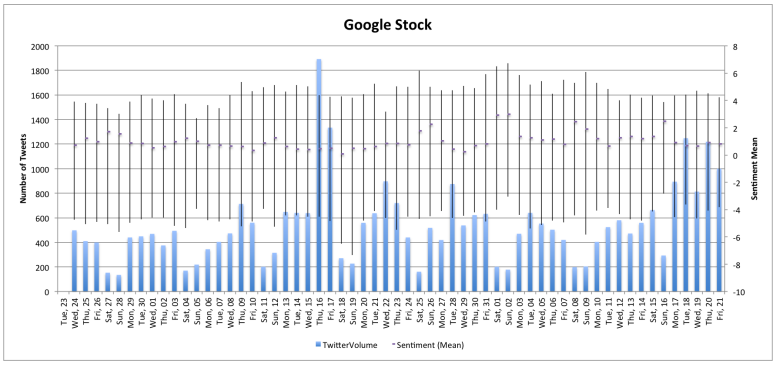



Figure 7.4: Twitter Volume, Mean of positive sentiment, Mean Negative sentiment and Daily sentiment mean for Google.

tweets text, which are often irrelevant for the study. Hence, it has been decided to consider only the tweets in which some extent of financial knowledge was observed, i.e., only the tweets in which appeared the stock's ticker. Every tweet that was not in English has been eliminated during this first step, since taking into account other languages was not relevant to the study per se - and because they represented a small portion of the dataset as well. Regarding how the scoring system works, the underlying algorithm assesses how positive or negative is the text of certain tweet. For this work, the range of this rating oscillates between -20 and +20, even if particular topic/text requires sometimes a higher/lower evaluation. The figure 7.5 shows the daily volatility of the scores with respect to the stock volatility. The first ones are quite stable, while of course the stock variations are really volatile.

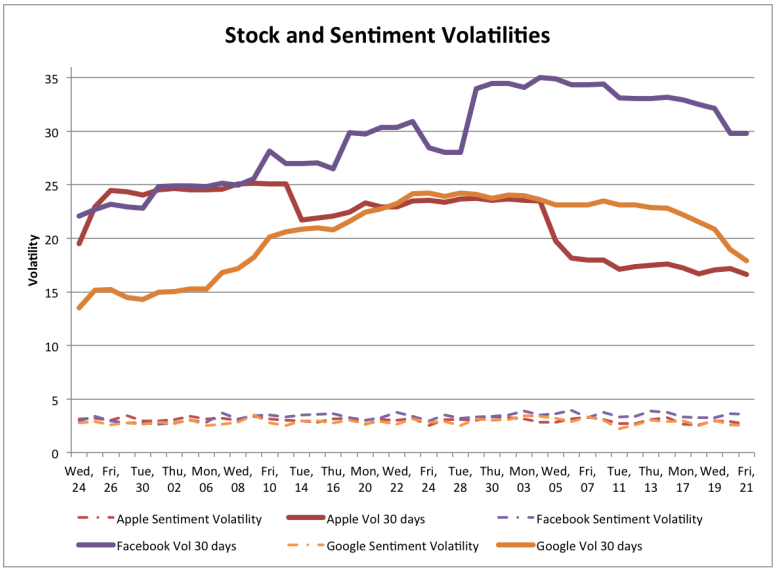



Figure 7.5: Stock volatility (lines) and sentiment score volatility (dashes) for Apple, Google and Facebook.

The following step was the construction of relevant variables for the empirical analysis. It has been indeed built a variable for the *sentiment mean*, taking the simple average for the tweets' scores on a minute basis; it was also computed the time volume moving average, and a sentiment moving average (SMMA), where both of them are five-minutes moving average; finally, two variables for tracking the Nasdaq-100 were created: the *sentiment index-tracking* (SIT) and the *weighted sentiment index-tracking* (SITw), respectively

$$SIT_t = \frac{SM_{Apple}TV_{Apple} + SM_{Facebook}TV_{Facebook} + SM_{Google}TV_{Google}}{3} \tag{7.8}$$

$$SITw_t = \frac{SM_{Apple}TV_{Apple} + SM_{Facebook}TV_{Facebook} + SM_{Google}TV_{Google}}{TV_{Apple} + TV_{Facebook} + TV_{Google}}. \tag{7.9}$$

There have then created the equivalent moving average variables, i.e., *SITma* and *SITwma*. Hence, two different set of regressions were run: the first one was the standard one, in which the dependent variable was always the Nasdaq value at a certain minute. The second block concerned instead the Nasdaq's variations, so in other words the *direction* or *trend* the index was assuming. Afterwards, the first thing has been setting the benchmark model, i.e., a simple autoregressive model such as

$$M1 : P_t = \alpha + \phi_1 P_{t-1} + \epsilon_t. \tag{7.10}$$

Secondly, it has been tested whether the hypothesis of grouping the three stocks was indeed useful, or if maybe each of them had a different impact on the Nasdaq price:

$$M2 : P_t = \alpha + \phi_1 P_{t-1} + \beta_1 SM_{Apple} + \beta_2 SM_{Facebook} + \beta_3 SM_{Google} + \epsilon_t. \tag{7.11}$$

For a robustness check, it was run the same model for the sentiment five-minutes moving averages:

$$M3 : P_t = \alpha + \phi_1 P_{t-1} + \beta_1 SMMA_{Apple} + \beta_2 SMMA_{Facebook} + \beta_3 SMMA_{Google} + \epsilon_t. \tag{7.12}$$

Finally, the other model embedding the simple sentiment index-tracking variables has been tested, i.e.,

$$M4 : P_t = \alpha + \phi_1 P_{t-1} + \phi_2 SIT_{T-1} + \epsilon_t \tag{7.13}$$

and then the same has been implemented for the weighted version, the moving average one, and finally the weighted moving average, respectively *M5*, *M6*, and *M7*. In a perfectly symmetric way the same has been done using, instead of price variables, the direction (or trend) variable, i.e., a simple dummy variable that took value one if the ratio between the prices today and yesterday price was greater than one, zero otherwise (*M8 - M14*).

### 7.2.3 Empirical results

It has been used an ordinary least square regression for the models from one to seven, while the linear probability model has been used for the regressions M8 - M14. The results from the regressions are shown in Table 8.10 and Table 8.12 , while Table 8.11 and 8.13 exhibit the root mean squared errors and the adjusted $R^2$ for all the models. These two tools can be used to compare the models at a glance.

Beginning from the price regressions, the models where the single stocks are taken into account (M2 - M3) seem to increase the accuracy of the forecasts with respect to the benchmark model, but unfortunately almost none of the results are statistically significant. On the other hand, the sentiment index-tracking variables give positive results: even if the moving averages are not significant as well, the simple SIT or the weighted one are significant and meaningful. Besides, Table 8.11 shows the improvement of using this model with respect to the benchmark. In particular, it turns out that the simplest one (M4), in which the easiest version of SIT has been used, is the one that performs the best. The results from the second block of LPM regressions do not show any inconsistency with what just claimed above. The outcomes and considerations are perfectly symmetric, and once again the SIT-model seems to be the most explicative one, able to reduce the RMSE and anticipate to some extent the market trend.

### 7.2.4 Conclusions

Both the academic literature and the industry professionals are approaching social media as source of interesting insights. A strong hidden value is contained in this new information channel, and financial markets could exploit it as well. Even if the results are quite simple, specific to the technological sector, and still preliminary because of the limitless work that could be done about it, they give us many foods for thoughts. The high frequency nature of this new dataset is indeed able to capture some price variations for the Nasdaq-100 way better than basic forecasting models. To prove it, it was built a dataset for a

two-months period using data from Twitter, and then it was created a synthetic way to track the general index about the Nasdaq though his three main companies, i.e., Apple, Google, and Facebook. Different analysis were run, and it was used a comparative augmented approach to evaluate their differences in performance. With respect to the simple autoregressive benchmarks, the explanatory power and the accuracy achieved by the sentiment-models are bigger, and this pushes us to create new models in which human judgment, sentiment and opinions play a role. The study is only a first look on this immense field, and many more improvements could be done in future: different sector or regions, longer time series, event studies related to corporate or market events, or particular situations of the market cycle (crisis, expansion, etc.).

## 7.3 Sentiment Analysis for Stock Market in Technology Sector

### Abstract

Bollen et al. ([10]) reintroduced the idea of formulating prediction based on the general sentiment of the investors, even if they originally exploited microblogging data. The purpose of this study is to verify whether social data may have a predictive power for the stock prices, returns, and volumes. The analysis has been implemented for different large technology companies, and the robustness has been tested through a ten-days rolling window. The evidence shows that there is some intrinsic value in these new features, and that both the sentiment and the amount of tweets posted online can improve the forecast given by a baseline autoregressive model.

### 7.3.1 Literature Review

The higher ability of storing large datasets is increasing our chance to exploit new kind of data, such as for instance the data coming from the web and the social networks. The applications for this innovative sources are manifold, such as epidemics ([16]), presidential elections ([42]), commercial sales ([13]), movies revenues ([31]) or music albums ([22]), but one that is becoming extremely interesting both from an academic and a practitioner point of view is how to use these datasets to implement robust forecasting models ([3]), particularly in order to build effective trading strategies.

The idea of investigating the investors sentiment in order to better understand capital markets is well-established in literature ([17], [18]), but of course unused data could provide new insights on the phenomenon. A comprehensive dissertation on the different kinds of investors sentiment and their use for tactical allocation has been treated in Fisher and Statman ([23]). On the other hand, Tetlock et al. ([40], [41]) already proposed how a different language used in the financial news could affect the stock returns. They actually proved that certain negative words in firm-specific news forecasts lower firm earnings, and also how prices shortly underreact to this information. Furthermore, they verified that the impact of the negative news was larger for the stories focused on the fundamentals. Instead, Baker and Wurgler ([4], [5]) showed a top-down approach to measure behavioural biases (i.e., limits to arbitrage) and investor sentiments in order to explain which kind or class of stocks would be more affected by investor's opinions. Clearly an extensive literature exists on how quantitative information are embedded into stock prices, and they could fill to some extent the gap left by the Efficient Market Hypothesis (EMH) failure. Indeed, it seems that a portion of the price forecasting may not be attributed purely to the fundamentals, but that also something additional may be used to increase the explanatory power of our models ([33], [36]).

While earlier studies focused on data coming from stock message boards ([2], [26]), from financial news ([38], [27]), from blogs ([21]), from web search queries ([11]), or from security analyst recommendations ([6]), more recent studies exploited eventually data from microblogs ([1]) and how microblogging activity may be correlated with financial time series ([37]). The reasons why the cutting edge research is shifting toward these new sources are manifold: first of all, social networks are not a static source of information, as for instance a television program, a commentary, or a radio broadcast are, but they evolve continuously and rapidly over the same day, and they are constantly updated. An important aspect of this dynamism is the ability to record not only pure information or opinions, but feedbacks as well, so that they can be thought as a complex neural system that changes every minute and where the signals go back and forth from a single point, modifying this point each time they reach it. In addition, they do not capture only the opinion of a single individual (as the blogs do, for instance), but they pull and gather a multitude of different viewpoints, covering basically the whole spectrum of different individuals and embedding all the various nuances. An extra consideration to be taken into account is that the opinions shared on the socials have an intrinsic predicting power since they are forward-looking, and can be used as ex-ante information for any forecasting model, while other sources such as TV, company's website, etc., are more ex-post information, and since they look retrospectively to the company's events and/or news, they run out of any predicting power. Hence, the socials are not only a new mechanism to disseminate information, but rather a way to create new information that was not available up to now.

For all these reasons, and of course for the incredibly huge amount of data the social networks are collecting and making available - and this is particularly true for Twitter, many researchers are using

them as the new main source of information and as a fundamental building block for innovative forecasting models. Indeed, Bollen et al., for instance, in several works ([10], [9], [28]), and Mittal and Goel in a later study ([32]), used the data coming from Twitter in order to forecast the stock prices for general index such as the Dow Jones Industrial Average index based on different investors' moods, or also for predicting international financial markets ([29]). Corea and Cervellati ([14]), and Corea ([15])showed instead how to use Twitter data in order to create a sentiment index with some predictive value for the Nasdaq-100, using only three of the biggest companies that compose the index. The same kind of analysis has been implemented in other publications (see for instance Zhang, [43], or Brown, [12]) and many of them showed very promising results: Oliveira et al. ([35]) concluded positively on the effect of the posting volume on robust forecasting, while Oh and Sheng ([34]) provided a model for irrational investor sentiment and finally recommended an investor approach using user-generated content deriving from microblogging activity. Furthermore, Sprenger et al. ([39]) claimed the sentiment (i.e., bullishness) of tweets to be associated with abnormal stock returns and message volume to predict next-day trading volume.

Hence, the purpose of this paper is to provide a simple pricing model that embeds new measures of sentiment index to improve the forecast ability of the standard analysis. Contrary to earlier studies, we focus on some particular larger companies in the technology sector in a short time frame, and we try to provide an individual pricing/forecasting model instead of a more general one concerning a big index such as the DJIA ([10]). This paper is then structured as follows: the next section is going to provide a detailed explanation of the dataset collected, on the modelling and the methodology applied. Section 7.3.3 shows the empirical results of the study, to finally draw some conclusions in section 7.3.4 and suggest future improvements.

### 7.3.2 Data and Methodology

We obtained the data from the DataSift provider, and we decided to collect data from September 24th to November 21st 2014 coming from Twitter English-speaker users. We chose to run our analysis for very large technology companies, mainly because of the great presence of tweets for these firms and because we had some priors on the effect of sentiment for this particular sector. We picked three large firms, i.e., Apple, Google and Facebook and, to avoid all the noise given by tweets written by individuals with no financial knowledge at all, we selected only the tweets that showed the companies' tickers, such as AAPL, GOOG and FB respectively.[3] Hence, we totally gathered almost 88,000 tweets for Apple, half of that for Facebook, and slightly less than 32,000 for Google.

In addition to standard information such as the text, the posting time and location, username, gender, and so on so forth, we were able to obtain some extra important information, such as the klout score - an influence scoring system - and the sentiment score. This scoring system allows a computer to consistently rate the positivity/negativity of a particular text. Within DataSift, the variation range allowed is typically between -20 and +20, although values outside this range do sometimes occur. Some basic statistics concerning these data are shown in the appendix. We show the breakdown of tweets per gender and date for each stock in Figure 8.10, and then the ratio per gender of positive and negative tweets per day for each stock in Figures 8.11 - 8.12. As it is possible to see from the pictures presented, men are usually more prone to post tweets and it does not seem possible to distinguish based on gender the positivity or negativity of the tweets, i.e., men do not exclusively post negative/positive tweets and neither do the women. Nevertheless, the proportions between women and men change between positive and negative tweets: on average, and even if it is quite variable from day to day, women seem to post more positive tweets on percentage with respect to men. Furthermore, it is possible to notice that there are some days with an extremely higher volume corresponding to important news about the firm, as for instance October, 20th for Apple (the expected debut of Apple Pay), October 28th-29th for Facebook

[3] These kind of data are easily collectable through the easy commands in CSDL coding environment - a specific code editor provided by DataSift for this purpose - that respectively are:

- twitter.symbols contains "AAPL" ;
- twitter.symbols in "GOOG, GOOGL" ;
- twitter.symbols contains "FB" .

(coinciding with the expected launch of Atlas platform and the third quarter in declining), and October, 16th for Google (for the news about medical services that it was planning to offer and for the third quarter closed missing earnings' expectations). Moreover, Figure 8.13 shows the tweets volume per day independently from the gender (the green bars), and the sentiment mean for all the daily tweets (the black lines). More specifically, the black lines are built in such a way that the lowest value represents the daily mean of the negative tweets, the highest value represents the daily mean of positive tweets, while the point are the daily overall mean and the red line represents the trend of the daily means. It looks intuitive how Facebook has the highest intraday variability in sentiment expressed through the Twitter channel, while for each stock the sentiment mean is pretty stable.

Instead, Figures 8.14 - 8.16 display the time series for prices and returns plotted against the sentiment trend. This has been done to have a sort of prior on the kind of correlation existing between the price/return and the general sentiment of the market. The returns seem to not be extremely correlated with the investors' sentiment, while on the other hand the price structure provides a better reflection of more the sentiment time series.

Finally, Figure 8.17 shows the daily average klout score for every stock. The klout score is an indicator of the degree of social influence for a person using a certain social media tool, and it may range between 1 and 100. The higher the value, the more influence a person has. It is very interesting to notice that in the period considered, on average, the influencers talk more about Google than the other two firms, and most of all that there is a strong decline in the last two weeks (7th - 21st of November).

Figure 8.18, on the other hand, concerns the volatility trends for every stock and for the relative sentiment score associated. As we can see, even if the stock are pretty volatile, the sentiment scores are instead quite stable.

For the sake of completeness, we should specify that the daily stock prices and volumes (the number of shares traded in a certain day) corresponding to that period were extracted from Bloomberg and eventually Yahoo! Finance has been used for counterchecking. The volatility selected from the provider was the volatility at 30 days. We then computed the returns directly from the prices with the standard log-formula:

$$r_t = ln\left(\frac{P_t}{P_{t-1}}\right) \tag{7.14}$$

where $P_t$ is the closing price for the day $t$.

We then constructed the following indicators to be tested, someone also similar to what proposed in Oliveira ([35]):

- *Sentiment Mean* (SM): the simple mean of the sentiment score per day;
- *Sentiment Ratio* (SR): the ratio between the Sentiment Mean at $t$ and $t-1$;
- *Bull-Bear Sentiment* (BBS) positive/negative: the daily Sentiment Mean only for positive/negative tweets;
- *Bull-Bear Sentiment Ratio* (BBSR): the ratio between the Bull-Bear Sentiment for positive and for negative tweets;
- *Bull-Bear Sentiment Percentage* (BBSper):

$$BBSper = \frac{BBSp}{(BBSp + BBSn)}; \tag{7.15}$$

- *Bull-Bear Sentiment Percentage Ratio* (BBSperR): the ratio between the Bull-Bear Sentiment percentage at $t$ and $t-1$;
- *Twitter Volume* (TV): the volume of tweets at a particular day $t$;
- *Volume Ratio* (VR): the ratio between the Twitter Volume at $t$ and $t-1$;

- *Bull-Bear Volume* (BBV) positive/negative: the daily Sentiment Volume only for positive/negative tweets;
- *Bull-Bear Volume Ratio* (BBVR): the ratio between the Bull-Bear Volume for positive and for negative tweets;
- *Twitter Volume 5-days Moving Average* (TVMA):

$$TVMA_t = \frac{1}{5} \sum_{i=t-4}^{t} TV_i; \tag{7.16}$$

- *Klout Score*: this is an indicator of the influence a certain person has in a social network. We computed the average of the score per day, so that the higher the score is, more influencers are expressing their opinions on the stock considered;
- *Sentiment Volatility* (SV): it has been computed as the standard deviation of the sentiment score;
- *Volatility Ratio* (VolR): it has been constructed as the ratio between the Sentiment Volatility and the 30-days stock volatility.

Instead of testing several models as in [35], we preferred to implement a variable selection model that would indicate to us which variable has to be included in the regression and which one should not. We decided to use a stepwise regression model, and more particularly a backward stepwise model. This case estimates the full model with all the explanatory variables and, if the least-significant term is statistically insignificant, it removes that variable and reestimates the model (otherwise it stops). The process is then reiterated. Furthermore, if the most-significant excluded term is statistically significant, it adds that variable and reestimates the model (otherwise it stops). The process is thus alternatively choosing the least significant variable to drop and then reconsidering all the variables dropped to be reintroduced in the model. This allows to retain only what matters to our model. We picked a significance level of 0.1 to be removed and 0.05 to be added back to the model. Moreover, for every model considered, the regression used in the stepwise process is a simple OLS regression, and in the case of the binary dependent variable, a linear probability model (LPM).

Another difference with respect to the literature ([35]) is that we do not believe in a perfect predictive power of online posting regarding prices or returns, but we consider microblogging as containing some relevant information that could be used in order to forecast the *direction* of the market and not the exact magnitude. This is the reason why we adopted, in addition to a more classic approach to be used as benchmark, an innovative approach which estimates as a dependent variable a binary variable, where 1 corresponds to an *up*-movement with respect to the price of the day before, while 0 means a *down*-movement.

Hence, to wrap up, the regression models used are of the form

$$y_t = \boldsymbol{x}_t \beta + \epsilon_t \tag{7.17}$$

or

$$y_t^* = \boldsymbol{x}_t \beta + \epsilon_t \tag{7.18}$$

where $y_t^*$ is a latent variable observable only in terms of its sign

$$y_t^* = \begin{cases} 0, & \text{if } (p_t - p_{t-1}) \leq 0 \\ 1, & \text{if } (p_t - p_{t-1}) > 0 \end{cases} \tag{7.19}$$

We explore both the case in which the dependent variable are the prices (in absolute value or as binary variable), and the returns as well.

Furthermore, we also run a rolling window analysis to test the robustness of our results. It would help us to assess the model's stability over time, since the key assumption of the parameters to be constant over time seems intuitive but needs to be verified. We thus decided to use a window of ten days to backtest our models and to evaluate their predictive accuracy.

### 7.3.3 Results

As it is possible to see from Tables 8.14 - 8.15, the variables selection gave different results for each stock considered, both for prices and returns. We included in all the tables every stock, and for each firm we computed the value forecasting (*Price*) and the direction forecasting (*Trend* or *Direction*, i.e. up or down for prices and returns respectively). We are going to focus more on the signs of the results obtained rather than the exact number, since as mentioned above, we strongly believe that social media data could help in capturing more of the direction of future movements than the accurate magnitude. We can thus summarize the models obtained in the following way:

- **Apple:**

  1. $P_t = \beta_1 P_{t-1} + \beta_2 BBSR_{t-1} + \beta_3 BBSp_{t-1} + \beta_4 VR_{t-1} + \beta_5 SV_{t-1} + \beta_6 BBVR_{t-1} + \beta_7 VolR_{t-1}$

     This model is exactly what we expected at the beginning: the price is thus a function of the price of yesterday, of some of the sentiment indicators and finally of the Twitter volume. While the posting volume has a positive impact on the price, even if with a lower magnitude, the sentiment indicators seem to be contrarian, since their sign is negative.

  2. $Trend_t = \beta_1 Trend_{t-1} + \beta_2 SM_{t-1} + \beta_3 SR_{t-1} + \beta_4 VolR_{t-1}$

     What this model represents is whether the price is going to be up tomorrow is related to what the market has done today (and this is why we liked to call the variable here *Trend*). It also depends on the sentiment mean, the sentiment ratio and finally on the volatility ratio. Hence, the higher the volatility within the sentiment score, the higher the chances that the price is going to increase tomorrow.

  3. $R_t = \beta_1 R_{t-1} + \beta_2 SM_{t-1} + \beta_3 SR_{t-1} + \beta_4 VR_{t-1} + \beta_5 BBSperR_{t-1} + \beta_6 BBSn_{t-1} + \beta_7 BBSper_{t-1}$

     The returns are also different from the previous model, since in addition to the "standard" variable of sentiment and volume already found before, they are also impacted by the the changing of the sentiment through time (i.e. BBSper, etc.).

  4. $Direction_t = \beta_1 Direction_{t-1} + \beta_2 SV_{t-1}$

     The returns trend is only a function of what happened yesterday, and of the volatility of the sentiment score through time.

- **Facebook:**

  1. $P_t = \beta_1 P_{t-1} + \beta_2 SV_{t-1} + \beta_3 TV_{t-1} + \beta_4 BBSperR_{t-1} + \beta_5 BBSper_{t-1} + \beta_6 SR_{t-1} + \beta_7 BBSn_{t-1}$

     Once again, it is interesting to see that the price does not only depend on the price of yesterday, but also on the posting volume, and most of all on the sentiment changes in time and on its volatility.

  2. $Trend_t = \beta_1 Trend_{t-1} + \beta_2 TV_{t-1} + \beta_3 BBSR_{t-1} + \beta_4 BBSp_{t-1} + \beta_5 BBSn_{t-1}$

     All the variables here are contrarian indicators, since they all have a negative sign. On the other hand, the magnitude is quite high.

  3. $R_t = \beta_1 SR_{t-1} + \beta_2 BBSperR_{t-1} + \beta_3 BBSper_{t-1} + \beta_4 BBSn_{t-1}$

     Oddly, the returns here seem to be predicted only by the evolution of the sentiment through time. It is strange since both in the literature and in the industry there is consensus within the

financial community regarding that the price/return of yesterday influences (or it is at least a good proxy) of the price/return of tomorrow.

4. $Direction_t = \beta_1 Direction_{t-1} + \beta_2 BBSperR_{t-1} + \beta_3 SM_{t-1} + \beta_4 SR_{t-1} + \beta_5 SV_{t-1} + \beta_6 TVMA_{t-1}$

   This model for returns forecasting shows that the direction of the Facebook returns is predicted well enough from sentiment score indicators. There is indeed only one variable that deals with the posting volume, and it is the Time-Varying Moving Average, and it has an extremely small impact of the forecasting.

- **Google:**

1. $P_t = \beta_1 P_{t-1} + \beta_2 TV_{t-1} + \beta_3 BBSR_{t-1} + \beta_4 BBSp_{t-1} + \beta_5 BBSn_{t-1}$

   The price here depends positively on yesterday's price, positively on the posting volume, and negatively on the changes in sentiment through time. It can also be observed that, even if statistically significants, the values for the *bull-bear sentiment* indicators are in this case not economically meaningful, so that extra data and a deeper analysis of corporate news or other items may be required to light up the issue.

2. $Trend_t = \beta_1 Trend_{t-1} + \beta_2 TVMA_{t-1} + \beta_3 BBSperR_{t-1} + \beta_4 BBSR_{t-1} + \beta_5 BBSp_{t-1} + \beta_6 VolR_{t-1} + \beta_7 TV_{t-1} + \beta_8 SR_{t-1}$

   The price trend is again correlated with the evolution of sentiment and volume in time, but once again the changes in the magnitude of positive or negative tweets play their role. The volatility ratio has a relevant effect on the forecasting, while the moving average seems to be significant but not so relevant in term of magnitude.

3. $R_t = \beta_1 R_{t-1} + \beta_2 SV_{t-1} + \beta_3 SM_{t-1} + \beta_4 BBVR_{t-1} + \beta_5 TV_{t-1} + \beta_6 TVMA_{t-1} + \beta_7 BBSp_{t-1}$

   Also here there is, even if it is very weakened, an explanatory power contained in both the posting volume and the moving average, while the sentiment score effect seems to be more relevant.

4. $Direction_t = \beta_1 Direction_{t-1} + \beta_2 VR_{t-1} + \beta_3 Klout_{t-1} + \beta_4 TV_{t-1} + \beta_5 BBSp_{t-1} + \beta_6 BBSper_{t-1} + \beta_7 TVMA_{t-1}$

   The returns trend is here given by the trend of yesterday, a strong component of the posting volume, the changes of the (positive) sentiment in time, and by the Klout score. In other words, it is not only important how much people talk about a certain stock, and which is their general sentiment about that, but also what they think today with respect to what they thought yesterday, and how many influencers talked about that stock as well.

There are some important aspects to notice to be similar across all the models: first of all, the autoregressive part of our models shows that the price lags are always positively correlated with the current prices, while the opposite is true for the returns. Furthermore, every time we use the binary dependent variable, we observe that their lags are negatively correlated with the current binary variable itself. This may be a clear sign of the mean reversion and the high daily fluctuations which characterise the stock market in general, and the technological sector more specifically. Secondly, the sentiment per se has a greater importance than the posting volume. This means that it is not true that "bad advertising does not exist", so stocks do not benefit from people talking about them if they say bad things. In the stock market, the stock has to create a positive sentiment and an inner trust, and it has not only to be on everyone's lips. The last interesting feature observable is that, in every model, the dependent variable is explained by at least one variable of sentiment and at least one variable of Twitter posting volume, even

if they may vary across model. This indicates unequivocally that the social networks sentiment is a road to be followed, because there is an intrinsic extra value for our model in using some variable that takes into account what people think or believe about a certain stock.

Finally, as it can be observed in the Figures 8.19 - 8.20, the betas for each model behave in a complete different way one from the others. Indeed, the betas of the price forecasting models seem to be pretty volatile and non-stable, except for Google in the long-run. The same, but maybe also more amplified, is true for the returns forecasting models. Hence, since the variables included differ completely from a stock to the others, and given the high variability of the betas' behaviour, it was not possible to identify a general model that took into account the social network sentiment as a proxy and indicator for the stock market movements.

### 7.3.4 Conclusions

In this work, microblogging data have collected and analyzed for their applications in the stock market. We actually built several indicators, and we used them in order to perform an efficient forecasting model, both for stock prices and returns. We decided to analyze three major technology companies, i.e., Apple, Facebook and Google, for a two-months period (Sep. - Nov. 2014). The results are promising, and it seems that, regardless of the model used or any company-specific factor, both the posting volume and the average sentiment expressed through social networks have definitely a value and a predictive power in the stock market. We were able to estimate different models for each stock considered, and we tested the robustness of our analysis through ten-days rolling windows. It is possible to conclude in favour of the importance and relevance of these innovative features for improved predictive models, although the variety of the specifications obtained and the non-constancy over time of the parameters makes difficult to create a unique model that works independently of the stock selected. The creation of single ad-hoc forecasting model is interesting for any practitioner indeed, but from the evidence presented it seems complicated to generalise the findings, as it is for instance for well-known models as the Fama-French one. Furthermore, the outcomes achieved give us an insight on many other different further analyses that can be implemented using the microblogging data, in particular the ones coming from Twitter. Of course, the first extension may be studying the phenomenon with longer time series and/or a different time frequency. In addition, in the wake of Michaely et al. ([30]) or De Bondt and Thaler ([19], [20]), an interesting field that can be tackled is the sentiment analysis regarding dividend initiations or omissions, or in general how this source of information can be used to assess the impact of company-specific changes and/or events. Furthermore, the socials could also be very helpful to exploit situations like the IPOs, in which a natural and usual underpricing is always present ([8]). Another potential field to investigate concerns how brand perception impacts the propensity of the investors to, for instance, hold a certain stock ([24]). Finally, it could analysed what kind of impact (if any) could be attributed to the gender.

## 7.4 Can Twitter proxy the investors' sentiment? The case for the technology sector

### Abstract

The stock market is influenced by several factors, such as macroeconomics, regulatory, purely speculative ones, and many others. However, one of the most relevant and meaningful is the general opinion and the overall investors' sentiment, i.e., what investors think about a certain firm and, as a consequence, about the relative stock. This investors' sentiment is here proxied by the Twitter content, and the study sums up to the recent outbreak of works that exploits sentiment analysis and Twitter data for stock market predictions. The sample analyzed concerns three major technology companies over a two-months period, on an minute basis. Using microblogging activities and a scoring algorithm for each tweet, it was possible to formulate interesting forecasting models identifying a new set of variables and indicators of the stock market future movements. A selection model has been used to implement the study, and the evidences found were encouraging, since it has been possible to draw the conclusion that this new source of data may increase the explanatory power of financial forecasting models. More in details, it looks like that the average sentiment associated to any tweet is not so relevant as expected in prediction terms, while the posting volume has a greater forecasting power and it could be used to augment the models.

### 7.4.1 Introduction

Big Data is becoming nowadays a buzzword used in several contexts and many different ways. Even if it presents controversial and still ambiguous definitions, it is generally identified as a common feature to all of these definitions the presence of a huge variety of high-speed unstructured data. Hence, probably one of the best examples of big data applications is the use of social media and web contents data, that is generally known as sentiment analysis (Agarwal et al., 2011). A manifold spectrum of utilizations of this new source of data has been studied over the last few years: in medical and epidemics contexts for instance (Culotta, 2010), or to try to predict the presidential elections (Tumasjan et al., 2010).

Although medical or political applications have been deeply explored, one of the most prolific fields of research concerned the use of social media for business and financial purposes. So, no matter whether it dealt with movie revenues (Mishne and Glance, 2006), commercial sales (Choi and Varian, 2012), or music albums forecasts (Dhar and Chang, 2009) from one hand, or with different social networks sources, such as blogging activities (De Choudhury et al., 2008), stock messages board (Antweiler and Frank, 2004; Koski et al., 2008), or web search queries (Bordino et al., 2012) from the other hand, the importance of this new available dataset has grown and it is currently used for trying to predict the future (Asur and Huberman, 2010).

Nevertheless business and finance in general were under a "social" attack in the last five years and a lot of different works have been implemented (e.g., Ruiz et al., 2012; Mao et al., 2015), a subset of them - the ones that regard the stock markets - have been particularly analyzed. The main instrument was the data coming from Twitter, and it has been extensively preferred to other sources, such as for instance analysts' recommendations (Barber et al., 2001), or financial news (Lavrenko et al., 2000; Schumaker and Chen, 2009), because of the tweets standard length, common language and symbolism, and high availability and variety.

Thus, Bollen et al. in a first place (2011), and then others in following works (Bollen and Mao, 2011; Mao et al., 2011; Mittal and Goel, 2012), used financial tweets and their associated investors mood in order to predict the Dow Jones Industrial Average Index. Corea (2015) and Corea and Cervellati (2015) instead used Twitter data about major technologies companies to predict the Nasdaq-100 movements, while Brown (2012) investigated how Twitter user's reputation could affect the stock market, and Oliveira et al. (2013) found a positive correlation between the tweets posting volume and the stock market variations.

Although the use of social media data in order to anticipate the stock markets oscillations is quite new, the idea of exploiting the investor sentiment and financial news to gain a competitive advantage is

well established in literature (Da et al., 2012; 2015). It has been showed that financial news with negative words (Tetlock, 2007; Tetlock et al., 2008), or investor sentiment (Baker and Wurgler, 2006; 2007) have a certain degree of prediction power for the stock markets, as well as it is for tactical allocation (Fisher and Statman, 2000). Finally, the gap between traditional finance view on the topic and sentiment analysis has been filled by Oh and Sheng (2011), Sprenger et al., (2010), as well as many others above-mentioned.

Hence, the purpose of this study is to sum up to the existing literature providing new insights and methods for sentiment analysis forecasting. Using data from three major technology companies over a two-months period, it will be provided a single high-frequency price-forecasting model for each of them, as well as a trend one, i.e., whether the prices are experiencing a bullishness or bearishness second by second. The work is then structured as follows: section 7.4.2 will deal with the data collection, variables creation, and methodology used, while section 7.4.3 will show some results from the analysis implemented. Section 7.4.4 will finally draw some conclusions, suggesting further future improvements for the field of study.

### 7.4.2 Methodologies and Dataset Construction

The data used in the study have been obtained through two different sources: the Twitter one comes from a data provider named DataSift, while the prices for the three stocks have been extracted by Bloomberg. The time period considered spanned over two months from September 24th to November 21st 2014, and only the English tweets regarding Apple, Facebook, and Google have been collected. Other languages represented a minority of tweets and were out of the scope of this analysis, and so there were not considered, while concerning the choice of the companies to analyze, the decision has been driven from two factors: the high presence of tweets on the selected companies, and the existing studies who proved that sentiment analysis works in the technology sector (Corea, 2015; Corea and Cervellati, 2015). As the frequency considered, the data were analyzed on a minute basis.

All the noise coming from meaningfulness tweets or information has been depurated taken into account only the tweets posted by individuals with some degree of financial literacy. This has been obtained considering only the tweets that showed the company ticker, where the presence of the ticker is meant to be a good proxy of individuals' financial knowledge. Hence, overall almost 88,000 thousands of tweets has been gathered for the Apple stock, about 44,000 for Facebook, and less than 32,000 concerning Google.

The Figures 7.6 - 7.8 illustrate the amount of tweets per minute relatively to each single stock. This gives an idea of the intensity of the microblogging phenomenon, and it could be used in future studies to deepen sentiment analysis with respect to specific tweets-intensive minutes (e.g., reaction to announcements). Furthermore, from the figures can be inferred that there are neither intraday patterns nor seasonality that might bias the results. The pictures also exclude any intuitive correlation in posting activities between stocks so similar. In the period considered, it seems indeed that no event affected all the stocks at the same time and with the same magnitude. In addition, the contagion effect that usually characterizes stock belonging to the same sector or geographic area seems to be missing here.

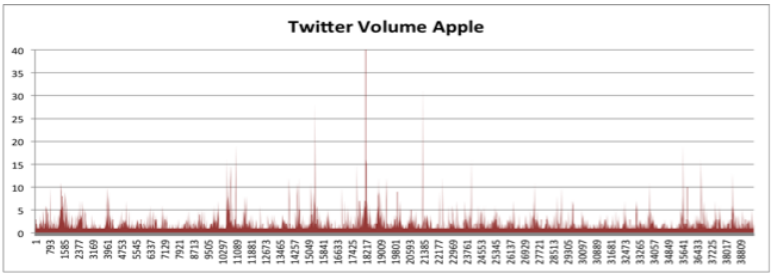



Figure 7.6: Amount of tweets posted for each minute about Apple stock.

Once the tweets have been extracted, their sentiment was assessed (by the data provider) through an algorithm that scored them with a value ranging from -20 to +20, depending on the strongly negativity

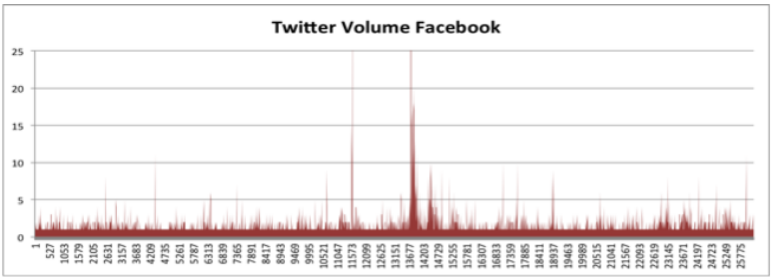



Figure 7.7: Amount of tweets posted for each minute about Facebook stock.

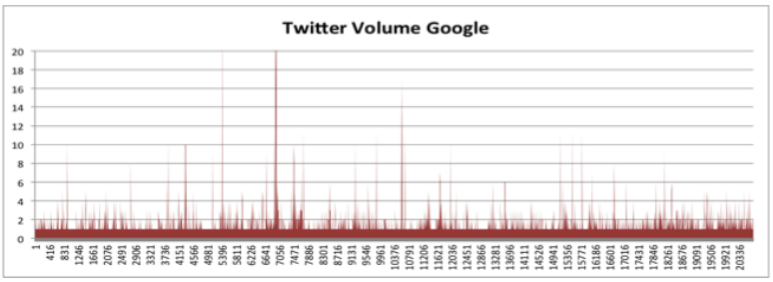



Figure 7.8: Amount of tweets posted for each minute about Google stock.

or positivity of each tweet content. A second different score - the klout score - has also been included in the dataset. This is a value that indicates the degree of social influence of certain individual in the social media world, and it varies between 1 and 100 - to a higher value corresponds a higher influence power.

In order to analyze not only the relations with the prices but also those ones with the trend, a set of different variable has been constructed, similarly to what previously observed in Oliveira et al. (2013):

- Sentiment Mean (SM): the simple mean of the sentiment score per minute;
- Sentiment Ratio (SR): the ratio between the Sentiment Mean at t and t-1;
- Bull-Bear Sentiment (BBS) positive/negative: the Sentiment Mean per minute only for positive/negative tweets;
- Bull-Bear Sentiment Ratio (BBSR): the ratio between the Bull-Bear Sentiment for positive and for negative tweets;
- Twitter Volume (TV): the volume of tweets at a particular minute t;
- Bull-Bear Volume (BBV) positive/negative: the Sentiment Volume per minute only for positive/negative tweets;
- Bull-Bear Volume Ratio (BBVR): the ratio between the Bull-Bear Volume for positive and for negative tweets;
- Twitter Volume 5-minutes Moving Average (TVMA):

$$TVMA_t = \frac{1}{5} \sum_{i=t-4}^{t} TV_i; \tag{7.20}$$

- Twitter Sentiment 5-minutes Moving Average (SMMA):

$$SMMA_t = \frac{1}{5} \sum_{i=t-4}^{t} SM_i; \tag{7.21}$$

- Klout Score: it has been computed the average of the score per day.

The regression models used in order to understand the relations between the prices and the tweets sentiment are respectively an ordinary least square (OLS) regression and a linear probability model (LPM):

$$y_t = \boldsymbol{x}_t\beta + \epsilon_t \tag{7.22}$$

and

$$y_t^* = \boldsymbol{x}_t\beta + \epsilon_t \tag{7.23}$$

where $y_t^*$ is a latent variable observable only in terms of its sign. In other words:

$$y_t^* = \begin{cases} 0, & \text{if } (p_t - p_{t-1}) \leq 0 \\ 1, & \text{if } (p_t - p_{t-1}) > 0 \end{cases} \tag{7.24}$$

This is one of the differences with respect to the literature so far mentioned: the work studies both the impact of the sentiment on the simple stock price but also on the directional trend that the stock is experiencing, that is whether it is growing or decreasing over the following minute. As it has been noticed, the dummy variable indeed assumes value 1 whether the prices are up-moving, while 0 if they are down-moving.

Furthermore, instead of selecting by hand which of those variables to be included in the model or testing different models, it has been decided to use a selection model that automatically inserts or excludes a certain variable on the base of a threshold significance level. In this case, the value for a variable to be part of the model is 0.05, while 0.1 for being removed. There are different types of stepwise regression model, and here the backward version has been implemented. The backward stepwise regression assumes to estimate the full model with all the explanatory variables in a first place. Then, if the least-significant term is statistically insignificant, it removes that variable and reestimates the model (otherwise it stops). The process is then reiterated. At the same time, for each step, if the most-significant excluded term is statistically significant, it adds that variable back and reestimates the model (otherwise it stops). The algorithm is thus alternatively choosing the least significant variable to drop and to be reintroduced in the model. It is a particular smart and convenient way to select the statistical meaningful variables on the base of pre-fixed significance threshold values without having to deal with each one by hand.

A difference with respect to some previous works is the choice of not taking into account any corporate information at this stage (Kearney and Liu, 2014), and only considering the information coming directly from Twitter, as well as the stock price.

### 7.4.3 Empirical Analysis and Discussion

Hence, two regressions have been run for each company's stock, one for the price - OLS - and one for trend - LPM. The results are shown in the Table 8.16.

As it can be observed from the Table 8.16, the results are quite mixed but encouraging, meaning that there is a single feature/variable that is present in every regression. Nonetheless, some consistency can be noted. The most interesting thing is that the simple sentiment mean has been excluded from each regression, and in general the sentiment, whether it is positive or negative, it is not so relevant every time and for every stock. In general, it is true that the variables where the sentiment was considered in any form have more predictive power in term of stock trend than stock point-forecasting. On the other hand though, the tweets volume seems to have a strong impact both in terms of price forecasting and directional prediction. This suggests that maybe is more valuable how much people talk about a certain stock with respect to what they actually think about it. To confirm this hypothesis, it can be observed that negative sentiments have a negative impact on the stock price - as intuitively should be - while an increase in the posting volume of negative tweets has anyway a positive impact on the stock price.

Moreover, variables that capture the sentiment mean have on average a higher magnitude with respect to the posting volume ones.

A second interesting consideration is that the Klout score is significant in more than one case. Hence, it seems that individuals with a higher influence power within the social media worlds can effectively influence the stock direction with their posts and opinions, although the magnitude is extremely low.

Finally, contrarily to Tetlock (2007) and Tetlock et al. (2008), negative sentiment causes a downward pressure of a lower magnitude than the upward push of a positive tweet.

### 7.4.4 Conclusions

New sources of data are daily used for trying to capture the stock market behavior. One of the currently most used and innovative is without a doubt the social media data. It has been analyzed here how Twitter in particular could be used in financial contexts. Different variables embedding tweets content sentiment and volume as well have been created and used for price and trend forecasting. Three major technology companies have been studied for a two months period. Consistently with previous studies, as it can be inferred from the extensive survey proposed by Kearney and Liu (2014), linear regressions perform often far better than more complicated regressions, and are then used also for the sake of this study. In spite of that, no previous work use selection models such as the stepwise regression, which seems to optimize the variable selection process in the present analysis.

The results provided gave an overview on the kind of insights that can be achieved through microblogging and social media more in general. The results are also quite mixed, probably reflecting structural and specific intrinsic differences for each company, but at the same time they show some degree of consistency and comparability.

Further implementations could be studied in the next future, such as considering longer timeframes, different companies and sectors, or analyzing special situations such as IPO and company's announcements (dividends, etc.). Of particular interest would also be studying the structure of the network who is talking about a certain stock or firm and assess how this affect the company's evaluation on the stock market.

## 7.5 Emotional Speculative Behaviour in the Option Market

### Abstract

Social media data have been proved to be effective in augmenting stock price forecasting models before (Bollen et al., 2011; Corea, 2015), but given the intrinsic speculative nature of traders who may use these innovative datasets, it appears more reasonable to investigate the relation between the Twitter data and the stock option prices. The underlying hypothesis is indeed that speculative trading strategies - as the ones based on social media inference are - may be more effective if evaluated on speculative instruments instead of simple stock prices. Consistently with previous works, it has been then studied for three major technology stocks over a two-months period the relation between investors' sentiment and basic financial products on an intraday basis. A set of different variables has been created to include different interactions between sentiment and option prices, and a statistical selection model has been put in charge of identifying the most relevant correlations. The results are quite mixed: social media data seem to be indeed useful for predicting some option prices but no others, and in particular are able to better explain single companies' option prices oscillations rather than the ones related to general indexes such as the Nasdaq-100.

### 7.5.1 Introduction

The amount of data produced nowadays is increasing exponentially, and as everyone already knows more than 90% of data available today have been generated in the last two-four years (SINTEF, 2013). Even if a great portion of these new data comes from the Internet of Things, from sensors, and from mobile applications, another relevant quantity is generated by social networks and through web contents. The analysis of these data (Agarwal et al., 2011) may have different impact in different fields: it could indeed help in predicting the presidential elections' outcome (Tumasjan et al., 2010), understanding a disease spread (Culotta, 2010), or assessing the success of a new music album release (Dhar and Chang, 2009). No matter the field of applications though, social media data seem to be fundamental to predict the future (Asur and Huberman, 2010).

The study of the investors' sentiment is deeply rooted in the financial literature (Baker and Wurgler, 2006; 2007), but the use of new sources of information has given traders and institutional investors a new important way to gain a competitive advantage (Da et al., 2012; 2015). It seems to be clear that Twitter represents one of the most important sources of social media data, given the standard format of the posting activities - every tweet is limited to 140 characters - and for the fast diffusion it had in the last few years, although many other informative channels have been exploited as well. Hence, it makes sense to split the preexisting works into three branches, based on the use of solo Twitter for financial markets applications, on the use of distinct social media data, or for a different final business purpose. In the first group, pioneering works have to be attributed to Bollen and Mao (2011), who started a flow of research that has been then adopted and enhanced by others (Mittal and Goel, 2012). They focused on interpreting and deducing human emotions from microblogging activity, in order to provide insights on the movements of the Dow Jones Industrial Average Index (Bollen et al., 2011; Mao et al., 2011). Similarly, Oliveira et al. (2013) discovered a positive correlation between the stock volume and the Twitter volume, while Corea (2015) and Corea and Cervellati (2015) selected three major technology stocks and built an indicator for predicting the variations in the Nasdaq-100. Brown (2012) instead enquired the importance of the user's reputation as a driver for stock market changes rather that focusing on the tweets content. Ruiz et al. (2012) concentrated their effort in explaining hidden correlations between microblogging activity and financial time series, while Oh and Sheng (2011) and Sprenger et al. (2010) put their emphasis in the microblogging informative power. Finally, Mao et al. (2015) readapted the previous approach to analyze international financial markets mixing both Twitter and Google data.

In the second branch, i.e., the use of social media data different from Twitter, Lavrenko et al. (2000) few decades ago - and Schumaker and Chen later on (2009) - analyzed how financial news release impacts on the financial markets. Tetlock (2007; Tetlock et al., 2008) dug more into this field focusing particularly on the effect of negative news, and Barber et al. (2001) showed instead how analysts' recommendations may be exploited for extrapolating future trends. Other meaningful progresses in the field have been done by Antweiler and Frank (2004), and Koski et al. (2008), that used stock messages board as primary

source of information, while web search (Bordino et al., 2012) and blogs (De Choudhury et al., 2008) have been alternatively considered for formulating stock market predictions.

The last group - the one that concerns different business applications for sentiment analysis - started with Fisher and Statman (2000), in a work in which they used investors' sentiment for asset tactical allocation, and then varies from movie revenues forecasting (Mishne and Glance, 2006) to commercial sales prediction (Choi and Varian, 2012).

Hence, the literature is quite vast, and in order to give a contribution, the aim of this work is to analyze both for single stocks and stock index (i.e., the Nasdaq-100), how option prices movements could be explained by changes into the tweets' sentiment. The analysis will be performed on an intraday basis, because on a daily one has been proved to be ineffective (Corea, 2015), and the paper will have the structured as follows: section 2 will take care of the data gathering and the methodology used. In section 3, there are going to be showed some achievements of the models proposed, while section 4 will finally draw the conclusions, providing suggestions for future researches and some insights from the current study.

### 7.5.2 Methodologies and Dataset Construction

For consistency and comparability with previous works (Bollen et al., 2011; Corea, 2015), the same stocks have been analyzed, i.e., Google, Apple, and Facebook, as well as the Nasdaq-100 (Corea and Cervellati, 2015). Contrarily to what previously done though, the following analysis has been focused on the study of the relationship between the social expressions and the option prices, and in particular on the impact of the former one on the latter. Hence, the stock prices for the three stocks and for the Nasdaq-100 have been extracted from Bloomberg, and the option prices have been derived through the following Black and Scholes set of equations (the notation is the standard used):

$$Call = SN(d_1) - K^{-rT}N(d_2) \tag{7.25}$$

$$Put = K^{-rT}N(-d_2) - SN(-d_1) \tag{7.26}$$

$$d_1 = \frac{ln(\frac{S}{K}) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} \tag{7.27}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{7.28}$$

where the 3-months interest rate has been obtained from the Federal Reserve website, and the options have consistently a 3-months expiration time.

On the other side, the tweets, with relative sentiment and klout scores, have been acquired from DataSift. The sentiment score measures the level of positivity (or negativity) of the human opinion explicated in the text, while the klout score canalizes in a single value the degree of social influence of an individual. It varies between 1 and 100, and to a higher value corresponds a higher influence power. The time period considered for the analysis goes from September 24th to November 21st 2014, and it was possible to collect the data second by second, and eventually aggregate them on a minute basis.

On the content of the messages, two relevant choices have been taken for the sake of the study: first of all, only English tweets have been pondered - because they represent the almost totality of the tweets universe - and secondly, in order to reduce the high volatility deriving from unrelated or misunderstood tweets, it has been decided to include only the messages strictly connected to the stock valuation. In other words, the financial literacy of the bloggers has been proxied through the selection of the tweets where the companies' ticker was mentioned. Totally, almost 88,000 thousands of tweets has been grouped for the Apple stock, about 44,000 for Facebook, and less than 32,000 for Google.

Regarding instead one of the key variable - the tweet sentiment - a scoring algorithm assigned a value ranging from -20 to +20, respectively to extremely negative or positive tweets.

In addition, a group of other variables has been ideated to take into account variations and nuances of the sentiment score above-mentioned. Thus, in a similar fashion as in Oliveira et al. (2013), it has been defined a simple sentiment mean (SM), and its ratio with the one-lagged value (SR). Furthermore, a set of indicators aimed to capture the bullishness (or bearishness) of the market has been suggested: the simple mean has been computed first individually for positive (BBSp) and negative tweets (BBSn), and then a ratio between the two has been proposed (BBSR). Finally, it has been added the klout score, and a 5-moving averages for the sentiment (SMMA).

It has been then used a simple ordinary least squares (OLS) regression model and a linear probability model (LPM) to assess the type of impact the general sentiment had on the option prices:

$$y_t = \boldsymbol{x}_t \beta + \epsilon_t \tag{7.29}$$

or

$$y_t^* = \boldsymbol{x}_t \beta + \epsilon_t \tag{7.30}$$

where $y_t^*$ is a latent variable observable only in terms of its sign, i.e.,

$$y_t^* = \begin{cases} 0, & \text{if } (p_t - p_{t-1}) \leq 0 \\ 1, & \text{if } (p_t - p_{t-1}) > 0 \end{cases} \tag{7.31}$$

According to Corea (2015), and Corea and Cervellati (2015), this variable has been called Trend, and it has been constructed as a dummy variable with a value of 1 that indicates an up-movement, while 0 a down-movement.

Furthermore, instead of selecting by hand which of those variables to be included in the model or testing different models, it has been decided to use a selection model that automatically inserts or excludes a certain variable on the base of a threshold significance level. In this case, the value for a variable to be part of the model is 0.05, while 0.1 for being removed. There are different types of stepwise regression model, and here the backward version has been implemented. The backward stepwise regression assumes to estimate the full model with all the explanatory variables in a first place. Then, if the least-significant term is statistically insignificant, it removes that variable and reestimates the model (otherwise it stops). The process is then reiterated. At the same time, for each step, if the most-significant excluded term is statistically significant, it adds that variable back and reestimates the model (otherwise it stops). The algorithm is thus alternatively choosing the least significant variable to drop and to be reintroduced in the model. It is a particular smart and convenient way to select the statistical meaningful variables on the base of pre-fixed significance threshold values without having to deal with each one by hand.

Concerning the Nasdaq estimation instead, another set of indicators has been integrated, with the aim of replicating synthetically the index - as already proposed in Corea (2015). The main instruments embedded in the analysis have been therefore obtained as the simple average of the three stocks' sentiment (SIT) and the weighted variation for their respective tweets volumes (SITw). The two relative moving-average versions have been also incorporated (SITma and SITwma). A different procedure has also been implemented. In order to predict the Nasdaq-100 option oscillations, three of the major technology companies of the index itself (i.e., Google, Apple, and Facebook) have been selected ex-ante because they are expected to have a stronger weight within the stocks bundle belonging to the Nasdaq index. Hence, it has been done a kind of qualitative principal component analysis, in order to take into account from the beginning only the stocks with a higher explanatory power for the index.

Afterwards, the following models have been tested:

$$M1 : P_t = \alpha + \phi_1 P_{t-1} + \epsilon_t \tag{7.32}$$

$$M2 : P_t = \alpha + \phi_1 P_{t-1} + \beta_1 SM_{Apple} + \beta_2 SM_{Facebook} + \beta_3 SM_{Google} + \epsilon_t \tag{7.33}$$

$$M3 : P_t = \alpha + \phi_1 P_{t-1} + \beta_1 SMMA_{Apple} + \beta_2 SMMA_{Facebook} + \beta_3 SMMA_{Google} + \epsilon_t \tag{7.34}$$

$$M4 : P_t = \alpha + \phi_1 P_{t-1} + \phi_2 SIT_{T-1} + \epsilon_t \tag{7.35}$$

and then the same has been studied for the weighted version, the moving average one, and finally the weighted moving average, respectively M5, M6, and M7.

### 7.5.3 Empirical Analysis and Discussion

As already previously explained, two regressions have been run for each company stock, one for the price (the OLS), and one for trend (the LPM). The Nasdaq is going to be considered first, and then the single companies' forecasts. However, only the results relative to the call options will be showed - for the sake of completeness, the analysis has been implemented also on put options, and the conclusions remain almost the same ones. The results have been though taken out from the current work because they did not add any further value neither generate new insights for the investigation.

The results for the Nasdaq data are then shown in the Tables 8.17 - 8.20. Tables 8.17 and 8.19 have the same meaning, and they show the results of the OLS and LPM regressions. Tables 8.18 and 8.20 show instead the adjusted $R^2$ and root mean squared error for all the models considered, in order to provide a fast way to assess whether the augmented models performed better and were more accurate with respect to the benchmark.

Differently with respect to previous results in the literature (Corea, 2015; Corea and Cervellati, 2015), it seems that the Twitter explanatory power is fairly low concerning the price estimation. Indeed, only the sentiment tracking index variable seems to be slightly statistically significant, and the benchmark autoregressive model performs better than any other more complex variations. The opposite is true though in the trending case, in which the microblogging proves once again to be relevant for increasing the forecasting ability of the statistical models.

The situation drastically changes when intraday data are instead considered for the single companies' option forecasting. Indeed, as it is provided in Table 8.21 the predictions are more complex and heterogeneous, and some of the forecasts are more accurate and complete for the directional models than with respect to the price estimations, while some others the other way round.

It can be noticed that, no matter the stock taken into account, the Klout score has a significant impact on the option price: influencing traders or investors who release their opinions on the web may actually affect the market's evolution. A second interesting fact is that simple indicator such as the sentiment mean has a low meaning for these forecasting models, while more refined variables (e.g., the bullishness-bearishness ratio) are more valuable to the analysis. In particular, it seems also that negative news influence more the option prices than positive ones.

### 7.5.4 Conclusions

A vast literature is exploring the implications of new data sources for different field applications, and relevant progresses have been done especially in financial markets. Twitter and social media data may represent a new frontier of quantitative financial modelling, and in this work it has been given a contribution to this area. Two months of tweets have been collected, with a specific focus on three big technology companies - Apple, Google, and Facebook - and used them for refining option pricing forecasting models. It has been tried first of all to predict the Nasdaq-100 variations, with poor results concerning the price forecasting and slightly more encouraging ones regarding the directional changes. Afterwards, single companies' options have been considered, and a set of indicators has been built in order

to augment simple autoregressive models. The achievements of the models in this case are relevant - even if on a modest scale, due mainly to length of the time series - and further works will investigate for sure longer time series, different and multiple stocks, and different sectors. If will prove to be consistent, these further adjustments would increase the model and techniques standardisation, making the analysis generally applicable and transferable to different environments and maybe additional speculative instruments.

# General References


[1] Aguilar, O., West, M. (2000). "Bayesian dynamic factor models and portfolio allocation". Journal of Business and Economic Statistics 18: 338-357.

[2] Ahmed, N. K., Atiya, A. F., El Gayar, N., El-Shishiny, H. (2010). "An Empirical Comparison of Machine Learning Models for Time Series Forecasting". Journal of Econometric Reviews 29 (5-6): 594-621.

[3] Bai, J., Wang, P. (2012). "Identification and estimation of dynamic factor models". MPRA Paper No. 38434.

[4] Baillie, R.T., Baltagi, B. H. (1999). "Prediction from the regression model with one-way error components". Chapter 10 in C. Hsiao, K. Lahiri, L.F. Lee and H. Pesaran. "Analysis of Panels and Limited Dependent Variable Models". Cambridge University Press, Cambridge: 255-267.

[5] Baltagi, B. H. (1988). "Prediction with a Two-Way Error Component Regression Model". Econometric Theory, Volume 4 (1): 171-181.

[6] Baltagi, B. H., Levin, D. (1992). "Cigarette taxation: Raising revenues and reducing consumption". Structural Change and Economic Dynamics, Volume 3 (2): 321-335.

[7] Baltagi, B. H. (2007). "Forecasting with Panel Data". Center for Policy Research Working Paper No. 91.

[8] Baltagi, B. H. (2008). "Econometric Analysis of Panel Data". John Wiley & Sons Inc.

[9] Baltagi, B. H. (2009). "Forecasting with Spatial Panel Data". IZA Discussion papers, No. 4242.

[10] Banbura, M., Giannone, D., and Reichlin, L. (2011). "Nowcasting". In M. P. Clements and D. F. Hendry, eds., Oxford Handbook of Economic Forecasting, chap. 7. Oxford University Press.

[11] Baum, C. F. (2013). "Panel data estimation and forecasting". Lecture at University of Mauritius, slides package.

[12] Beratung, J., Eickmeier, S. (2005). "Dynamic Factor Models". Deutsche Bundesbank Discussion Paper Series 1: Economic Studies, No. 38.

[13] Bishop, C. M. (1995). "Neural Networks for Pattern Recognition". Oxford University Press.

[14] Bontempi, G. (2008). "Long term time series prediction with multi- input multi-output local learning". In Proceedings of the 2nd European Symposium on Time Series Prediction: 145-154.

[15] Bontempi, G. (2013). "Machine Learning for Time Series Prediction". Machine Learning Summer School 2013 presentation, Hammamet: 1-128.

[16] Bontempi, G., Birattari, M., Bersini, H. (1999). "Local learning for iterated time-series prediction". In I. Bratko and S. Dze- roski, editors, Machine Learning: Proceedings of the Sixteenth International Conference: 32-38.

[17] Bontempi, G., Taieb, S. B. (2011). "Conditionally dependent strategies for multiple-step-ahead prediction in local learning". International Journal of Forecasting, 27(3): 689-699.

[18] Bontempi, G., Taieb, S. B., Le Borgne, Y. (2013). "Machine Learning Strategies for Time Series Forecasting". Lecture Notes in Business Information Processing Volume 138: 62-77.

[19] Bosson Brou, Kouassi, E., Kymn, K. O. (2011). "Double Autocorrelation in Two Way Error Component Models". Open Journal of Statistics, No. 1: 185-198.

[20] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L. (2014). "Inferring causal impact using Bayesian structural time-series models". Annals of Applied Statistics (in press).

[21] Castle, J., Qin, X., Reed, R. (2009). "How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms". Working Paper No. 13/2009, University of Canterbury.

[22] Castle, J., Doornik, J. A., Hendry, D. F. (2010). "Evaluating Automatic Model Selection". Discussion paper series, Department of Economics of University of Oxford.

[23] Cheng, H., Tan, P., Gao, J., Scripps, J. (2006). "Multistep-ahead time series prediction". In Pacific Asia Knowledge Discovery and Data Mining: 765-774.

[24] Choi, H., Varian, H. (2009). "Predicting the present with Google Trends". Technical Report. Google.

[25] Choi, H., Varian, H. (2012). "Predicting the present with Google Trends". Economic Record 88: 2-9.

[26] Clements, M., Henry. (2005). "A Companion to Economic Forecasting". Blackwell Publishers, Oxford.

[27] De Mol, C., Giannone, D., Reichlin, L. (2006). "Forecasting using a large number of predictors: is Bayesian regression a valid alternative to principal components?". Deutsche Bundesbank Discussion Paper Series 1: Economic Studies No 32.

[28] Deaton, A. (1985). "Panel Data from Time Series of Cross-Sections". Journal of Econometrics 30: 109-126.

[29] Diebold, F. X. (2004). "Elements of Forecasting". South-Western, Cincinnati.

[30] Durbin, J., Koopman, S. J. (2001). "Time Series Analysis by State Space Methods". Oxford University Press.

[31] Eklund, J., Kapetanios, G. (2008). "A Review of Forecasting Techniques for Large Data Sets". Queen Mary Department of Economics Working Paper No. 625: 1-16.

[32] Forni, M., Reichlin, L. (1998). "Let's get real: A factor analytical approach to disaggregated business cycle dynamics". Review of Economic Studies 65: 453-473.

[33] Forni, M., M., Hallin, M., Lippi, Reichlin, L. (2000). "The generalized dynamic factor model: Identification and estimation". Review of Economics and Statistics 82: 540-554.

[34] Frees, E.W., Miller, T.W. (2004). "Sales forecasting using longitudinal data models". International Journal of Forecasting 20, 99-114.

[35] Gencay, R., Qi, M. (2001). "Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping and bagging". IEEE Transactions on Neural Networks, 12: 726-734.

[36] Goldberger, A. S. (1962). "Best Linear unbiased prediction in the generalised linear regression model". Journal of the American Statistical Association 57: 369-375.

[37] Gourieroux, C., Jasiak, J. (2001). "Dynamic Factor Models". Econometric Reviews 20 (4): 385-424.

[38] Greene, W. H. (2011). "Econometric Analysis". Prentice Hall, 7th Edition.

[39] Grouber, M. (1998). "Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators".CRC Press.

[40] Guo, M., Bai, Z., An, H. Z. "Multi-step prediction for non-linear autoregressive models based on empirical distributions". Statistica Sinica: 559-570.

[41] Hamilton, J. D. (1994). "Time Series Analysis". Princeton University Press.

[42] Harvey, A. C. (1989). "Forecasting, structural time series models and the Kalman filter". Cambridge University Press.

[43] Harvey, A. C., Shephard, N. (1993). "Structural Time Series Models". Handbook of Statistics, Vol. 11: 261-302.

[44] Hastie, T., Tibshirani, R., Friedman, J. (2013). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer. Springer Texts in Statistics.

[45] Holtz-Eakin, D., Newey, W., Rosen, H. (1988). "Estimating Vector Autoregressions with Panel Data". Econometrica. Vol. 56, No. 6: 1371-1395.

[46] Hsiao, C. (2003). "Analysis of Panel Data". Econometric Society Monographs Book 34. Cambridge University Press.

[47] Ishwaran, H., Rao, S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies". The Annals of Statistics, Vol. 33, No. 2: 730-773.

[48] Issler, J. V., Lima, L. R. (2009). "A panel data approach to economic forecasting: The bias-corrected average forecast". Journal of Econometrics. Volume 152 (2): 153-164.

[49] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). "An Introduction to Statistical Learning with Applications in R". Springer. Springer Texts in Statistics (Book 103).

[50] Kalman, R.E. (1960). "A new approach lo linear filtering and prediction problems". Journal of Basic Engineering, Transactions of the ASME 82: 35-45.

[51] Karlsson, S., Skoglund, J. (2001). "Specification and estimation of random effects models with serial correlation of general form". SSE/EFI Working Paper Series in Economics and Finance No. 433.

[52] Karlsson, S., Skoglund, J. (2004). "Maximum-likelihood based inference in the two-way random effects model with serially correlated time effects". Empirical Economics 29, 79-88.

[53] Kholodilin, K. A. , Siliverstovs, B., Kooths, S. (2007). "Dynamic Panel Data Approach to the Forecasting of the GDP of German Lander". DIW-Diskussionspapiere, No. 664.

[54] Kim, H. H., Swanson, N. R. (2013). "Mining Big Data Using Parsimonious Factor and Shrinkage Methods". Working Papers, Department of Economics, Rutgers, The State University of New Jersey, No. 16.

[55] Kline, D. M. (2004). "Methods for multi-step time series forecasting with neural networks". In G. Peter Zhang, editor, Neural Networks in Business Forecasting: 226-250.

[56] Kouassi, E., Sango, J., Bosson Brou, J. M., Teubissi, F. N., Kymn, K. O. (2011). "Prediction from the Regression Model with Two-Way Error Components". Journal of Forecasting, Volume 30: 541-564.

[57] Kouassi, E., Sango, J., Bosson Brou, J. M., Teubissi, F. N., Kymn, K. O. (2012). "Prediction from the One Way Error Components Model with AR(1) Disturbances". Journal of Forecasting, Volume 31, Issue 7: 617-638.

[58] Kyung, M., Gill, J., Ghosh, M., Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos". Bayesian Analysis 5 (2): 369-412.

[59] Liao, SH., Chu, PH., Hsiao, PY. (2012). "Data mining techniques and applications. A decade review from 2000 to 2011". Expert Systems with Applications 39 (2012): 11303-11311.

[60] Lykou, A., Ntzoufras, I. (2013). "On Bayesian lasso variable selection and the specification of the shrinkage parameter". Statistics and Computing, Volume 23 (3): 361-390.

[61] MacKay, D. J. C. (1992). "Bayesian Interpolation". Neural Computation, 4: 415-447.

[62] MacKay, D. J. C. (1992). "A Practical Bayesian Framework for Backpropagation Networks". Neural Computation, 4: 448-472.

[63] McNames, J. (1998). "A nearest trajectory strategy for time series prediction". In Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modelling: 112-128.

[64] Mitchell, T. (1997). "Machine Learning". McGraw-Hill.

[65] Mitchell, T. (2006). "The Discipline of Machine Learning". CMU-ML-06-108.

[66] Nadaraya, E. A. (1964). "On estimating regression". Theory of Probability and its Applications 10: 186-190.

[67] Ng, V., Engle, R. F., Rothschild, M. (1992). "A multi-dynamic-factor model for stock returns". Journal of Econometrics 52: 245-266.

[68] Nyman, R., Ormerod, p., Smith, R., Tuckett, D. (2014). "Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis". Submission to ECB Workshop on Big Data for Forecasting and statistics. Frankfurt 7/8 April.

[69] Park, T., Casella, G. (2008). "The Bayesian Lasso". Journal of the American Statistical Association Vol. 103, No. 482: 681-686.

[70] Powell, M. J. D. (1987). "Radial basis functions for multivariable interpolation: A review". In Algorithms for Approximation, J. C. Mason and M. G. Cox, Eds. Oxford: Clarendon: 143-168.

[71] Qian, W., Yang, Y. (2013). "Model selection via standard error adjusted adaptive lasso". Annals of the Institute of Statistical Mathematics, Volume 65 (2): 295-318.

[72] Rasmussen, C. E., Williams, C. K. L. (2006). "Gaussian Processes for Machine Learning". MIT Press.

[73] Saetrom, J., Omre, H. (2011). "Ensemble Kalman filtering with shrinkage regression techniques". Computational Geosciences, Volume 15 (2): 271-292.

[74] Saetrom, J., Omre, H. (2012). "Ensemble Kalman Filtering in a Bayesian Regression Framework". Working Paper presented at Ninth International Geostatistics Congress, Oslo, Norway.

[75] Schmalensee, R., Stoker, T. M., Judson, R.A. (1998). "World carbon dioxide emissions: 1950-2050". MIT Press Journal.

[76] Scott, S., Varian, H. R. (2013). "Predicting the Present with Bayesian Structural Time Series". Google Technical Report: 1-21.

[77] Scott, S., Varian, H. R. (2013). "Bayesian Variable Selection for Nowcasting Economic Time Series". Slides from talk at Berkeley University.

[78] Scott, S., Varian, H. R. (2014). "Bayesian Variable Selection for Nowcasting Economic Time Series". Google Technical Report: 1-22.

[79] Sorjamaa, A., Lendasse, A. (2006). "Time series prediction using DirRec strategy". In M. Verleysen, editor, ESANN06, European Symposium on Artificial Neural Networks: 143-148.

[80] Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., Lendasse, A. (2007). "Methodology for long-term prediction of time series". Neurocomputing, 70 (16-18): 2861-2869.

[81] Stock, J. H., Watson, M. W. (2002). "Macroeconomic forecasting using diffusion indexes". Journal of Business and Economic Statistics 20: 147-162.

[82] Stock, J. H., Watson, M. W. (2002). "Forecasting Using Principal Components from a Large Number of Predictors". Journal of the American Statistical Association. Vol. 97, no. 460: 1167-1179.

[83] Stock, J. H., Watson, M. W. (2005). "An empirical comparison of methods for forecasting using many predictors". Harvard and Princeton University Working paper.

[84] Stock, J. H., Watson, M. W. (2005). "Implications of Dynamic Factor Models for VAR Analysis". NBER Working Paper Series, no. 11467.

[85] Stock, J. H., Watson, M. W. (2006). "Forecasting with Many Predictors". Chapter 10 in Handbook of Economic Forecasting, Volume 1: 515-554, by Elliott, G., Granger, C. W. J., Timmermann, A.

[86] Stock, J. H., Watson, M. W. (2007). "Forecasting in Dynamic Factor Models Subject to Structural Instability". In The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry, by Castle, J., Shephard, N.

[87] Stock, J. H., Watson, M. W. (2010). "Dynamic Factor Models". Working Paper, prepared for Oxford Handbook of Economic Forecasting.

[88] Stock, J. H., Watson, M. W. (2012). "Generalized Shrinkage Methods for Forecasting Using Many Predictors". Journal of Business and Economic Statistics, 30 (4): 481-493.

[89] Stock, J. H., Watson, M. W. (2014). "Introduction to Econometrics, Update". Prentice Hall, 3rd edition.

[90] Taieb, S.B., Sorjamaa, A., Bontempi, G., Lendasse, A. (2009). "Long-term prediction of time series by combining direct and mimo strategies". International Joint Conference on Neural Networks.

[91] Taieb, S.B., Sorjamaa, A., Bontempi, G. (2010). "Multiple-output modelling for multi-step-ahead forecasting". Neuro-computing, 73: 1950-1957.

[92] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". Journal of the Royal Statistical Society. Series B (Methodological). Vol. 58 (1): 267-288.

[93] Tran, V. T., Yang, B. S., Tan, A. C. C. (2009). "Multi-step ahead direct prediction for the machine condition prognosis using regression trees and neuro-fuzzy systems. Expert Systems with Applications, 36(5): 9378-9387.

[94] Tsay, R. S. (2010). "Analysis of Financial Time Series". 3rd Edition. John Wiley & Sons Inc.

[95] Van Houwelingen, J. C. (2001). "Shrinkage and penalized likelihood as methods to improve predictive accuracy". Statistica Neerlandica, Vol. 55, nr. 1: 17-34.

[96] Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives, 28(2): 3-28.

[97] Varian, H. R. (2014). "Machine Learning and Econometrics". Slides package from talk at University of Washington.

[98] Varian, H. R. (2014). "Beyond Big Data". Transcription of invited talk at NABE, San Francisco.

[99] Weigend, A.S., Gershenfeld, N.A. (1994). "Time Series Prediction: forecasting the future and understanding the past". Addison Wesley, Harlow.

[100] Woolridge, J. M. (2010). "Econometric Analysis of Cross Section and Panel Data". The MIT Press.

[101] Woolridge, J. M. (2012). "Introductory Econometrics: A Modern Approach". Cengage Learning.

[102] Wu, J., Zhu, L. (2011). "Testing for serial correlation and random effects in a two-way error component regression model". Economic Modelling 28: 2377-2386.

# Insurance Analytics References


[1] Akerlof, G. A. (1970). "The Market for Lemons: Quality Uncertainty and the Market Mechanism". The Quarterly Journal of Economics 84 (3): 488-500.

[2] Borsch-Supan, A., A. Brugiavini, H. Jurges, J. Mackenbach, J. Siegrist and G. Weber. (2005). "Health, ageing and retirement in Europe - First results from the Survey of Health, Ageing and Retirement in Europe". Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

[3] Borsch-Supan, A. and H. Jurges (Eds.). (2005). "The Survey of Health, Ageing and Retirement in Europe-Methodology". Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

[4] Borsch-Supan, A., A. Brugiavini, H. Jurges, A. Kapteyn, J. Mackenbach, J. Siegrist and G. Weber. (2008). "First results from the Survey of Health, Ageing and Retirement in Europe (2004-2007). Starting the longitudinal dimension". Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

[5] Borsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S. (2013). "Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE)". International Journal of Epidemiology.

[6] Bryan, M. L. & Jenkins, S. P. (2013). "Regression Analysis of Country Effects Using Multilevel Data: A Cautionary Tale". IZA Discussion Paper Series No. 7583.

[7] Cardon, J., & Hendel, I. (2001). "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey". RAND Journal of Economics 32: 408-427.

[8] Cawley, J., & Philipson, T. (1999). "An empirical examination of information barriers to trade in insurance". American Economic Review, 89 (4): 827-846.

[9] Chiappori, P. A., & Salanie, B. (2000). "Testing for Asymmetric Information in Insurance Markets". Journal of Political Economy, 108(1): 56-78.

[10] Chiappori, P. A., Jullien, B., Salanie, B. & Salanie, F. (2006). "Asymmetric Information in Insurance: General Testable Implications". Rand Journal of Economics, 37(4): 783-98.

[11] Cohen, A., & Einav, L. (2007). "Estimating risk preferences from deductible choice". American Economic Review, 97 (3) : 745-788.

[12] Cutler, D. M., & Zeckhauser, R. (2000). "The Anatomy of Health Insurance". In Handbook of Health Economics, Volume 1A, ed. A. Culyer and J. Newhouse, 563-643. Amsterdam: Elsevier.

[13] Cutler, D. M., Finkelstein, A. & McGarry K. (2008). "Preference Heterogeneity and Insurance Markets: Explaining a Puzzle of Insurance". American Economic Review: Papers & Proceedings 98 (2): 157-162.

[14] De Meza, D., Webb, D. C. (2001). "Advantageous selection in insurance markets". RAND Journal of Economics 32 (2): 249-262.

[15] Dionne, G., Gouriéroux, C. & Vanasse, C. (2001). "Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment". Journal of Political Economy 109 (2): 444-453.

[16] Einav, L., Finkelstein, A. & Schrimpf, P. (2007). "The Welfare Cost of Asymmetric Information: Evidence from the U.K. Annuity Market". NBER Working Paper No. 13228.

[17] Einav, L., Finkelstein, A. & Levin, J. (2010). "Beyond Testing: Empirical Models of Insurance Markets". Annual Review of Economics 2: 311-336.

[18] Einav, L., Finkelstein, A. & Levin, J. (2011). "Selection in Insurance Markets: Theory and Empirics in Pictures". Journal of Economic Perspectives, 25 (1): 115-138.

[19] Ettner, S. (1997). "Adverse Selection and the Purchase of Medigap Insurance by the Elderly". Journal of Health Economics, 16 (5): 499-624.

[20] Fang, H., Keane, M. & Silverman, D. (2006). "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market". NBER Working Paper No. 12289.

[21] Finkelstein, A. & McGarry K. (2006). "Private Information and its Effect on Market Equilibrium: New Evidence from Long-Term Care Insurance". American Economic Review 96 (4): 938-58.

[22] Finkelstein, A. & McGarry K. (2006). "Multiple dimensions of private information: evidence from the long-term care insurance market". American Economic Review 96 (4): 938-958.

[23] Finkelstein, A., & Poterba, J. (2002). "Selection Effects in the Market for Individual Annuities: New Evidence from the United Kingdom". Economic Journal, 112 (476): 28-50.

[24] Finkelstein, A., & Poterba, J. (2004). "Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market". Journal of Political Economy, 112 (1): 183-208.

[25] Finkelstein, A., & Poterba, J. (2006). "Testing for Asymmetric Information Using 'Unused Observables' in Insurance Markets: Evidence from the U.K. Annuity Market". NBER Working Paper No. 12112.

[26] Hemenway, D. (1990). "Propitious selection". Quarterly Journal of Economics 105 (4): 1063-1069.

[27] Hurd, M. & McGarry, K. (1997). "Medical Insurance and the Use of Health Care Services by the Elderly". Journal of Health Economics, 16 (2): 129-154.

[28] McCarthy, D., & Mitchell, O. S. (2003). "International Adverse Selection in Life Insurance and Annuities". NBER Working Paper No. 9975.

[29] Mitchell, O. S., Poterba, J. M., Warshawsky, M. J., & Brown, J. R. (1999). "New Evidence on the Money's Worth of Individual Annuities". American Economic Review 89: 1299-1318.

[30] Rothschild, M. & Stiglitz, J. (1976). "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information". Quarterly Journal of Economics 90: 630-649.

# Behavioral Insurance References


[1] Aunon-Nerin, D., Ehling, P. (2008). "Why firms purchase property insurance". Journal of Financial Economics, 90: 298-312.

[2] Eeckhoudt, L., Gollier, C., Schlesinger, H. (1996). "Changes in background risk and risk taking behavior". Econometrica 64 (3): 683-689.

[3] Froot, K. A., Scharfstein, D. S., Stein, J. C. (1993). "Risk Management: Coordinating Corporate Investment and Financing Policies". The Journal of Finance, 68(5): 1629-1658.

[4] Garven, J. R., MacMinn, R. D. (1993). "The underinvestment problem, bond covenants, and insurance". The Journal of Risk and Insurance: 635-646.

[5] Gollier, C. (2010). "The Demand of the Insurance Demand by Firms". Mimeo.

[6] Guiso, L., Schivardi, F. (2010a). "Copertura assicurativa e accesso al credito bancario: evidenza da un campione di piccole e medie imprese italiane". Working paper.

[7] Guiso, L., Schivardi, F. (2010b). "La domanda di assicurazione delle imprese. Risultati dall'Indagine Ania sull'Assicurazione nelle Piccole Imprese Italiane". Working Paper.

[8] Han, L. M., MacMinn, R. D. (2006). "Stock options and the corporate demand for insurance". Journal of Risk and Insurance 73 (2): 231-260.

[9] Hoyt, R. E., Khang, H. (2000). "On the demand for corporate property insurance". The Journal of Risk and Insurance, 67(1): 91-107.

[10] Hubbard, R. (1998). "Capital-market imperfections and investment". Journal of Economic Literature, 36: 193-225.

[11] Jensen, M. C., Smith, C. W. (1985). "Stockholder, Manager, and Creditor Interests: Applications of Agency Theory". In Theory of the Firm 1(1), 2000.

[12] MacMinn, R. D. (1987). "Insurance and corporate risk management". The Journal of Risk and Insurance: 658-677.

[13] MacMinn, R. D., Garven, J. R. (2013). "On the Demand for Corporate Insurance: Creating Value". Handbook of Insurance: 487-516.

[14] Mayers, D., Smith, C. W. (1982). "On the Corporate Demand for Insurance". The Journal of Business, 55(2): 281-296.

[15] Mayers, D., Smith, C. W. (1987). "Corporate Insurance and the Underinvestment Problem". Journal of Risk and Insurance, 54(1): 45-54.

[16] Zou, H., Adams, M. B. (2006). "The corporate purchase of property insurance: Chinese evidence". Journal of Financial Intermediation, 15: 165-196.

[17] Zou, H., Adams, M. B. (2008). "Debt capacity, cost of debt, and corporate insurance". Journal Of Financial And Quantitative Analysis, 43(2): 433-466.

# Sentiment Analysis References


[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). "Sentiment Analysis of Twitter Data". LSM '11 Proceedings of the Workshop on Languages in Social Media: 30-38.

[2] Antweiler, W., Frank, M.Z. (2004). "Is all that talk just noise? The information content of internet stock message boards". The Journal of Finance 59 (3): 1259-1294.

[3] Asur, S., Huberman, B. (2010). "Predicting the Future With Social Media". Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on Volume 1: 492-499.

[4] Baker, M., Wurgler, J. (2006). "Investor Sentiment and the Cross-Section of Stock Returns". The Journal of Finance Volume 61 (4): 1645-1680.

[5] Baker, M., Wurgler, J. (2007). "Investor Sentiment in the Stock Market". Journal of Economics Perspectives Volume 21, No. 2: 129-151.

[6] Barber, B., Lehavy, R., McNichols, M., Trueman, B. (2001). "Can Investors Profit from the Prophets? Security Analyst Recommendations and Stock Returns". The Journal of Finance 56 (2): 531-563.

[7] Barber B. M., Odean, T. (2008). "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors". Review of Financial Studies 21 (2): 785-818.

[8] Berg, J., Neumann, G. R., Rietz, T. A. (2009). "Searching for Google's Value: Using Prediction Markets to Forecast Market Capitalisation Prior to an Initial Public Offering". Journal Management Science Volume 55 (3): 348-361.

[9] Bollen, J., Mao, H. (2011). "Twitter Mood as a Stock Market Predictor". IEEE Computer. Vol. 44 (10): 91-94.

[10] Bollen, J., Mao, H., Zeng, X. (2011). "Twitter mood predicts the stock market". Journal of Computational Science Volume 2 (1): 1-8.

[11] Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I. (2012). "Web search queries can predict stock market volumes". PloS one 7 (7), e40014.

[12] Brown, E. D. (2012). "Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market". SAIS 2012 Proceedings. Paper 7.

[13] Choi, H., Varian, H. (2012). "Predicting the Present with Google Trends". Economic Record, Special Issue: Selected Papers from the 40th Australian Conference of Economists Volume 88, Issue Supplement s1, pages 2-9.

[14] Corea, F., Cervellati, E. M. (2015). "The Power of Micro-Blogging: How to Use Twitter for Predicting the Stock Market". Eurasian Journal of Economics and Finance, 3 (4): 1-6.

[15] Corea, F. (2015). "Why social media matters: the use of Twitter in portfolio strategies". International Journal of Computer Applications, 128 (6): 25-30.

[16] Culotta, A. (2010). "Towards detecting influenza epidemics by analysing Twitter messages". Proceedings of the First Workshop on Social Media Analytics: 115-122.

[17] Da, Z., Engelberg, J., Gao, P. (2012). "In search of attention". The Journal of Finance 66 (5): 1461-1499.

[18] Da, Z., Engelberg, J., Gao, P. (2015). "The Sum of All FEARS Investor Sentiment and Asset Prices". Review of Financial Studies 28 (1), 1-32.

[19] De Bondt, W., Thaler, R. (1985). "Does the stock market overreact?". Journal of Finance 40: 1837-1864.

[20] De Bondt, W., Thaler, R. (1989). "Anomalies: A mean-reverting walk down Wall Street". Journal of Economic Perspectives 3 (1): 189-202.

[21] De Choudhury, M., Sundaram, H., John, A., Seligmann, D. D. (2008). "Can blog communication dynamics be correlated with stock market activity?". Proceedings of the nineteenth ACM conference on Hypertext and hypermedia: 55-60.

[22] Dhar, V., Chang, E. A. (2009). "Does Chatter Matter? The Impact of User-Generated Content on Music Sales". Journal of Interactive Marketing Volume 23 (4): 300-307.

[23] Fisher, K. L., Statman, M. (2000). "Investor Sentiment and Stock Returns". Financial Analysts Journal, Vol. 56, No. 2: 16-23.

[24] Frieder, L., Subrahmanyan, A. (2005). "Brand perceptions and the market for common stock". Journal of Financial and Quantitative Analysis, Volume 40 (1): 57-85.

[25] Kearney, C., Liu, S. (2014). "Textual Sentiment in Finance: A Survey of Methods and Models". International Review of Financial Analysis, 33: 171-185.

[26] Koski, J. L., Rice, E. M., Tarhouni, A. (2008). "Day Trading and Volatility: Evidence from Message Board Postings in 2002 vs. 1999". Working paper under review by Management Science.

[27] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J. (2000). "Language Models for Financial News Recommendation". Proceedings of the ninth international conference on Information and knowledge management: 389-396.

[28] Mao, H., Bollen, J., Counts, S. (2011)."Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data". Working Paper.

[29] Mao, H., Counts, S., Bollen, J. (2015)."Quantifying the effects of online bullishness on international financial markets". ECB Statistics Paper Series, 9.

[30] Michaely, R., Thaler, R. H., Womack, K. L. (1995). "Price reactions to dividend initiations and omissions: Overreaction or drift?". The Journal of Finance 50 (2), 573-608.

[31] Mishne, G., Glance, N. (2006). "Predicting movie sales from blogger sentiment". In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.

[32] Mittal, A., Goel, A. (2012). "Stock Prediction Using Twitter Sentiment Analysis". Working Paper Stanford University CS 229.

[33] Nofsinger, J.R. (2005). "Social mood and financial economics". The Journal of Behavioural Finance 6 (3): 144-160.

[34] Oh, C., Sheng, O. R. L. (2011). "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement". ICIS 2011 Proceedings.

[35] Oliveira, N., Cortez, P., Areal, N. (2013). "On the Predictability of Stock Market Behaviour Using StockTwits Sentiment and Posting Volume". Progress in Artificial Intelligence, Lecture Notes in Computer Science Volume 8154: 355-365.

[36] Peterson, R.L. (2007). "Affect and financial decision-making: How neuroscience can inform market participants". The Journal of Behavioural Finance 8 (2): 70-78.

[37] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. (2012). "Correlating financial time series with micro-blogging activity". Proceedings of the fifth ACM international conference on Web search and data mining: 513-522.

[38] Schumaker, R.P., Chen, H. (2009). "Textual analysis of stock market prediction using breaking financial news: The azfin text system". ACM Transactions on Information Systems (TOIS) 27 (2), 12.

[39] Sprenger, T., Welpe, I. (2010). "Tweets and trades: The information content of stock microblogs". Social Science Research Network Working Paper Series: 1-89.

[40] Tetlock, P. C. (2007). "Giving content to investor sentiment: The role of media in the stock market". The Journal of Finance 62 (3): 1139-1168.

[41] Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S. (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals". Journal of Finance 63, 1437-1467.

[42] Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M. (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[43] Zhang, L. (2013). "Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation". Thesis, Department of Computer Science, The University of Texas at Austin.

# Chapter 8

# Appendix

## Appendix Insurance

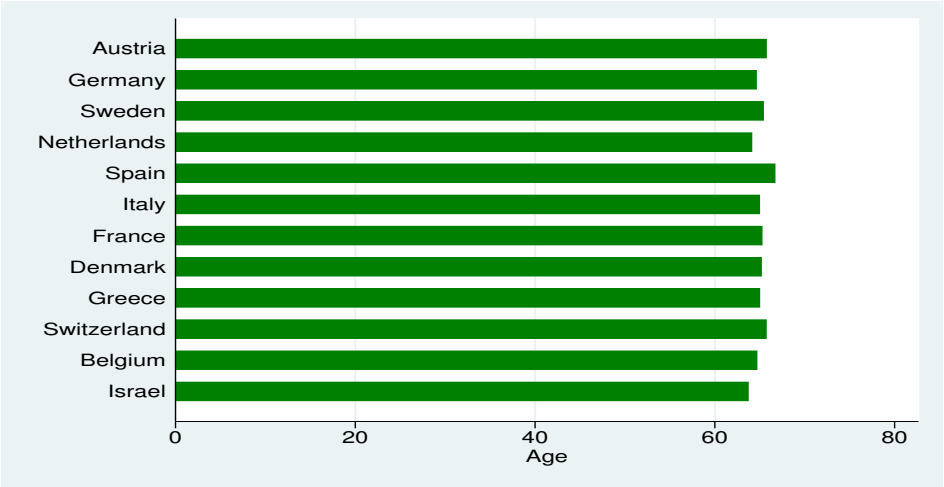



Figure 8.1: Key summary statistics for average age per country.

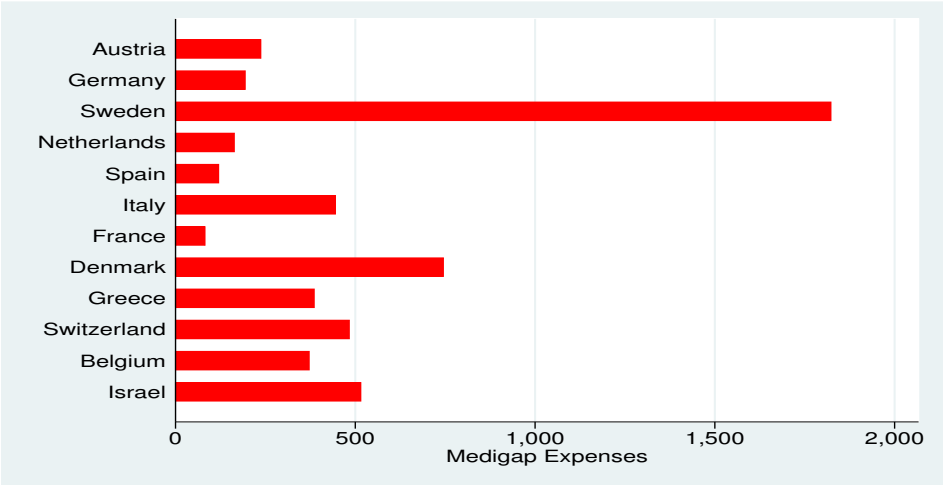



Figure 8.2: Key summary statistics for medigap expenses per country (%).

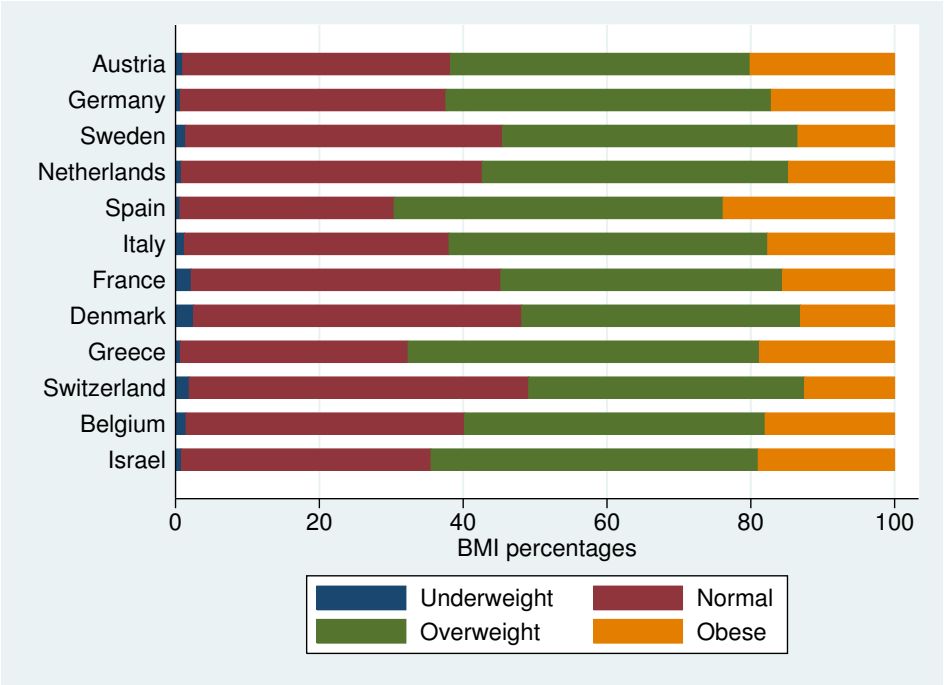
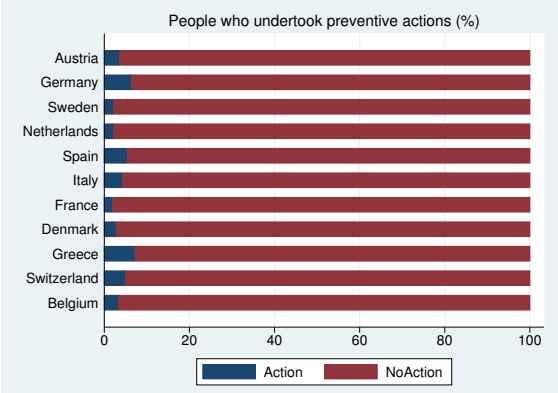
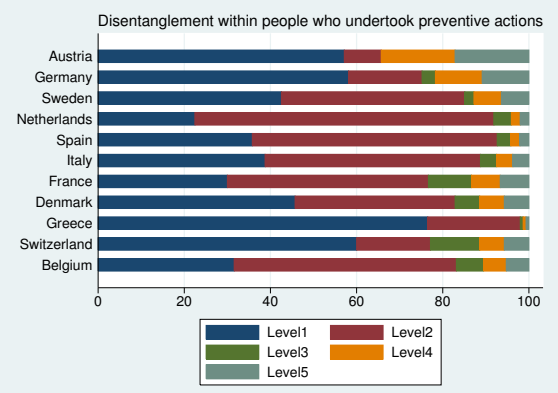



Figure 8.3: Key summary statistics for bmi index per country (%).

Figure 8.4: Key summary statistics for different level of prevention (number of preventive actions) per country (%).



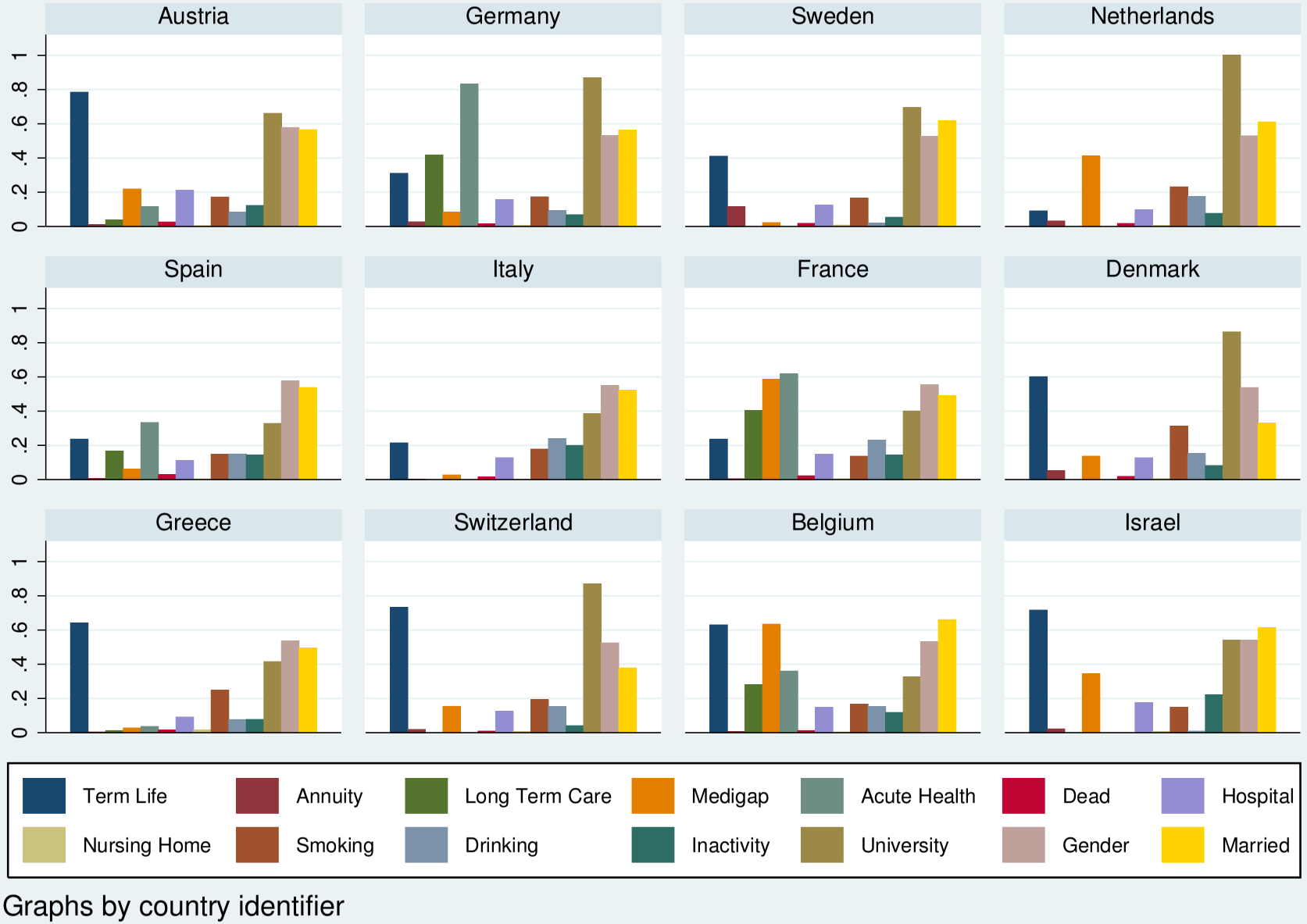



Figure 8.5: Key summary statistics for other variables (%).

Table 8.1: Relation between Insurance and Risky behaviours (Pooled Probit regression).

| | Term life | | Annuity | | Lt care | | Medigap | | Acute health | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| main | | | | | | | | | | |
| Smoking | 0.110* | 0.130** | 0.0389 | 0.0276 | -0.296*** | -0.277*** | -0.103 | -0.138 | -0.0322 | -0.0308 |
| | (2.37) | (3.25) | (0.37) | (0.29) | (-5.43) | (-4.25) | (-1.08) | (-1.94) | (-0.49) | (-0.65) |
| Drinking | -0.285*** | -0.310*** | -0.0729 | -0.0678 | -0.175 | -0.286 | 0.241 | 0.243 | -0.387*** | -0.389** |
| | (-4.25) | (-5.28) | (-0.68) | (-0.66) | (-1.15) | (-1.88) | (1.43) | (1.87) | (-3.44) | (-3.26) |
| BMI | 0.0283 | 0.0122 | -0.115** | -0.116 | -0.0138 | -0.114 | -0.151*** | -0.154*** | -0.311*** | -0.311*** |
| | (0.59) | (0.23) | (-2.62) | (-1.91) | (-0.16) | (-1.70) | (-4.28) | (-3.33) | (-4.40) | (-4.70) |
| Preventive | -0.0431 | -0.0363 | -0.100** | -0.110*** | 0.139 | 0.183 | -0.0598 | -0.0525 | 0.189** | 0.189** |
| | (-1.17) | (-1.06) | (-3.25) | (-4.47) | (1.52) | (1.53) | (-1.37) | (-1.29) | (3.03) | (2.75) |
| Inactivity | -0.179 | -0.215 | -0.408*** | -0.468*** | -0.716* | -0.819** | -0.146 | -0.0985 | -0.0624 | -0.0683 |
| | (-0.97) | (-1.50) | (-7.19) | (-6.72) | (-2.13) | (-3.07) | (-1.49) | (-1.73) | (-0.62) | (-0.36) |
| *N* | 2657 | 2657 | 22221 | 22221 | 1269 | 1269 | 22233 | 22233 | 1269 | 1269 |

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8.2: Relation between Risk occurrence and Risky behaviours (Pooled LPM regression).

| | Dead | | Alive | | Nursing Home | | Medigap Exp | | Hospital | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| Smoking | -0.00399 | 0.00529* | -0.0217 | -0.0285 | -0.00198 | -0.000663 | -56.18 | -8.348 | -0.0223** | -0.0238*** |
| | (-1.86) | (2.47) | (-1.33) | (-1.84) | (-1.42) | (-0.42) | (-1.20) | (-0.31) | (-3.92) | (-4.64) |
| Drinking | 0.00745 | 0.0000605 | 0.0507 | 0.0483 | -0.00166 | -0.00272 | -45.09 | -42.96 | -0.00409 | -0.00232 |
| | (1.87) | (0.01) | (1.87) | (1.93) | (-1.05) | (-1.23) | (-1.66) | (-1.28) | (-0.60) | (-0.29) |
| BMI | -0.00285 | -0.00398 | 0.0202** | 0.0161** | -0.000794 | -0.000787 | -77.96 | -71.38 | 0.00773 | 0.00825 |
| | (-0.88) | (-1.00) | (3.88) | (3.17) | (-0.42) | (-0.44) | (-0.96) | (-0.88) | (0.84) | (0.89) |
| Preventive | 0.00267 | 0.00301 | -0.000823 | 0.00228 | -0.000145 | -0.000224 | -6.056 | -16.64 | 0.0262*** | 0.0260*** |
| | (0.72) | (1.20) | (-0.11) | (0.29) | (-0.10) | (-0.16) | (-0.39) | (-1.08) | (9.14) | (8.91) |
| Inactivity | 0.0777*** | 0.0635*** | -0.103* | -0.0841* | 0.0164 | 0.0151 | 253.3** | 190.1** | 0.169*** | 0.173*** |
| | (12.06) | (10.11) | (-3.08) | (-2.68) | (1.81) | (1.90) | (4.05) | (3.93) | (17.17) | (13.96) |
| *N* | 22233 | 22233 | 22233 | 22233 | 15040 | 15040 | 22233 | 22233 | 22226 | 22226 |

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

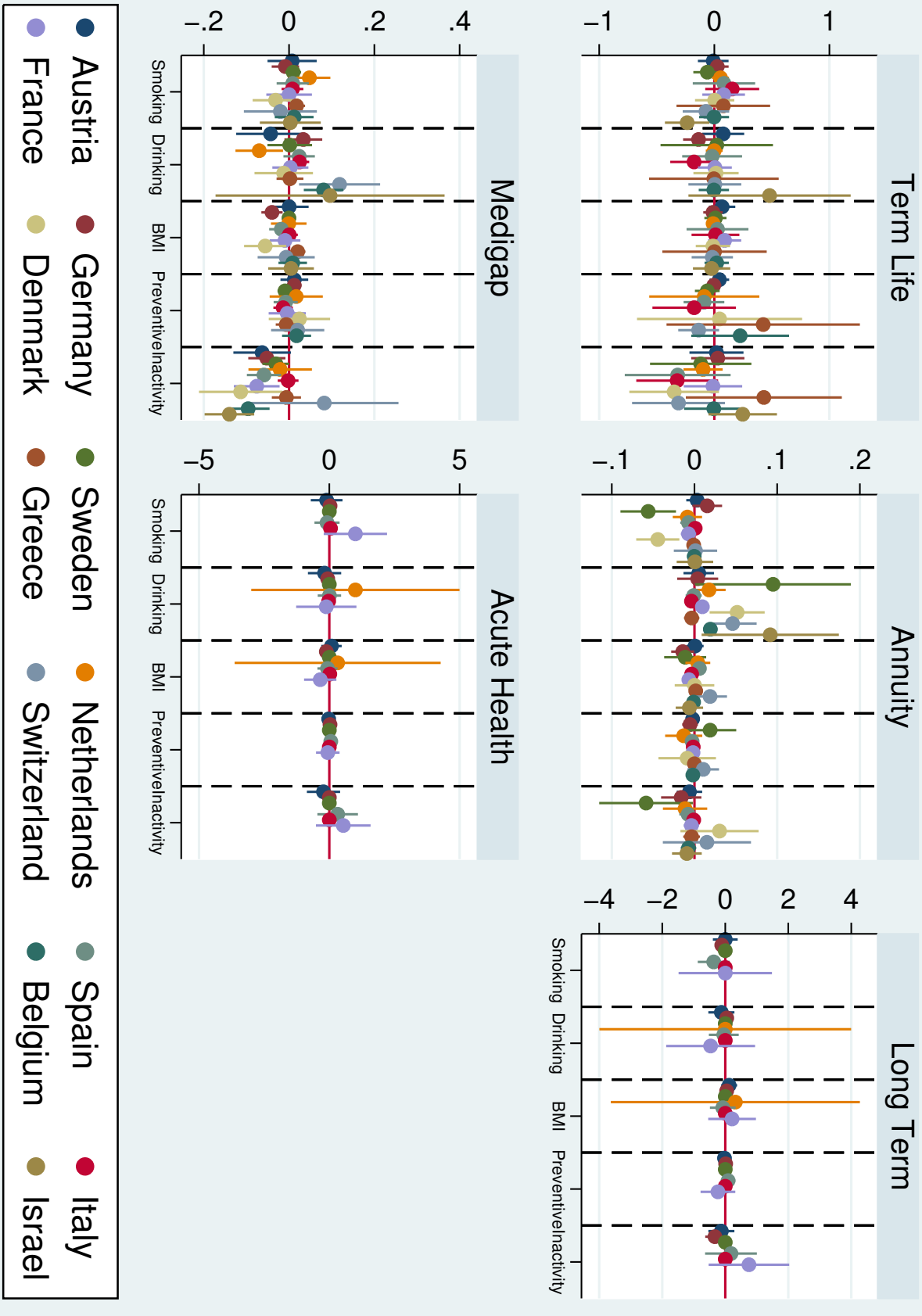



Figure 8.6: Relation between Insurance and Risky behaviours (LPM regression) per country.

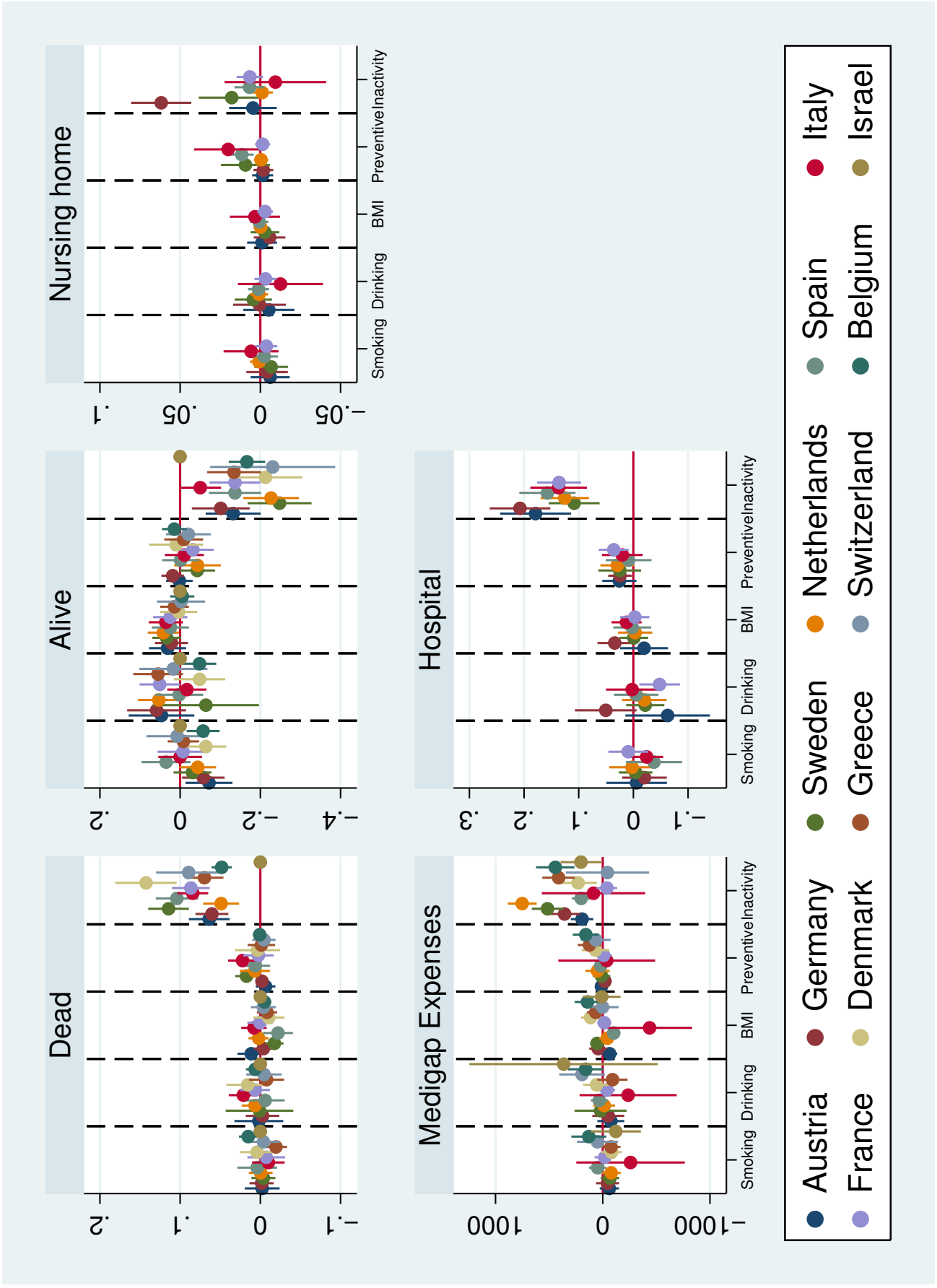



Figure 8.7: Relation between Risk occurrence and Risky behaviours (LPM regression) per country.

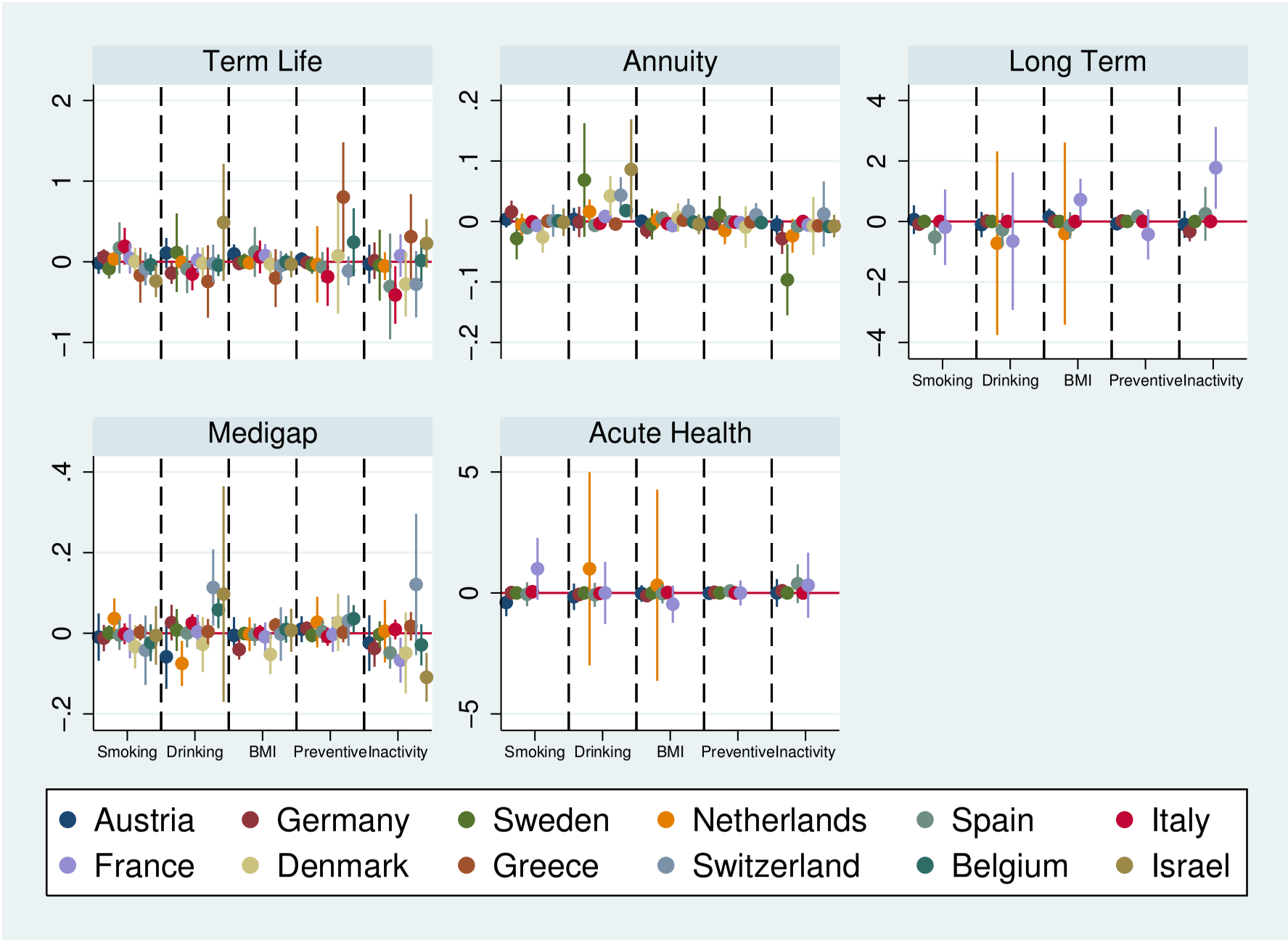



Figure 8.8: Relation between Insurance and Risky behaviours (LPM regression) per country with control variables.

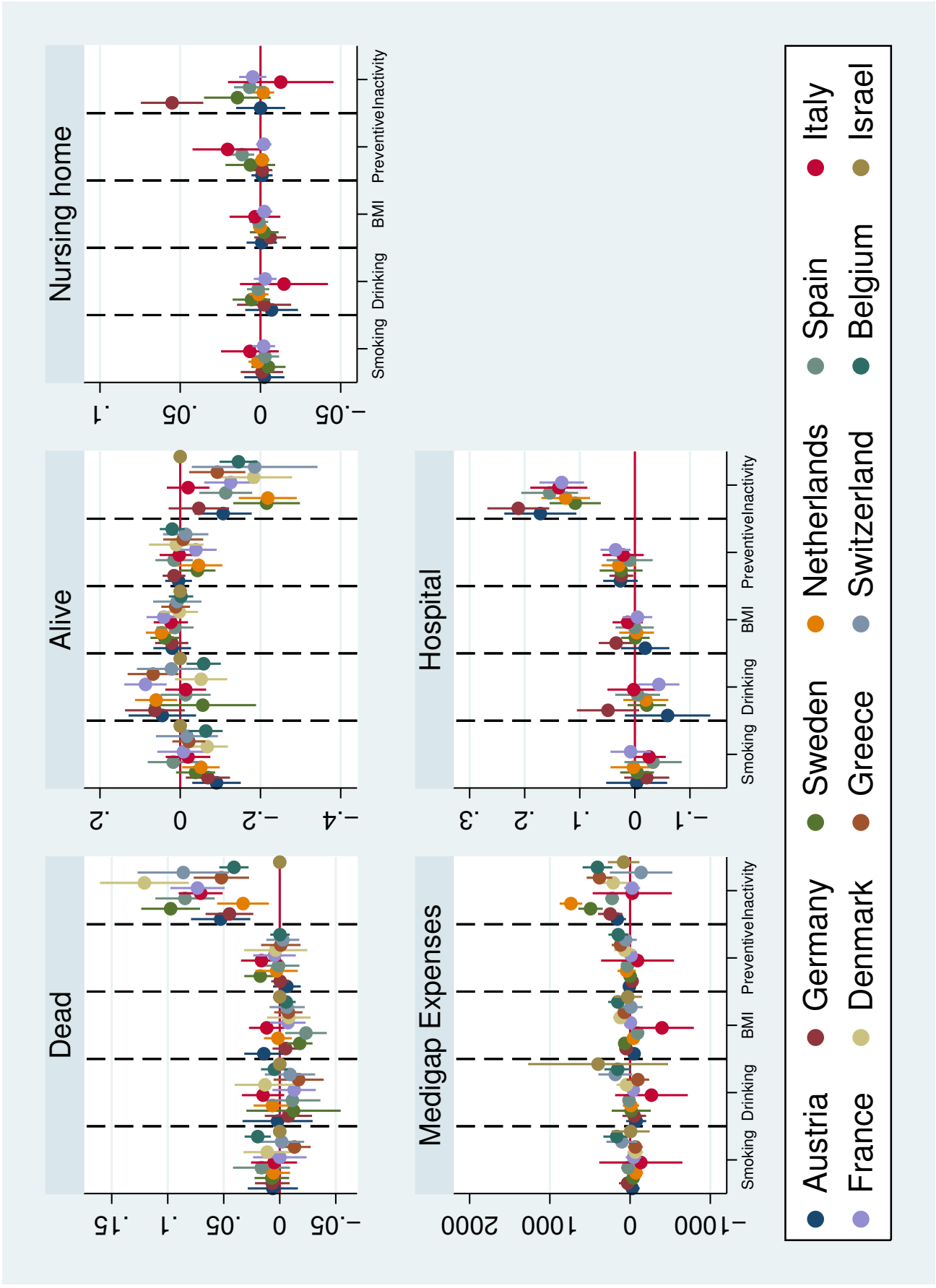



Figure 8.9: Relation between Risk occurrence and Risky behaviours (LPM regression) per country with control variables.

Table 8.3: Relation between Insurance and Risky behaviours with fixed-effect (Probit regression).

| | Term life | | Annuity | | Lt care | | Medigap | | Acute health | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| main | | | | | | | | | | |
| Smoking | 0.0947* | 0.120** | -0.00152 | 0.0190 | -0.347*** | -0.315*** | 0.0218 | -0.0340 | 0.0961 | 0.0445 |
| | (2.40) | (2.70) | (-0.01) | (0.18) | (-5.78) | (-3.99) | (0.86) | (-1.68) | (1.29) | (0.70) |
| Drinking | -0.194 | -0.227* | 0.161 | 0.0964 | 0.0955* | -0.0229 | 0.0988* | 0.0764 | -0.0981 | -0.0613 |
| | (-1.76) | (-2.22) | (1.68) | (0.98) | (2.23) | (-0.33) | (2.00) | (1.74) | (-0.68) | (-0.37) |
| BMI | 0.0435 | 0.0270 | -0.123 | -0.128 | 0.0999* | -0.00520 | -0.0804 | -0.0709 | -0.459*** | -0.464*** |
| | (0.83) | (0.50) | (-1.90) | (-1.87) | (2.48) | (-0.12) | (-1.81) | (-1.49) | (-13.05) | (-14.13) |
| Preventive | -0.0428 | -0.0376 | -0.111** | -0.104*** | 0.0757 | 0.102 | 0.0229 | 0.0337 | 0.150*** | 0.138** |
| | (-1.18) | (-1.31) | (-3.10) | (-3.59) | (1.33) | (1.38) | (0.87) | (1.75) | (3.38) | (2.96) |
| Inactivity | -0.193 | -0.241 | -0.290*** | -0.426*** | -0.835** | -0.890** | -0.234*** | -0.140* | 0.113 | 0.333*** |
| | (-1.03) | (-1.58) | (-4.98) | (-5.02) | (-2.65) | (-3.12) | (-4.49) | (-2.38) | (0.79) | (3.53) |
| *N* | 2657 | 2657 | 22221 | 22221 | 332 | 332 | 22233 | 22233 | 390 | 390 |

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8.4: Relation between Risk occurrence and Risky behaviours with fixed-effect (LPM regression).

| | Dead | | Alive | | Nursing Home | | Medigap Exp | | Hospital | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| Smoking | -0.00379 | 0.00536* | -0.0232 | -0.0344* | -0.00285* | -0.00141 | -67.52 | -20.02 | -0.0205** | -0.0200** |
| | (-1.75) | (2.65) | (-1.60) | (-2.33) | (-2.76) | (-1.21) | (-1.66) | (-0.72) | (-4.08) | (-4.18) |
| Drinking | 0.00629 | -0.00132 | 0.0212 | 0.0279 | 0.000407 | -0.000893 | -70.66 | -76.88 | 0.00319 | 0.00362 |
| | (1.29) | (-0.25) | (1.31) | (1.46) | (0.31) | (-0.48) | (-1.28) | (-1.32) | (0.31) | (0.36) |
| BMI | -0.00290 | -0.00383 | 0.0248*** | 0.0235*** | -0.00106 | -0.00121 | -82.38 | -71.16 | 0.00858 | 0.00831 |
| | (-0.84) | (-0.95) | (8.55) | (6.04) | (-0.57) | (-0.70) | (-1.01) | (-0.91) | (1.05) | (0.97) |
| Preventive | 0.00268 | 0.00300 | 0.00378 | 0.00498 | -0.000324 | -0.000322 | -4.803 | -15.51 | 0.0248*** | 0.0247*** |
| | (0.74) | (1.21) | (0.42) | (0.65) | (-0.21) | (-0.23) | (-0.38) | (-1.11) | (9.41) | (9.56) |
| Inactivity | 0.0780*** | 0.0639*** | -0.107*** | -0.0757** | 0.0183 | 0.0164 | 201.2* | 141.7 | 0.173*** | 0.172*** |
| | (11.91) | (9.88) | (-5.18) | (-3.37) | (1.97) | (2.06) | (3.03) | (2.12) | (13.38) | (12.98) |
| *N* | 22233 | 22233 | 22233 | 22233 | 15040 | 15040 | 22233 | 22233 | 22226 | 22226 |

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

# Appendix Behavioral Insurance

Table 8.5: Number of insurance policies underwritten.

| | Number of Policies |
|---|---|
| Employee | 0.382*** |
| | (7.06) |
| Age | 0.00785*** |
| | (3.52) |
| Probability being damaged | 0.122*** |
| | (2.72) |
| Consultant | -0.0856* |
| | (-1.65) |
| Damaged (last 5 years) | -0.466*** |
| | (-3.15) |
| Having damaged (last 5 years) | -0.458** |
| | (-2.47) |
| Building | -0.429** |
| | (-2.24) |
| Savings | -0.514*** |
| | (-4.15) |
| PersonalDamage | -0.390*** |
| | (-3.39) |
| Bankruptcy | -0.00523* |
| | (-1.71) |
| Export | 0.0128*** |
| | (5.24) |
| Factories | 0.0182*** |
| | (5.07) |
| AdmOffice | 0.312** |
| | (2.17) |
| Overconfidence | 0.248** |
| | (2.00) |
| Optimism | 0.0530* |
| | (1.77) |
| Stubborness | 0.0691** |
| | (2.16) |
| Business Name | -0.00518** |
| | (-2.54) |
| Trust vs Insurance | 0.0780*** |
| | (3.02) |
| Loan | 0.268** |
| | (2.42) |
| PersonalLD | -0.225* |
| | (-1.95) |
| OwnerOffice | 0.268* |
| | (1.83) |
| Foreign Management | 1.152*** |
| | (3.98) |

$t$ statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table 8.6: Logistic regression for different risks to be insured.

| | Fire | Theft | Goods | Credit | Expo | Bus. Inter. | Third Liab. | Product Liab. | Envir. Liab. | Employee Liab. | Tech. Liab |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Employee | 0.167*** | | 0.259*** | 0.101* | 0.172** | 0.247** | 0.264*** | 0.223*** | 0.202*** | 0.170*** | 0.149*** |
| | (2.65) | | (5.37) | (1.81) | (2.43) | (2.44) | (5.44) | (4.47) | (3.37) | (3.14) | (3.23) |
| Age | 0.00783** | | | | | 0.00767** | | | 0.00836*** | | |
| | (2.06) | | | | | (2.49) | | | (3.75) | | |
| Loan | 0.449*** | 0.150* | | | | 0.643*** | 0.223** | 0.285*** | -0.349*** | | 0.173* |
| | (3.29) | (1.65) | | | | (3.12) | (2.31) | (2.78) | (-2.69) | | (1.77) |
| Savings | -0.527*** | -0.225** | -0.336*** | -0.293** | -0.576*** | -0.519** | | -0.307*** | | | -0.260** |
| | (-3.19) | (-2.09) | (-3.28) | (-2.45) | (-3.89) | (-2.48) | | (-2.79) | | | (-2.42) |
| Damaged (last 5 years) | -0.728*** | -0.268** | | | | | | | -0.340** | | -0.294** |
| | (-2.77) | (-2.02) | | | | | | | (-2.24) | | (-2.29) |
| AdmOffice | 0.415** | | | 0.257** | | | -0.244* | | | | |
| | (2.39) | | | (2.32) | | | (-1.86) | | | | |
| Passion | 0.0468* | | | 0.0409** | 0.0745*** | | | | | | |
| | (1.94) | | | (2.08) | (2.72) | | | | | | |
| Production | 0.536*** | | 0.392*** | | | | -0.189* | 0.552*** | | | |
| | (3.61) | | (3.39) | | | | (-1.92) | (5.15) | | | |
| Trade | 0.790*** | 0.466*** | 0.352*** | | -0.347* | -0.478* | | | -0.436*** | | -0.207* |
| | (4.70) | (4.30) | (2.84) | | (-1.65) | (-1.71) | | | (-2.70) | | (-1.79) |
| OwnerOffice | 0.298* | | | | | | -0.241* | | | | |
| | (1.73) | | | | | | (-1.82) | | | | |
| Business Name | -0.00594*** | -0.00509*** | | -0.00818** | | | | | | | |
| | (-3.10) | (-3.09) | | (-2.32) | | | | | | | |
| Age entrepreneur | 0.0141** | | | | | | | | | | |
| | (2.23) | | | | | | | | | | |
| Trust vs Insurance | | 0.0486** | 0.0507** | | | 0.103* | | | 0.0729** | | |
| | | (2.01) | (2.16) | | | (1.83) | | | (2.37) | | |
| Probability being damaged | | 0.0904** | | 0.214*** | | 0.157** | | | 0.166*** | | |
| | | (2.22) | | (4.34) | | (2.24) | | | (3.60) | | |
| PersonalHealth | | -0.184* | | -0.248** | | | | | | | |
| | | (-1.78) | | (-2.05) | | | | | | | |
| Trust vs Entrepreneurs | | 0.0481* | | | | -0.114* | | | | | 0.0588** |
| | | (1.78) | | | | (-1.89) | | | | | (2.30) |
| PersonalLD | | -0.218** | | | | | -0.195* | | | | -0.319*** |
| | | (-2.34) | | | | | (-1.92) | | | | (-3.43) |
| Factories | | 0.0585** | | 0.0240* | | 0.0113*** | | 0.00997** | | 0.0109* | 0.0124** |
| | | (2.47) | | (1.86) | | (2.99) | | (2.50) | | (1.95) | (2.21) |
| Education | | | -0.105*** | | | | | | | 0.114** | |
| | | | (-2.67) | | | | | | | (2.36) | |
| Probability damaging others | | | 0.0790** | -0.135** | | | 0.116*** | | | | |
| | | | (2.04) | (-2.37) | | | (2.63) | | | | |
| PersonalDamage | | | -0.205** | | | | -0.396*** | -0.296*** | -0.352*** | | |
| | | | (-2.26) | | | | (-3.85) | (-3.04) | (-2.88) | | |
| Transportation | | | 1.073*** | | | | | | | | -0.579** |
| | | | (4.97) | | | | | | | | (-2.24) |
| Foreign Management | | | 0.565** | | | 0.970*** | | | | 1.076*** | |
| | | | (2.42) | | | (2.97) | | | | (4.66) | |
| Export | | | 0.0101*** | | 0.0192*** | 0.00667* | | 0.00531** | | | |
| | | | (4.73) | | (7.70) | (1.90) | | (2.42) | | | |
| Married | | | -0.183* | | | | | | | | |
| | | | (-1.72) | | | | | | | | |
| Having damaged (last 5 years) | | | | -0.833*** | | | | | -0.391** | | |
| | | | | (-5.15) | | | | | (-2.12) | | |
| Majority share | | | | 0.00236** | | -0.00557** | | | | | |
| | | | | (1.98) | | (-2.38) | | | | | |
| Consultant | | | | -0.113** | | | | | -0.126** | | -0.0831* |
| | | | | (-2.07) | | | | | (-2.07) | | (-1.77) |
| PersonalLP | | | | 0.228* | | | | | | | |
| | | | | (1.91) | | | | | | | |
| Overconfidence | | | | 0.271** | | 0.851*** | | | | | |
| | | | | (2.00) | | (2.99) | | | | | |
| Height | | | | 0.0168** | | | | | 0.0143* | | |
| | | | | (2.43) | | | | | (1.87) | | |
| Trust vs StockMarket | | | | 0.0436* | 0.0731** | | | | | | |
| | | | | (1.66) | (2.10) | | | | | | |
| Listed | | | | | | -1.339** | | | | | |
| | | | | | | (-2.43) | | | | | |
| Optimism | | | | | | 0.222*** | | | | 0.120*** | |
| | | | | | | (3.43) | | | | (3.44) | |
| PersonalLI | | | | | | 0.615** | | | | | |
| | | | | | | (2.35) | | | | | |
| EV/family assets | | | | | | | -0.00430** | | | | |
| | | | | | | | (-2.51) | | | | |
| Stubborness | | | | | | | 0.0679** | 0.0788*** | 0.0915** | | |
| | | | | | | | (2.53) | (2.65) | (2.32) | | |
| Bankruptcy | | | | | | | -0.00512* | -0.00767** | | | |
| | | | | | | | (-1.95) | (-2.49) | | | |
| Gender | | | | | | | 0.211** | | | | |
| | | | | | | | (2.15) | | | | |
| Energy/Water/Telco | | | | | | | | 1.040** | | | |
| | | | | | | | | (2.46) | | | |
| Trust vs Others | | | | | | | | 0.0386* | | | |
| | | | | | | | | (1.73) | | | |
| Mining | | | | | | | | 0.595** | | | 0.582** |
| | | | | | | | | (1.96) | | | (2.01) |
| Ambiguity Aversion | | | | | | | | -0.0829** | | | |
| | | | | | | | | (-2.52) | | | |
| Building | | | | | | | | | | | -0.568*** |
| | | | | | | | | | | | (-2.98) |

$t$ statistics in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 8.7: Perceived Likelihood of Suffering/Causing Damages.

| | Pr. Suffering Damages | Pr. Causing Damages |
|---|---|---|
| Mining | 0.832*** | |
| | (3.58) | |
| Savings | -0.306*** | -0.159* |
| | (-3.73) | (-1.94) |
| Damaged (last 5 years) | -0.365*** | |
| | (-3.64) | |
| EV/family assets | 0.00407*** | 0.00324** |
| | (3.11) | (2.40) |
| Listed | 0.309** | |
| | (2.05) | |
| Overconfidence | -0.202** | |
| | (-2.40) | |
| Stubborness | -0.0655*** | -0.0596*** |
| | (-3.07) | (-2.78) |
| Bankruptcy | 0.0105*** | 0.00511* |
| | (5.15) | (1.91) |
| Employee | | 0.103*** |
| | | (2.87) |
| Age | | -0.00347** |
| | | (-2.23) |
| Trade | | -0.180* |
| | | (-1.95) |
| Having damaged (last 5 years) | | -0.390*** |
| | | (-3.10) |
| Passion | | -0.0321** |
| | | (-2.35) |
| Bankruptcy (competitors) | | 0.00334* |
| | | (1.72) |
| PersonalDamage | | -0.198*** |
| | | (-2.78) |
| Trust vs Others | | 0.0359** |
| | | (2.30) |
| Production | | -0.215** |
| | | (-2.51) |
| Transportation | | 0.311* |
| | | (1.75) |

$t$ statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

# Appendix Sentiment Analysis

## Appendix A

Table 8.8: OLS regressions results for model 1-6.

| | M1<br>Nasdaq Price | M2<br>Nasdaq Price | M5<br>Nasdaq Price | M3<br>Nasdaq Volume | M4<br>Nasdaq Volume | M6<br>Nasdaq Volume |
|---|---|---|---|---|---|---|
| $Nasdaq\ Price_{t-1}$ | 0.983*** | 0.914*** | 0.949*** | | | |
| | (20.21) | (14.90) | (19.18) | | | |
| $Apple\ sentiment_{t-1}$ | | 35.86 | | | | |
| | | (1.67) | | | | |
| $Google\ sentiment_{t-1}$ | | 8.272 | | | | |
| | | (0.35) | | | | |
| $Facebook\ sentiment_{t-1}$ | | 26.31 | | | | |
| | | (0.92) | | | | |
| $SIT_{t-1}$ | | | 0.0431* | | | -251659.0 |
| | | | (2.11) | | | (-1.95) |
| $Nasdaq\ Volume_{t-1}$ | | | | 0.704*** | 0.738*** | 0.606*** |
| | | | | (6.17) | (6.37) | (5.01) |
| $Apple\ TV_{t-1}$ | | | | | 59933.1 | |
| | | | | | (1.23) | |
| $Google\ TV_{t-1}$ | | | | | -235036.6 | |
| | | | | | (-1.71) | |
| $Facebook\ TV_{t-1}$ | | | | | 35732.7 | |
| | | | | | (0.48) | |

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8.9: Adjusted $R^2$ and root mean square error for all the models.

| | M1 | M2 | M3 | M4 | M5 | M6 |
|---|---|---|---|---|---|---|
| Adj. $R^2$ | 0.9065 | 0.9102 | 0.4690 | 0.4778 | 0.9137 | 0.5029 |
| RMSE | 43.382 | 42.51 | 2.6e+08 | 2.6e+08 | 41.672 | 2.5e+08 |

# Appendix B

Table 8.10: OLS regressions results for model 1-7.

| | (1) Nasdaq Price | (2) Nasdaq Price | (3) Nasdaq Price | (4) Nasdaq Price | (5) Nasdaq Price | (6) Nasdaq Price | (7) Nasdaq Price |
|---|---|---|---|---|---|---|---|
| $Nasdaq\ Price_{t-1}$ | 1.000*** | 1.000*** | 1.000*** | 1.000*** | 1.000*** | 1.000*** | 1.000*** |
| | (11975.78) | (5787.09) | (4635.19) | (5790.16) | (5786.15) | (4653.34) | (4640.86) |
| $Apple\ sentiment_{t-1}$ | | 0.0316*** | | | | | |
| | | (2.68) | | | | | |
| $Google\ sentiment_{t-1}$ | | -0.00736 | | | | | |
| | | (-0.72) | | | | | |
| $Facebook\ sentiment_{t-1}$ | | 0.00401 | | | | | |
| | | (0.42) | | | | | |
| $Apple\ SMMA_{t-1}$ | | | 0.0178 | | | | |
| | | | (0.60) | | | | |
| $Google\ SMMA_{t-1}$ | | | -0.00828 | | | | |
| | | | (-0.31) | | | | |
| $Facebook\ SMMA_{t-1}$ | | | -0.0000612 | | | | |
| | | | (-0.00) | | | | |
| $SIT_{t-1}$ | | | | 0.0123* | | | |
| | | | | (1.90) | | | |
| $SITw_{t-1}$ | | | | | 0.0287* | | |
| | | | | | (1.82) | | |
| $SITma_{t-1}$ | | | | | | 0.0180 | |
| | | | | | | (0.97) | |
| $SITwma_{t-1}$ | | | | | | | 0.0210 |
| | | | | | | | (0.53) |

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8.11: Adjusted $R^2$ and root mean square error for all the models.

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| Adj. $R^2$ | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| RMSE | 1.5641 | 1.3978 | 1.4401 | 1.3983 | 1.3984 | 1.4392 | 1.4394 |

Table 8.12: LPM regressions results for model 8-14.

| | (8) Nasdaq Trend | (9) Nasdaq Trend | (10) Nasdaq Trend | (11) Nasdaq Trend | (12) Nasdaq Trend | (13) Nasdaq Trend | (14) Nasdaq Trend |
|---|---|---|---|---|---|---|---|
| $Nasdaq\ Trend_{t-1}$ | 0.117*** | 0.158*** | 0.155*** | 0.159*** | 0.159*** | 0.156*** | 0.156*** |
| | (16.22) | (8.99) | (7.08) | (9.02) | (9.02) | (7.11) | (7.11) |
| $Apple\ sentiment_{t-1}$ | | 0.00814** | | | | | |
| | | (1.98) | | | | | |
| $Google\ sentiment_{t-1}$ | | 0.000834 | | | | | |
| | | (0.23) | | | | | |
| $Facebook\ sentiment_{t-1}$ | | 0.000173 | | | | | |
| | | (0.05) | | | | | |
| $Apple\ SMMA_{t-1}$ | | | 0.000981 | | | | |
| | | | (0.10) | | | | |
| $Google\ SMMA_{t-1}$ | | | -0.00519 | | | | |
| | | | (-0.58) | | | | |
| $Facebook\ SMMA_{t-1}$ | | | 0.00392 | | | | |
| | | | (0.49) | | | | |
| $SIT_{t-1}$ | | | | 0.00414* | | | |
| | | | | (1.83) | | | |
| $SITw_{t-1}$ | | | | | 0.00973* | | |
| | | | | | (1.77) | | |
| $SITma_{t-1}$ | | | | | | 0.00137 | |
| | | | | | | (0.22) | |
| $SITwma_{t-1}$ | | | | | | | 0.00405 |
| | | | | | | | (0.30) |

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8.13: Adjusted $R^2$ and root mean square error for all the models.

| | M8 | M9 | M10 | M11 | M12 | M13 | M14 |
|---|---|---|---|---|---|---|---|
| Adj. $R^2$ | 0.0135 | 0.0254 | 0.0227 | 0.0257 | 0.0256 | 0.0234 | 0.0234 |
| RMSE | 0.49459 | 0.48774 | 0.48862 | 0.48767 | 0.48769 | 0.48844 | 0.48843 |

# Appendix C

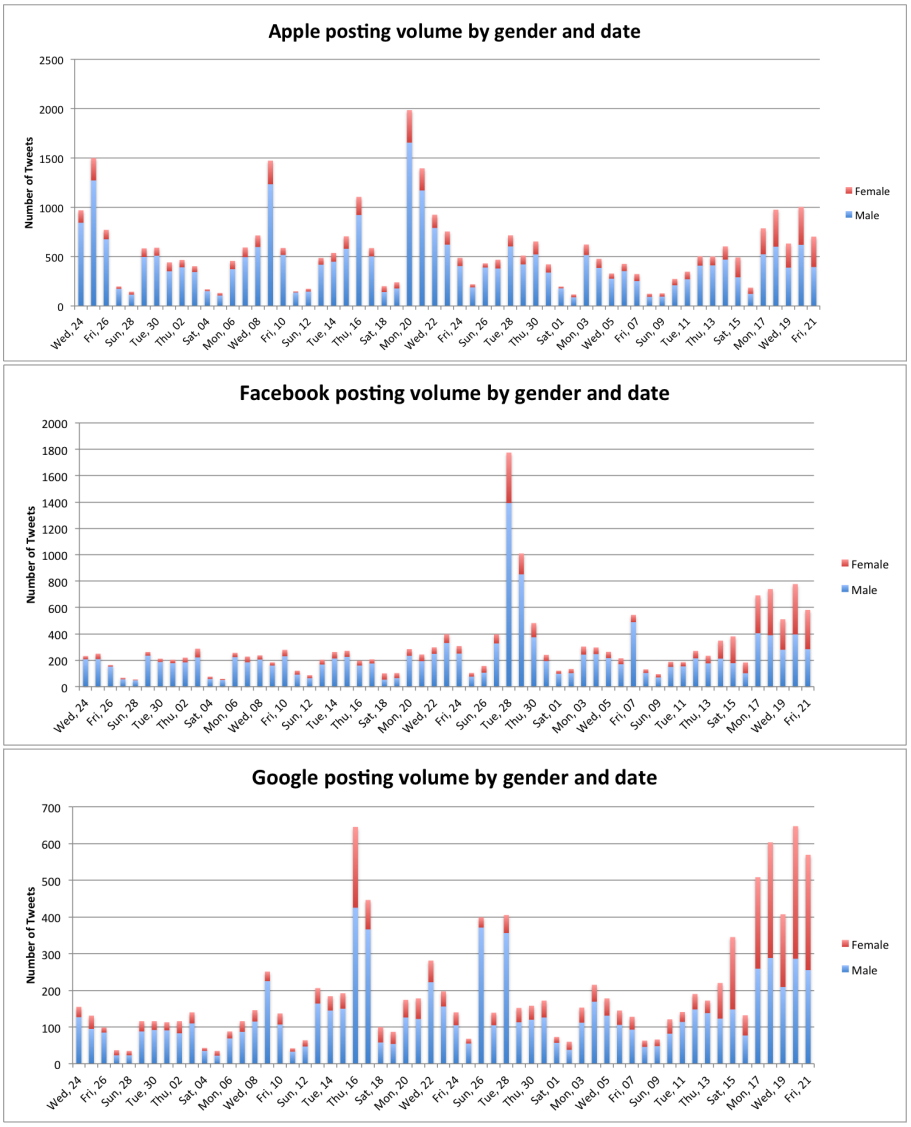



Figure 8.10: Breakdown of posting volume per gender and date for each firm.

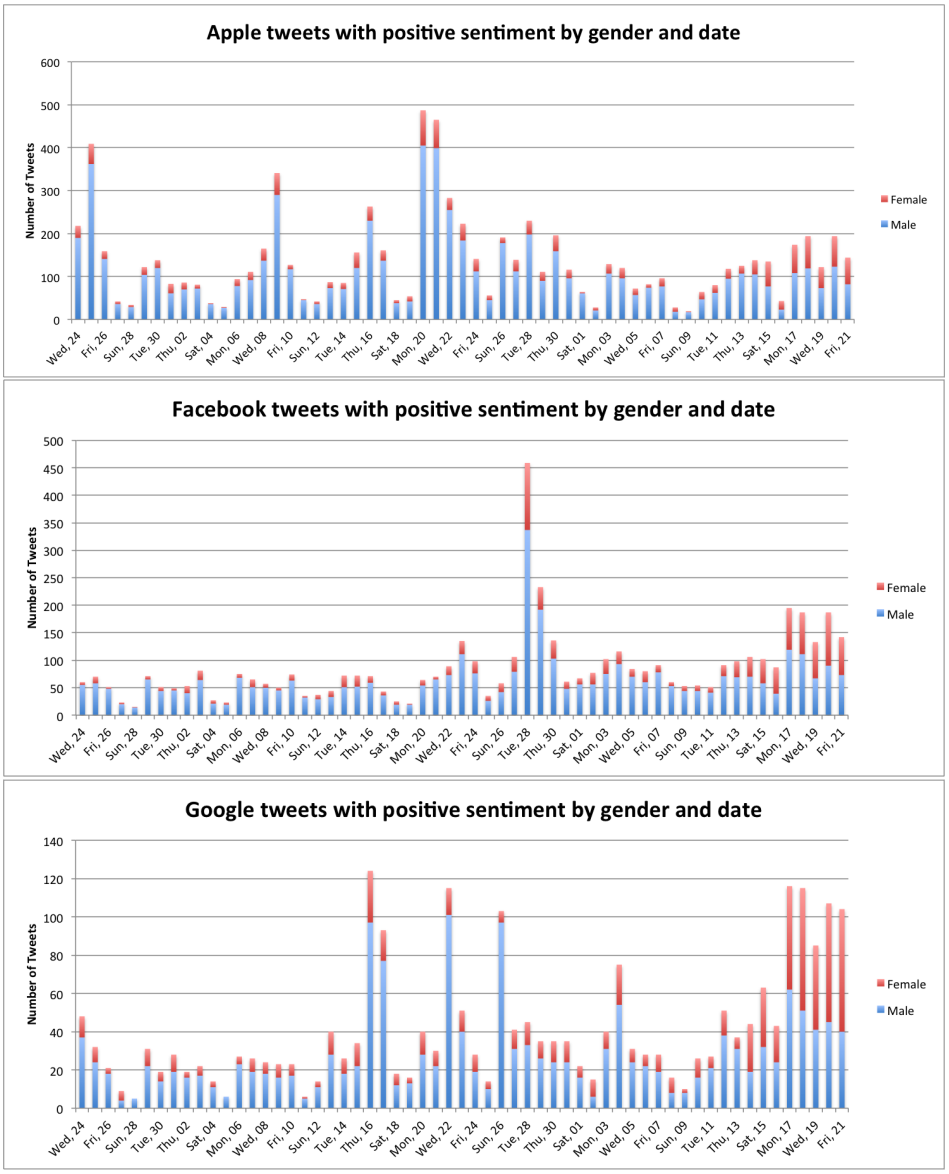



Figure 8.11: Breakdown of positive tweets per gender and date for each firm.

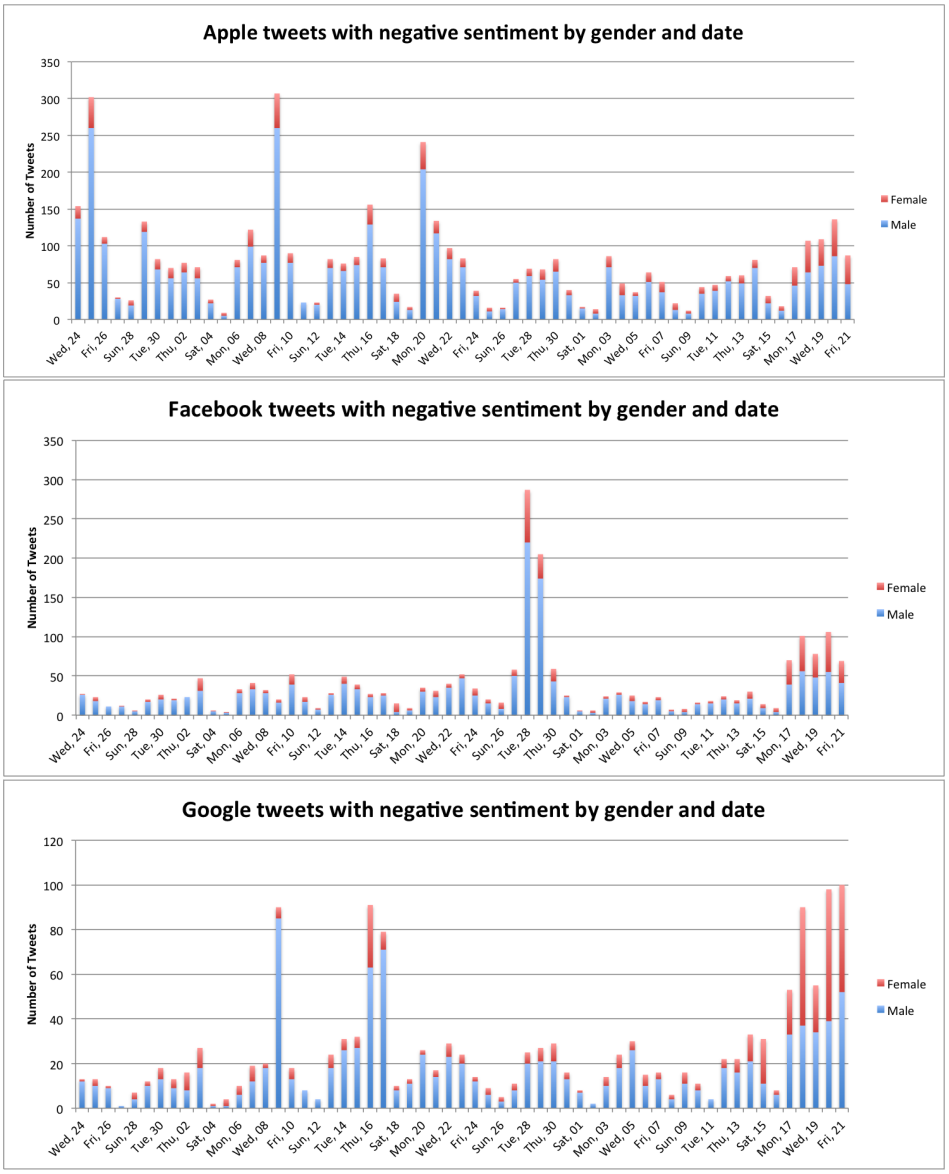



Figure 8.12: Breakdown of negative tweets per gender and date for each firm.

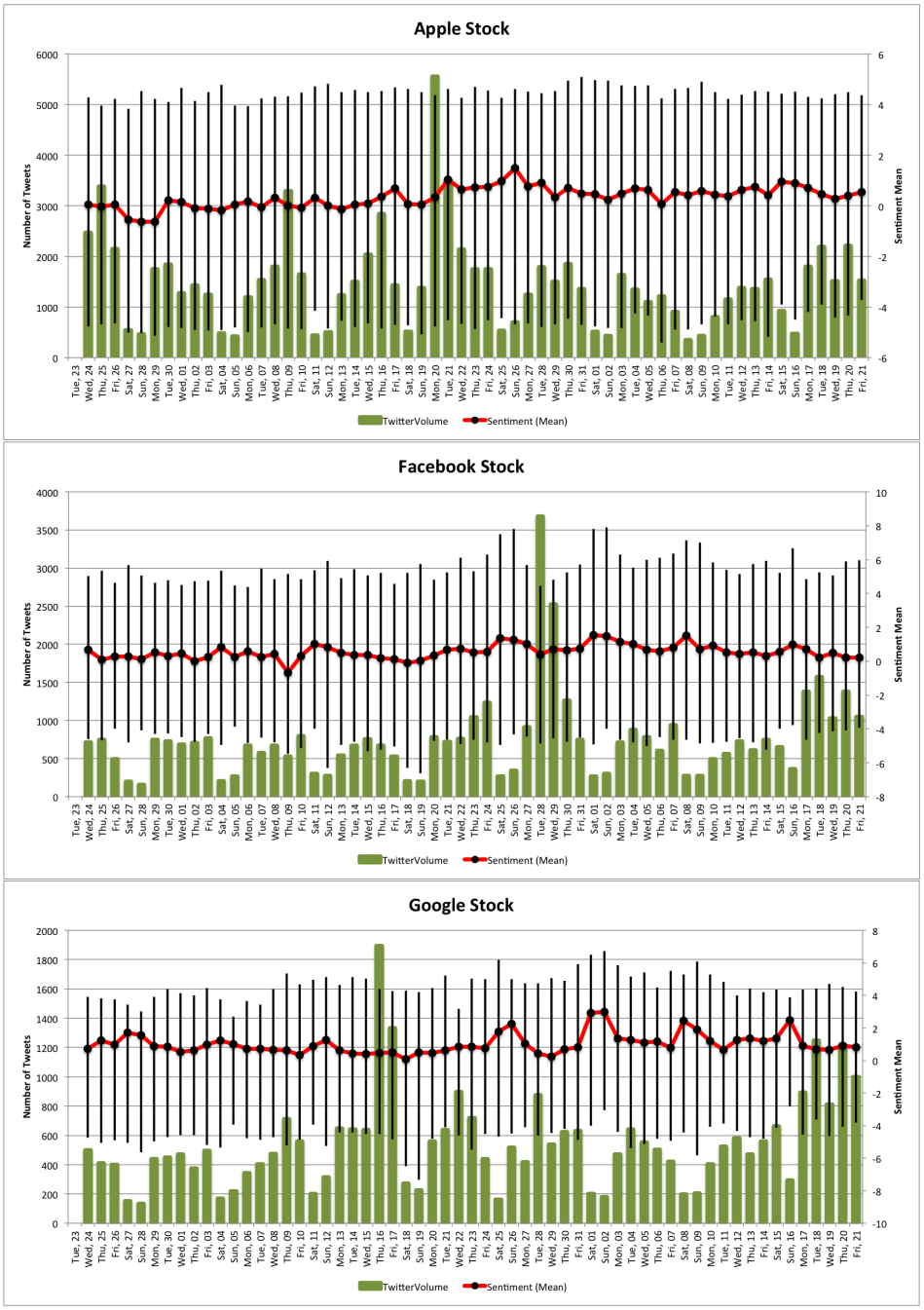
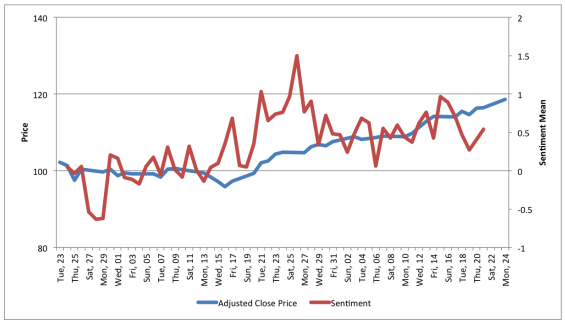



Figure 8.13: Twitter Volume, Mean of positive sentiment, Mean Negative sentiment and Daily sentiment mean for Apple, Facebook and Google.



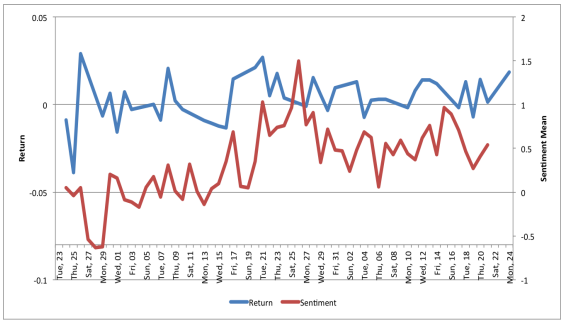




Figure 8.14: Time series for prices (on the left) and returns (on the right) plotted against the average daily sentiment, for Apple.

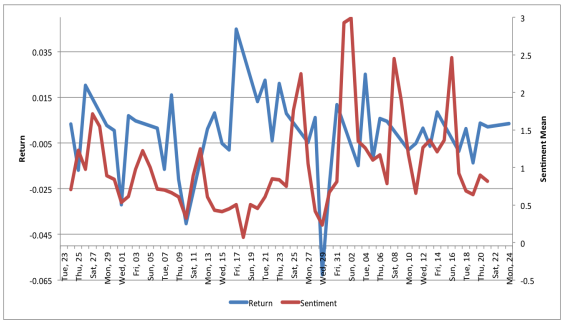




Figure 8.15: Time series for prices (on the left) and returns (on the right) plotted against the average daily sentiment, for Facebook.

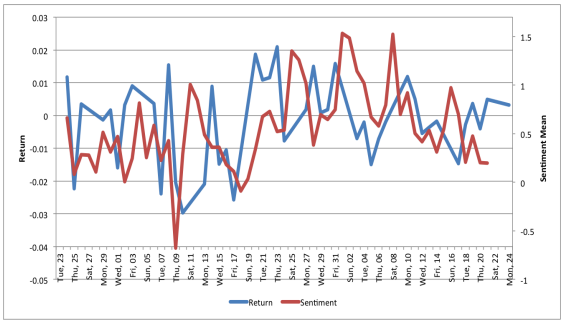




Figure 8.16: Time series for prices (on the left) and returns (on the right) plotted against the average daily sentiment, for Google.

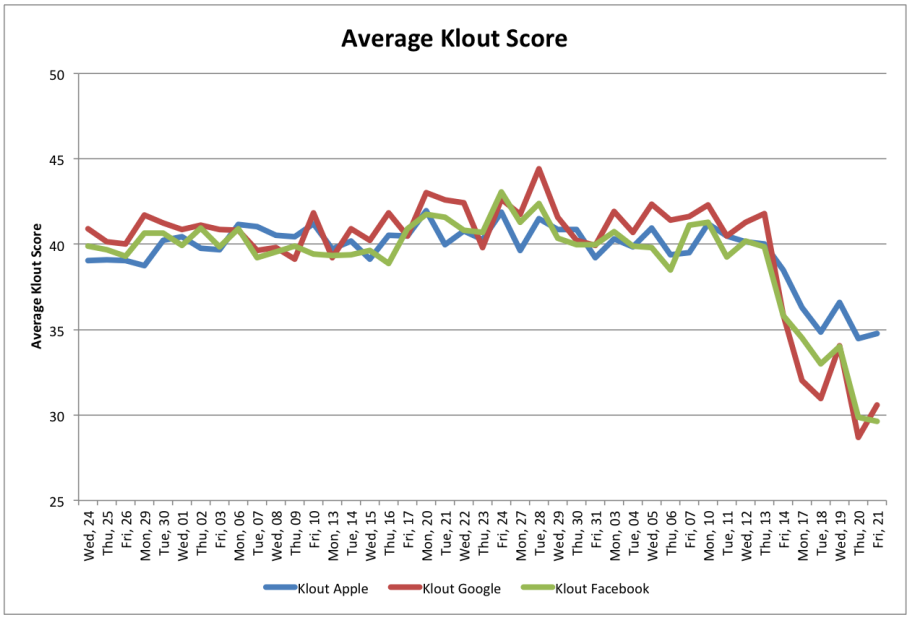



Figure 8.17: Average Klout score for Apple, Google and Facebook.

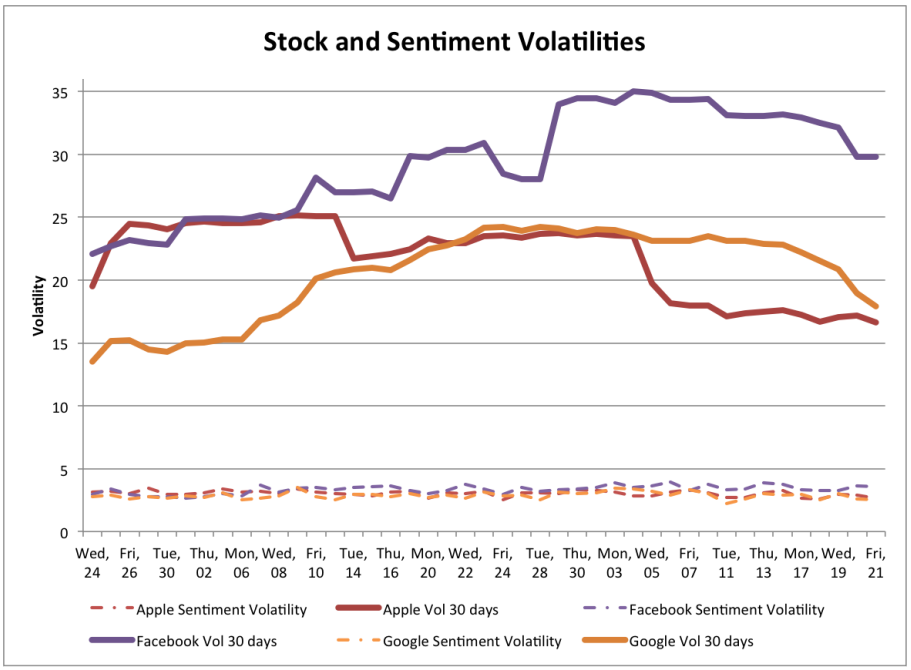



Figure 8.18: Stock volatility (lines) and sentiment score volatility (dashes) for Apple, Google and Facebook.

Table 8.14: Stepwise Variable Selection for the prices.

| | Apple | | Facebook | | Google | |
|---|---|---|---|---|---|---|
| | Price | Trend | Price | Trend | Price | Trend |
| Price | 0.912*** | | 1.005*** | | 0.772*** | |
| | (23.22) | | (36.41) | | (10.90) | |
| BBVR | 0.721* | | | | | |
| | (2.34) | | | | | |
| SV | 1.740* | | 0.974* | | | |
| | (1.76) | | (1.97) | | | |
| BBSp | -2.158* | | | -0.900** | -36.71*** | -0.648*** |
| | (-1.82) | | | (-2.82) | (-4.05) | (-3.73) |
| VolR | 26.83* | 3.480*** | | | | 5.064* |
| | (2.20) | (4.40) | | | | (2.69) |
| BBSR | -8.541* | | | -3.684** | -152.7*** | -2.439** |
| | (-2.27) | | | (-3.43) | (-3.97) | (-3.51) |
| VR | 0.770* | | | | | |
| | (2.44) | | | | | |
| Trend | | -0.357* | | -0.438** | | -0.292* |
| | | (-2.42) | | (-3.05) | | (-2.09) |
| SM | | 0.948*** | | | | |
| | | (3.70) | | | | |
| SR | | -0.0635* | 0.811* | | | 0.0185* |
| | | (-2.30) | (2.24) | | | (1.78) |
| BBSperR | | | -0.0474*** | | | -0.0387* |
| | | | (-3.67) | | | (-1.78) |
| BBSn | | | 0.785* | -0.351* | -32.14** | |
| | | | (1.71) | (-2.36) | (-3.45) | |
| TV | | | -0.00113*** | -0.000247* | -0.00989** | -0.00107*** |
| | | | (-3.89) | (-2.02) | (-2.78) | (-3.71) |
| BBSper | | | 0.00410** | | | |
| | | | (3.08) | | | |
| TVMA | | | | | | 0.00188*** |
| | | | | | | (4.30) |
| $R^2$ | 0.999 | 0.721 | 0.998 | 0.719 | 0.999 | 0.768 |
| *RMSE* | 1.046 | 0.429 | 0.996 | 0.436 | 5.759 | 0.387 |

Every indipendent variable as a lag 1 with respect to the dependent one. t statistics in parentheses.
* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

Table 8.15: Stepwise Variable Selection for the returns.

| | Apple | | Facebook | | Google | |
|---|---|---|---|---|---|---|
| | Return | Trend | Return | Trend | Return | Trend |
| Return | -0.371* | | | | -0.274* | |
| | (-2.62) | | | | (-2.24) | |
| SM | 0.0202** | | | -0.616* | 0.0236* | |
| | (3.28) | | | (-2.50) | (2.30) | |
| SR | -0.00118* | | 0.0135** | 0.341* | | |
| | (-1.74) | | (2.73) | (2.13) | | |
| BBSperR | 0.000316* | | -0.000654** | -0.00437* | | |
| | (2.05) | | (-3.51) | (-1.79) | | |
| BBSn | 0.00239* | | 0.00346** | | | |
| | (2.60) | | (2.77) | | | |
| VR | 0.00835** | | | | | -0.738** |
| | (2.81) | | | | | (-3.35) |
| BBSper | -0.0000243* | | 0.0000598** | | | 0.00682* |
| | (-1.95) | | (3.17) | | | (2.02) |
| Trend | | -0.386* | | -0.393* | | -0.322* |
| | | (-2.67) | | (-2.67) | | (-2.59) |
| SV | | 0.203*** | | 0.238** | 0.0208** | |
| | | (6.52) | | (2.96) | (2.91) | |
| TVMA | | | | -0.000327* | 0.0000579*** | -0.00102* |
| | | | | (-1.84) | (4.67) | (-2.07) |
| BBSp | | | | | -0.0153*** | -0.362*** |
| | | | | | (-3.80) | (-3.79) |
| BBVR | | | | | -0.00967* | |
| | | | | | (-2.09) | |
| TV | | | | | -0.0000327*** | 0.00126** |
| | | | | | (-4.00) | (2.90) |
| Klout | | | | | | 0.0689*** |
| | | | | | | (5.59) |
| $R^2$ | 0.444 | 0.538 | 0.319 | 0.521 | 0.544 | 0.676 |
| *RMSE* | 0.0097 | 0.468 | 0.146 | 0.399 | 0.109 | 0.372 |

Every indipendent variable as a lag 1 with respect to the dependent one. t statistics in parentheses.
* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

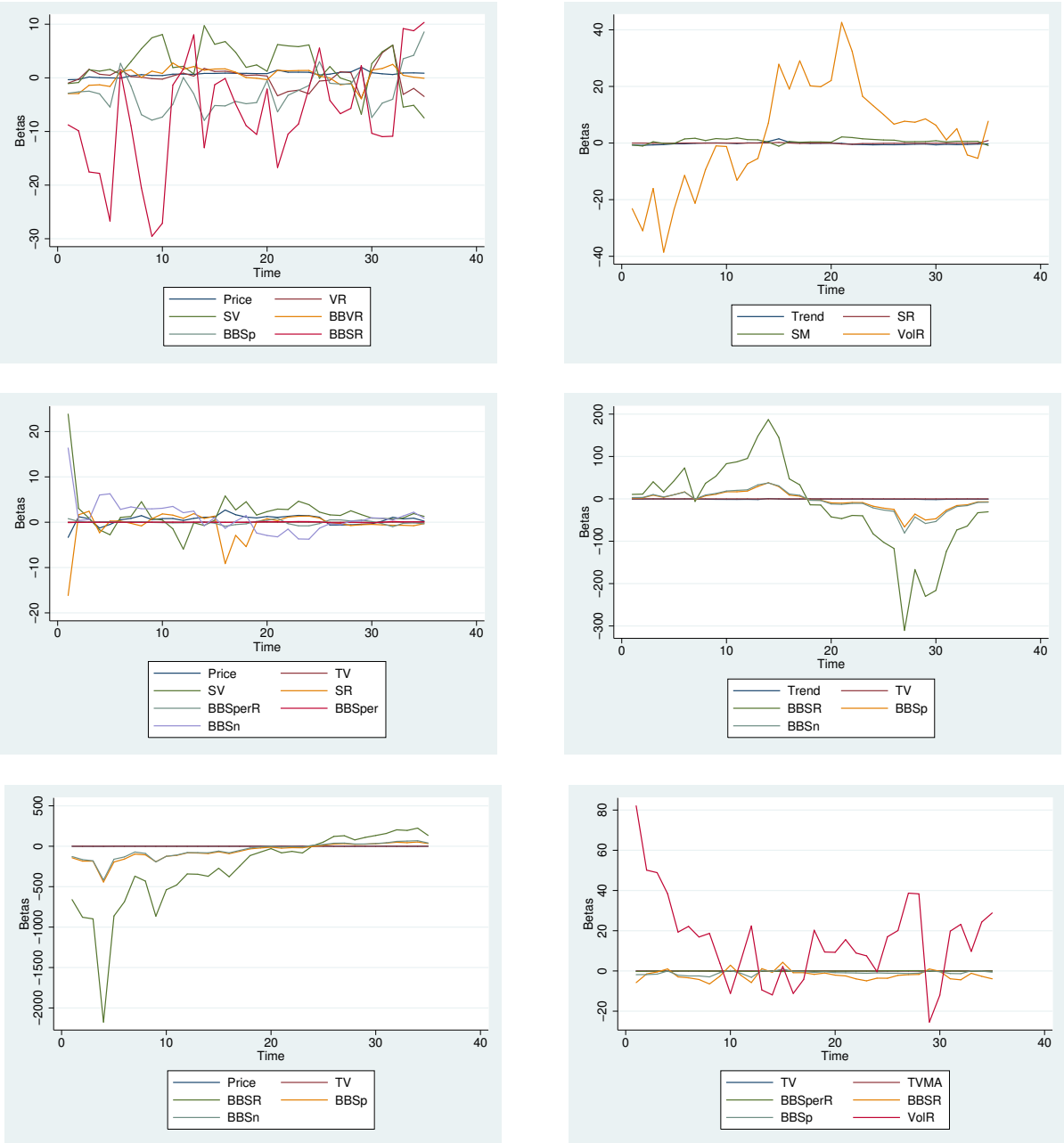



Figure 8.19: Rolling window for both prices value (left) and direction forecasting (right) respectively for Apple, Facebook and Google.

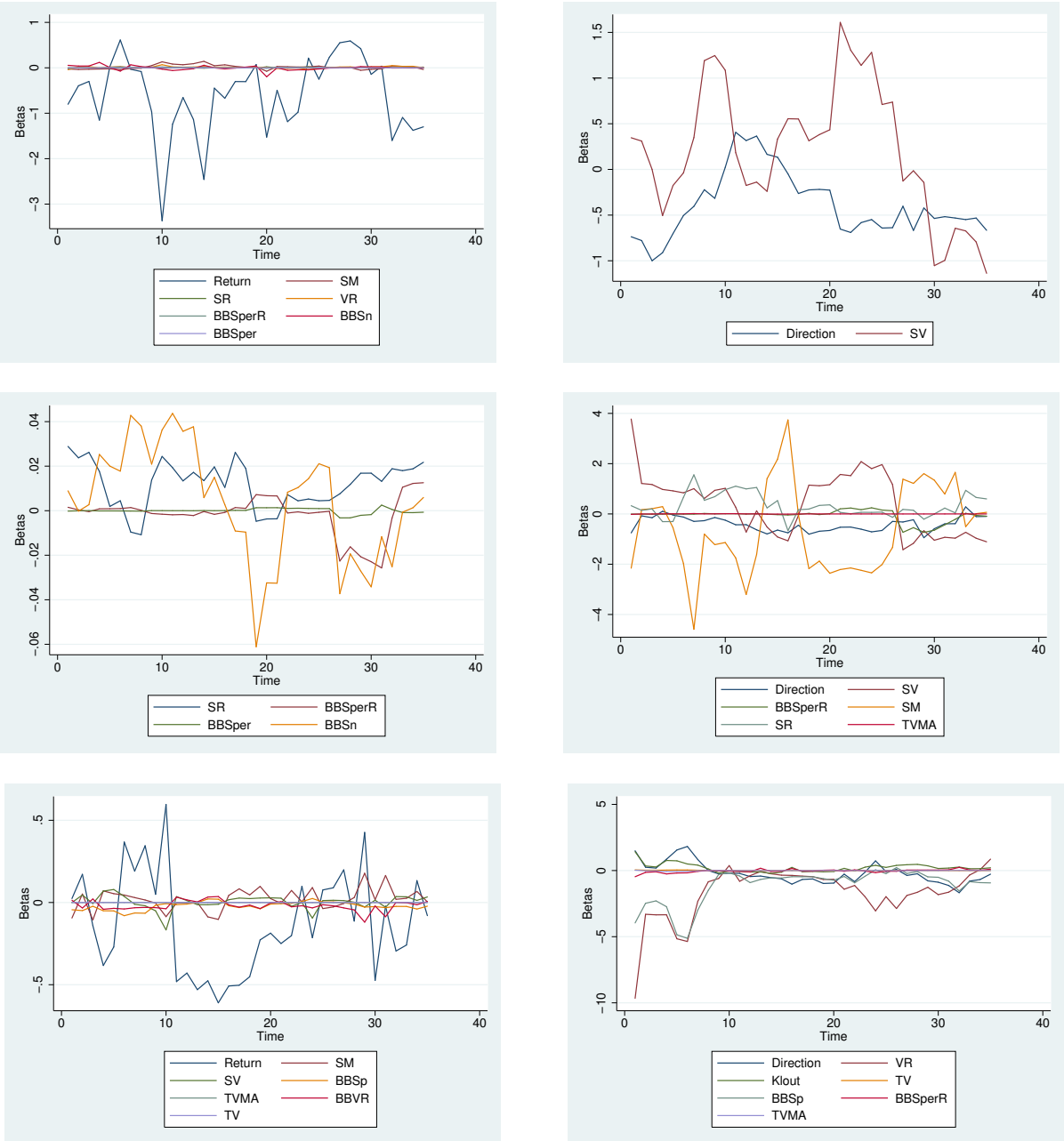



Figure 8.20: Rolling window for both returns value (left) and direction forecasting (right) respectively for Apple, Facebook and Google.

# Appendix D

Table 8.16: Stepwise Variable Selection for the high-frequency prices and trends.

| | Apple | | Facebook | | Google | |
|---|---|---|---|---|---|---|
| | Price | Trend | Price | Trend | Price | Trend |
| Price | 1.000*** | | 1.000*** | | 1.000*** | |
| | (22483.99) | | (4341.11) | | (13475.16) | |
| BBVR | 0.00504* | | 0.0141* | | | 0.0463* |
| | (2.01) | | (1.69) | | | (1.72) |
| BBVn | 0.00319* | | | | | 0.0850** |
| | (2.18) | | | | | (2.72) |
| TV | -0.00218** | | | | | |
| | (-2.97) | | | | | |
| Trend | | 0.395*** | | 0.486*** | | 0.413*** |
| | | (17.36) | | (12.52) | | (5.14) |
| Klout | | 0.00836*** | | 0.00862*** | | |
| | | (10.82) | | (9.47) | | |
| SR | | 0.00860* | -0.0115* | | | |
| | | (1.99) | (-2.28) | | | |
| BBSp | | 0.0160*** | | | | |
| | | (3.68) | | | | |
| BBSn | | -0.00739* | 0.00748* | | | -0.0448** |
| | | (-1.65) | (2.47) | | | (-3.21) |
| BBVp | | 0.0115* | | | | |
| | | (2.03) | | | | |
| SMMA | | | 0.00875* | | | |
| | | | (2.26) | | | |
| TVMA | | | | 0.00615* | | |
| | | | | (2.27) | | |
| BBSR | | | | -0.0354** | | -0.0666** |
| | | | | (-2.80) | | (-3.11) |
| $R^2$ | 1.000 | 0.810 | 1.000 | 0.868 | 1.000 | 0.842 |

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$

# Appendix E

Table 8.17: OLS regressions results for Nasdaq model 1-7.

| | (1) Call Price | (2) Call Price | (3) Call Price | (4) Call Price | (5) Call Price | (6) Call Price | (7) Call Price |
|---|---|---|---|---|---|---|---|
| $Nasdaq\ Price_{t-1}$ | 0.998*** | 0.998*** | 0.997*** | 0.998*** | 0.998*** | 0.997*** | 0.997*** |
| | (2328.36) | (923.86) | (655.98) | (923.50) | (923.14) | (656.77) | (656.59) |
| $Apple\ sentiment_{t-1}$ | | 0.149 | | | | | |
| | | (0.73) | | | | | |
| $Google\ sentiment_{t-1}$ | | -0.382** | | | | | |
| | | (-2.15) | | | | | |
| $Facebook\ sentiment_{t-1}$ | | 0.288* | | | | | |
| | | (1.75) | | | | | |
| $Apple\ SMMA_{t-1}$ | | | -0.256 | | | | |
| | | | (-0.45) | | | | |
| $Google\ SMMA_{t-1}$ | | | 0.214 | | | | |
| | | | (0.42) | | | | |
| $Facebook\ SMMA_{t-1}$ | | | 0.138 | | | | |
| | | | (0.31) | | | | |
| $SIT_{t-1}$ | | | | 0.0132 | | | |
| | | | | (0.12) | | | |
| $SITw_{t-1}$ | | | | | 0.0425 | | |
| | | | | | (0.16) | | |
| $SITma_{t-1}$ | | | | | | 0.0690 | |
| | | | | | | (0.19) | |
| $SITwma_{t-1}$ | | | | | | | 0.103 |
| | | | | | | | (0.13) |

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8.18: Adjusted $R^2$ and root mean square error for all the models.

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| Adj. $R^2$ | 0.9965 | 0.9963 | 0.9953 | 0.9963 | 0.9963 | 0.9953 | 0.9953 |
| RMSE | 23.357 | 24.231 | 27.733 | 24.249 | 24.249 | 27.722 | 27.722 |

Table 8.19: LPM regressions results for model 8-14.

| | (1) Call Trend | (2) Call Trend | (3) Call Trend | (4) Call Trend | (5) Call Trend | (6) Call Trend | (7) Call Trend |
|---|---|---|---|---|---|---|---|
| $Nasdaq\ Trend_{t-1}$ | 0.679*** | 0.703*** | 0.713*** | 0.702*** | 0.703*** | 0.713*** | 0.713*** |
| | (127.66) | (55.21) | (45.79) | (55.21) | (55.24) | (45.77) | (45.76) |
| $Apple\ sentiment_{t-1}$ | | 0.00217 | | | | | |
| | | (0.72) | | | | | |
| $Google\ sentiment_{t-1}$ | | 0.00315 | | | | | |
| | | (1.21) | | | | | |
| $Facebook\ sentiment_{t-1}$ | | 0.00315 | | | | | |
| | | (1.31) | | | | | |
| $Apple\ SMMA_{t-1}$ | | | -0.00548 | | | | |
| | | | (-0.76) | | | | |
| $Google\ SMMA_{t-1}$ | | | 0.0152** | | | | |
| | | | (2.35) | | | | |
| $Facebook\ SMMA_{t-1}$ | | | -0.000516 | | | | |
| | | | (-0.09) | | | | |
| $SIT_{t-1}$ | | | | 0.00276* | | | |
| | | | | (1.67) | | | |
| $SITw_{t-1}$ | | | | | 0.00877** | | |
| | | | | | (2.19) | | |
| $SITma_{t-1}$ | | | | | | 0.00171 | |
| | | | | | | (0.38) | |
| $SITwma_{t-1}$ | | | | | | | 0.00790 |
| | | | | | | | (0.81) |

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8.20: Adjusted $R^2$ and root mean square error for all the models.

| | M8 | M9 | M10 | M11 | M12 | M13 | M14 |
|---|---|---|---|---|---|---|---|
| Adj. $R^2$ | 0.4611 | 0.4933 | 0.5091 | 0.4932 | 0.4935 | 0.5083 | 0.5084 |
| RMSE | 0.3666 | 0.3558 | 0.3504 | 0.3558 | 0.3557 | 0.3506 | 0.3506 |

Table 8.21: Stepwise Variable Selection for the high-frequency prices and trends.

| | Apple | | Facebook | | Google | |
|---|---|---|---|---|---|---|
| | Price | Trend | Price | Trend | Price | Trend |
| Price | 1.000*** | | 0.443*** | | 0.900*** | |
| | (32313.78) | | (7.11) | | (27.97) | |
| Trend | | 0.403*** | | 0.624*** | | 0.459*** |
| | | (17.79) | | (17.50) | | (5.93) |
| Klout | | 0.00859*** | 0.711*** | 0.00713*** | 0.762* | 0.00532* |
| | | (11.35) | (9.10) | (8.67) | (2.63) | (1.94) |
| SR | | 0.00910* | | | -7.883** | |
| | | (2.11) | | | (-2.92) | |
| BBSp | | 0.0166*** | | | -3.206* | |
| | | (3.81) | | | (-1.97) | |
| BBSn | | -0.00778* | -4.147*** | | -5.151* | -0.0378* |
| | | (-1.74) | (-8.59) | | (-2.31) | (-1.95) |
| SM | | | -1.970* | | | |
| | | | (-2.56) | | | |
| BBSR | | | -7.655*** | -0.0207* | -13.20* | -0.0455* |
| | | | (-8.31) | (-1.85) | (-2.38) | (-1.75) |
| $R^2$ | 1.000 | 0.810 | 0.987 | 0.895 | 0.999 | 0.845 |

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.01$, *** $p < 0.001$