Evaluating Econometric Models and Expert Intuition

ISBN: 978 90 361 0291 9

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 530 of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Evaluating Econometric Models and Expert Intuition

Het evalueren van econometrische modellen en intuïtie van experts

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 10 mei 2012 om 15.30 uur

door

Rianne Legerstee
geboren te Rotterdam

Promotiecommissie

Promotor:	Prof.dr. Ph.H.B.F. Franses
Overige leden:	Prof.dr. R. Paap
	Prof.dr. D.J.C. van Dijk
	Prof.dr. A. Timmermann

Acknowledgements

When I decided to pursue a PhD, I knew it would be challenging. I reasoned that I needed a lot of perseverance and discipline to obtain a PhD, but I always considered myself blessed with those qualities. But as I now know, I can also be quite naive and lighthearted, because it turned out to be far more challenging than I could have ever imagined.

In retrospect, I am not sure I would do it again. Every day was a struggle and the days I was satisfied with what I had done were very rare. Considering quitting did not only cross my mind one month after starting my PhD, but also after one year, two years and even when I had only half a year to go. But in the end, I am so glad that I started and did not quit. I learned incredibly much: about econometrics, statistics and forecasting, about economics, about teaching and presenting, about doing research and the academic world, about the English language and last but not least about myself. I truly feel that I have grown and that I gained confidence in whatever I will do next. Furthermore, and not less important, I had a lot of fun too. For example during the trips abroad for conferences and during the endless (lunch) breaks with friends and colleagues that became friends.

That this journey was so heavy, but at the same time so rewarding for me, means that I am truly indebted to the people surrounding me in work life and in private. No perseverance and discipline would have been enough, if it was not for the following people.

First and foremost I would like to thank Philip Hans Franses, my supervisor, whose

bottomless enthusiasm, positivity and creativity have been a true motivation and inspiration to me. I believe that finishing your PhD in time is for a large part influenced by your supervisor and while I am writing this I have still a few days left of the four years scheduled for it. Philip Hans, thank you for always finding time in your busy schedule to meet or email me, for putting so much trust in me and for never hesitating to show me that trust.

Second, I would like to thank Richard Paap. As co-author of two of the chapters in this thesis he contributed significantly. Richard, thanks for your research ideas and help on methodology whenever I needed it.

Many thanks also go to Christiaan Heij and Adriana Gabor. I truly enjoyed assisting you both twice in teaching the courses Econometrie 1 and Kansrekening. Having such great examples and receiving the right amount of responsibility made that I have learned a lot from it.

The Econometric Institute and the Tinbergen Institute have provided a very nice work environment. I am grateful to all the staff members at these institutes for their friendly support throughout the years.

I am also grateful to the two remaining members of the inner committee, Dick van Dijk and Allan Timmermann, for taking the time to review this thesis and for providing their useful comments.

Starting to cross the line between work life and private life, I would like to thank my fellow PhD students Jorn, Jeanine, Anne and Sjoerd. Sjoerd for the long chats in the late afternoon and the pool nights together with Anne and Jorn. Anne for being such a sociable, cheerful, funny and noisy roommate. In my last years of being a PhD student I could not have wished for a more suitable roommate. Jeanine for becoming my one and only true girl-PhD-friend. I felt you were the most alike person I could find on the 9th floor and sometimes that is all one could wish for. Jorn, since the very first beginning of my PhD trajectory, you have been a true friend and especially in the first two years you were one of the few people I could talk to and who would understand my troubles, complaints and struggles as a PhD student. And thanks in advance Jeanine

and Jorn, for standing beside me as my paranymphs at the defence ceremony of this thesis!

Furthermore, I would like to thank my non-PhD friends, whom, almost without exception, I all know since high school, which by itself already tells a lot. Thanks Caroline, Stephanie, Rianne, Fréderike, Rik, Renée, Julia, Jurjen, Hans, Maimouna, Rogier, Abdi, Ryan, Marlous, Daan, Hubert, Egle, Hedwig and Raymond for your interest, distraction and just for being friends. The holidays to for example Portugal, Brussel and Brugge, the numerous dinners, movies and chill-nights, the book club evenings and reading the books for it and more of that stuff all played their part in preventing me from going crazy sometimes. Special thanks go to Rianne and Julia for also providing the necessary distraction during work time.

I am not only blessed with great friends, but also with a great family. Aunts, uncles, nieces and nephews, and especially grandparents, of whom sadly enough only one is still around, all deserve gratitude for their love and interest. Thanks to my family in-law for their support and concern, for always providing help wherever needed and for the many nice, festive, excessive and nourishing meals. Thanks Marcel and Zalihe, for being such a good brother and sister-in-law and for cheering up many Friday nights, as well as many other (family) occasions. Marcel, for being one of the first in my life who taught me to be skeptical and for always being available to discuss and evaluate every aspect of life. Mom and dad, thanks for being the parents all one could wish for, for being proud no matter what, for always supporting my decisions in life, for always being available whenever I need a listening ear, for raising me to what I am today.

Finally, Tolga, you are essential in my life and were thereby essential in the development of this thesis. Dear Tolga, I want you to know that your endless love, support and admiration are noticed and truly appreciated and I want you to know that I love, support and admire you just as much or even more.

Rianne Legerstee

Rotterdam, February 2012

Contents

Chapter 1

Introduction and outline

An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination. (Andrew Lang)

1.1 Introduction

Numerous economic forecasts are produced every day. Macroeconomic forecasts on, for example, Gross Domestic Product, inflation and housing prices are released by governmental institutions all over the world, such as the Federal Reserve Bank, the European Central Bank and the Netherlands Bureau for Economic Policy Analysis (CPB). Also non-governmental organizations, such as banks, insurance companies, consultancy firms and universities produce forecasts for the same macroeconomic variables, sometimes bundled in anonymous surveys, such as the Survey of Professional Forecasters in the USA, and sometimes quoted in high impact media, not anonymously. Furthermore, there are many microeconomic forecasts, like forecasts at the business level, on, for example, expected sales in order to plan production and the amount of products to be kept in stock.

Although we see these forecasts every day, use them, count on them and often evaluate their accuracy afterwards, for people other than the forecasters themselves

very little is known about how these forecasts are created. We could assume that the forecasts are created with the use of an econometric model, based on past data of the target variable and on other possibly explanatory variables. However, to what extent such an econometric model is actually used is usually not clear. It is conceivable that the final forecast quoted by the forecaster is for 100% based on an econometric model, but it is also possible that the forecast is a purely intuitive forecast, based on no computations at all and for 0% based on an econometric model. And, it can be anything in between.

Professional economic forecasts are often not for 100% the outcome of an econometric model. Often, when a statistical model forecast is produced an expert (with domain knowledge) intervenes and adjusts the forecast on grounds that are not fully systematic and often not documented. This happens both in microeconomic forecasting (see for example Mathews and Diamantopoulos, 1986; Fildes et al., 2009) and in macroeconomic forecasting (see for example McNees, 1990; Turner, 1990). In recent work presented in Franses et al. (2011a) it is shown that the CPB uses a large macroeconomic model to create forecasts of various important macroeconomic variables, but that the outcomes of this model are almost invariably adjusted by experts. Moreover, only the final expert forecasts are made available to the general public. These types of adjustments are often called judgmental adjustments and can be viewed as the result of expert intuition.

Forecasters should not rely only on econometric models, past and current data and statistics, and this view is implied by the quote at the beginning of this chapter. Indeed, the rationale for judgmental adjustments is that experts can take into account exceptional circumstances that cannot be captured by the statistical model (Goodwin, 2000). Quite some literature exists on the forecast accuracy of these by judgment adjusted forecasts, but conclusions are mixed. In some instances the adjustments seem to improve the accuracy of the forecasts (see for example Fildes et al., 2009; Franses et al., 2011a), but in other situations the adjustments sometimes dramatically deteriorate forecasts, decreasing overall forecast accuracy (Franses and Legerstee, 2010).

However, before using the forecasts and analyzing their accuracy, more should be known about what it is that the experts do when they decide to somehow deviate from econometric model forecasts. How are these final forecasts created? What is the extent to which the statistical model is used and how is intuition integrated? What are the motivations of the expert for creating the forecasts? Not only the econometric models and intuition about the value of the variable to be forecasted may influence the value of the forecasts, also, for example, the reputation of the forecasters (Laster et al., 1999), forecast uncertainty (Lahiri and Sheng, 2010), the way valuable time and resources are used (Mankiw et al., 2003) and the necessity to prevent being out off stock from a supply chain management point of view (see Chapter 4). If we have answers to such questions, the forecasts can be placed in the right perspective and it would provide tools to evaluate the quality of the forecasts. For example, by linking what it is that experts do to forecast accuracy we can say more about good and bad adjustment practices, which is the focus of Chapter 2.

Also Chapter 4 is directly concerned with finding out what it is that the experts do. Is the loss function of experts symmetric, that is, is every type of inaccuracy in forecasting, positive and negative, equally important to the forecaster? Because if it is not, this should be taken into account when evaluating the accuracy of the forecasts. If a symmetric loss function is assumed and forecasts are evaluated on that assumption, the forecasting practices might turn out to be poor, while the underlying reason might be that the forecaster had different objectives.

We also do not know much about the ability of experts to adapt to new insights. If we would know what good and bad adjustment practices are, does proper feedback results in improved adjusted forecasts? The feedback literature is largely based on laboratory experiments (Lawrence et al., 2006), but in Chapter 3 we analyze a unique natural experiment to find out what the impact is of feedback on judgmental adjustments.

Because there are often a large amount of forecasts available from different experts for the same event, especially in macroeconomics, another question is how to use such

a plethora of forecasts. Most research has focused on using the mean or median of the forecasts (Armstrong, 2001b; Elliott and Timmermann, 2005), but as argued above, we do not know how all these forecasts are created. Various theories exist on what factors might influence the forecasts and on how these factors affect the forecasts. It is argued in Chapter 5 that, as a result of these different kind of influences, disagreement between forecasters has predictive value too.

1.2 Outline

The chapters of this thesis are self-contained and can thus be read independently. All chapters are about forecasting situations in which econometric models and expert intuition are involved, although the last chapter is somewhat different from the other three chapters. The first three chapters are about what it is that experts do when they adjust statistical model forecasts and what might possibly improve that adjustment behavior. Although the techniques described and the results obtained in these chapters can be applied and might be generalized to other micro- and macroeconomic data sets, we applied the techniques to and obtained the results for stock keeping unit (SKU) level sales data. The empirical sections of these chapters are all based on (parts of) the same unique data set, obtained from a large pharmaceutical company with its headquarter in The Netherlands and local offices in various countries.

The final chapter is devoted to research on how to make use in an optimal way of multiple forecasts produced by multiple experts for one and the same event. The situation in which multiples of such forecasts are available occurs mainly in macroeconomic forecasting and the empirical part of this research is based on forecasts from the Survey of Professional Forecasters (SPF), which are publicly available.

In more detail the outline of this thesis is as follows.

Chapter 2 is based on Legerstee et al. (2011). In the situation as analyzed in this chapter, experts can rely on statistical model forecasts when creating their own forecasts and it is not documented what the experts do. In this chapter we focus on three

questions, which we try to answer given the availability of expert forecasts and model forecasts. First, is the expert forecast related to the model forecast and how? Second, how is this potential relation influenced by other factors? Third, how does this relation influence forecast accuracy?

We propose a new and innovative two-level Hierarchical Bayes model to answer these questions. We apply our proposed methodology to the large data set of forecasts and realizations of SKU-level sales data from the pharmaceutical company. We find that expert forecasts can depend on model forecasts in a variety of ways. Average sales levels, sales volatility, and the forecast horizon influence this dependence. We also demonstrate that theoretical implications of expert behavior on forecast accuracy are reflected in the empirical data. In general, experts who follow some simple rules, which optimize forecast performance under optimal circumstances, outperform the model forecasts in our data set.

Chapter 3 is based on Legerstee and Franses (2011). Here we again analyze the behavior of experts who quote forecasts for monthly SKU-level sales data, where we now compare data before and after the moment that experts received different kinds of feedback on their behavior. We have data for 21 experts located in as many countries who make SKU-level forecasts for a variety of pharmaceutical products for October 2006 to September 2007. We study the behavior of the experts by comparing their forecasts with those from an automated statistical program, and we report the forecast accuracy over these 12 months. In September 2007 these experts were given feedback on their behavior and they received a training at the headquarters' office, where specific attention was given to the ins and outs of the statistical program used to create the model forecasts. Next, we study the behavior of the experts for the three months after the training session, that is, October 2007 to December 2007. Our main conclusion is that in the second period the experts' forecasts deviated less from the statistical forecasts and that their accuracy improved substantially.

Chapter 4 is based on Franses et al. (2011b). In this chapter a new and simple methodology is proposed to estimate the loss function associated with experts' fore-

casts. Under the assumption of conditional normality of the data and the forecast distribution, the asymmetry parameter of the lin-lin and linex loss function can easily be estimated using a linear regression. This regression also provides an estimate for potential systematic bias in the forecasts of the expert. The residuals of the regression are the input for a test for the validity of the normality assumption.

We apply our approach again to the data set of SKU-level sales forecasts made by experts and we compare the outcomes with those for statistical model-based forecasts of the same sales data. We find substantial evidence for asymmetry in the loss functions of the experts, with underprediction penalized more than overprediction.

Chapter 5 is based on Legerstee and Franses (2010). Forecasts from various experts are often used in forecasting models and in forecast combinations by taking the mean or median of the survey data. In the present study we take a different stance as we examine the predictive power of potential disagreement amongst forecasters. The premise is that the degree of disagreement could signal upcoming structural or temporal changes in an economic process or in the predictive power of the survey forecasts.

In our empirical work, we examine a variety of macroeconomic variables, and we use different kinds of measurements for the degree of disagreement, together with measures for location of the survey data and autoregressive components. Forecasts from simple linear models and forecasts from Markov regime-switching models with constant and with time-varying transition probabilities are constructed in real-time and compared on forecast accuracy. Our main finding is that disagreement can indeed have predictive value, especially when used in Markov regime-switching models.

Chapter 2

Do experts incorporate statistical model forecasts and should they?

Joint work with Philip Hans Franses and Richard Paap.

2.1 Introduction

In many forecasting situations there are two forecasts available. First, a statistical model is used to produce a model forecast, which is based on available (past) data and possibly other variables. Second, an expert creates an expert forecast. Usually it is assumed that an expert first looks at the model-based forecast and then decides to make an adjustment and, if so, decides on the size of the adjustment.

The literature on judgmental adjustments to model forecasts is extensive and growing, in particular due to the fact that more detailed factual data become available. Most literature focuses on the quality improvement or deterioration caused by the adjustments. In theory, judgmental adjustments by experts could make expert forecasts more accurate than model-based forecasts. One of the main justifications for judgmental adjustment is that experts can recognize rare events that might influence the variable under consideration but that are too irregular to be incorporated in statistical models

(Goodwin, 2000).

A few of the earlier studies on forecast adjustment using actual case study data are Mathews and Diamantopoulos (1986, 1989, 1990, 1992, 1994), Diamantopoulos and Mathews (1989) and Blattberg and Hoch (1990). In general, these authors conclude that forecast adjustments lead to more accurate forecasts on average. More recent work by Fildes et al. (2009), and research based on macroeconomic data in for example McNees (1990) and Turner (1990), also indicates that in general expert adjustments improve forecasting accuracy. However, all studies suggest that there is room for further improvement. For example, Fildes et al. (2009) find that for only three out of the four investigated companies judgmental adjustments increased accuracy on average. Furthermore, the above studies all document a general tendency towards making positive adjustments.

There are also studies which report that expert forecasts are not necessarily better than model forecasts. In an extensive study, in which adjusted forecasts made by different managers are analyzed, Franses and Legerstee (2010) document that managers do not deteriorate forecast accuracy at best, but that often model forecasts outperform the expert-adjusted forecasts. Franses and Legerstee (2011b) show that similar results hold for a range of different forecast horizons. These two studies, and also Sanders (1992) and Fildes and Goodwin (2007b), suggest that model-based forecasts may need less adjustment and that experts perhaps put too much weight on their own contribution.

In sum, in theory, expert-adjusted forecasts should outperform model-based forecasts and in some cases they appear to do so. However, there is also evidence that experts can reduce the forecast quality of model-based forecasts. These conflicting findings trigger the natural question: what is it exactly that the experts do? And, how does this behavior result in improvement or deterioration of forecast accuracy?

Although some recent studies have tried to answer these questions, there is no study that takes all possible expert behavior into account. For example, Fildes et al. (2009) and Trapero et al. (2010) focus on positive versus negative adjustments and on the size of the adjustments when they evaluate what kind of forecast adjustments generate more

accurate forecasts. But what if experts do not look at the model forecasts at all? In that case they are not making (positive or negative) adjustments and there is no relationship between model and expert forecasts at all. If this is the case, how should we evaluate forecast accuracy? Boulaksil and Franses (2009) used a questionnaire to find out what experts do with the model forecasts and how they create final forecasts. Interestingly, part of the experts state that they do not look at the model forecast before they create a forecast themselves. The empirical results in Franses and Legerstee (2009) emphasize the possibility that model forecasts are only partially taken into account in creating the expert forecasts.

This leads to the next natural question: what would be optimal for experts to do? How should they optimally incorporate the model forecasts in the final forecasts? A structured discussion of this issue is absent from the current literature on this subject. We believe it is important though, as insight into optimal behavior can guide methods to evaluate and improve forecasts.

In this paper, we therefore focus on the following three questions which we address given the availability of model forecasts, expert forecasts and realizations: (a) Is the final expert forecast related to the model forecast and how? (b) How is this relation influenced by other factors? (c) How does this relation influence forecast accuracy? In this paper we rely on theoretical arguments and we match these with actual data using a model that is new to the literature.

Central to our approach is the relation

$$EF = \alpha + \beta MF + I, \tag{2.1}$$

where EF is the final forecast of the expert, MF is the statistical model forecast and I is what we will call the intuition of the expert. This equation will turn out to be key to understanding and analyzing expert forecasts. As we will argue, estimating the parameters of this relation provides an answer to the first research question. Interesting cases are when α is close to 0 and β is close to 1, indicating that the expert closely follows the model forecasts, and when α and/ or β deviate from these values consid-

erably. Besides the values of these parameters, it is also interesting to examine the relation between intuition I and the model forecasts. Are there any factors influencing the model forecasts that also influence through I? If this is the case, one could have evidence for double counting, a phenomenon also described in Bunn and Salo (1996).

Relating α and β to various factors can provide an answer to our second research question. For these factors one can think of characteristics of the realized data, R, like the average size and volatility of R, and of personal characteristics of the expert. It is here where we shall introduce our two-level hierarchical Bayes model.

Finally, relating to research question three, we show that the values of α and β, the correlation between I and R, and the correlation between I and MF influence forecast accuracy of EF. We provide theoretical arguments and we hold that against our empirical data.

As we have actual data for individual forecasters for various variables and various forecast horizons, we propose a two-level Hierarchical Bayes model. Its first level is an extended version of (2.1) whereas the second level consists of equations that relate the parameters in (2.1) to characteristics of the variable being forecasted, of the forecasts and of the experts. Furthermore, we take into account the possible endogeneity of the model forecasts in (2.1), that is, potential correlation between MF and I, which slightly complicates parameter estimation.

For our case study we use a large data set containing model forecasts and expert forecasts of different experts for stock keeping unit (SKU) level sales data of various medical products. We document that values for α and β differ substantially across products and experts. Factors such as average sales level, sales volatility, and forecast horizon appear to influence the size of α and β. We also draw conclusions on the optimal values for α and β in terms of forecast accuracy. As such, our study is the first to relate expert behavior with expert performance using non-experimental data.

The remainder of the paper is structured as follows. In the next two sections we formulate the hypotheses which are the starting point of our data analysis and which follow from theory and previous research. In Section 2.4 we describe the models that

we develop to test the hypotheses. Section 2.5 describes the data and the results of our case study. The final section concludes.

2.2 Modeling expert behavior

What is it that experts do with model forecasts when they create their own forecasts and how is this behavior influenced by other factors? We discuss these two questions, where we assume that there are no records available of this behavior, and hence that we have to use the actual forecasts and realizations to answer the questions.

Although most that we put forward in this section is true for any kind of forecasts from experts, we focus in this section on forecasts for SKU-level sales data as this matches our empirical illustration.

2.2.1 What do experts do with model forecasts?

We assume a linear relation between expert forecasts and model forecasts, that is

$$EF_{t+h|t} = \alpha + \beta MF_{t+h|t} + I_{t+h|t}, \tag{2.2}$$

where $EF_{t+h|t}$ is the expert forecast created at origin t for $t+h$, where h is the forecast horizon, $MF_{t+h|t}$ is the model forecast created at the same origin, for the same variable and with the same forecast horizon and where $I_{t+h|t}$ is the intuition of the expert at origin t. We assume that for all t, $\mathrm{E}[I_{t+h|t}] = 0$, where E is the expectation operator. In later sections we describe and estimate a model for which (2.2) is our main building block, where we assume availability of $EF_{t+h|t}$ and $MF_{t+h|t}$ for $t = 1, 2, ..., T$.

One typical situation captured by this model is when $\alpha = 0$ and $\beta = 1$. This can be seen as the benchmark situation, in which the expert closely follows the model forecasts. On average over time, if the model forecasts increase (decrease) the expert forecasts increase (decrease) by the same amount. The expert forecasts are on average not higher nor lower than the model forecasts (if the model forecasts are unbiased, the

expert forecasts are unbiased like the model forecasts) and the only differences between model forecasts and expert forecasts are captured by the intuition of the expert $I_{t+h|t}$. $I_{t+h|t}$ covers factors that influence the expert forecasts otherwise than model forecasts. In this situation a forecaster closely follows the model forecasts and apparently trusts the model forecasts, but might decide to increase or decrease the model forecasts based on factors captured in $I_{t+h|t}$.

A second interesting variant of (2.2) is when $\alpha \neq 0$ and $\beta = 1$. Although the expert still follows the model closely, the expert forecasts are on average higher ($\alpha > 0$) or lower ($\alpha < 0$) than the model forecasts. Thus there is a constant deviation from the model forecasts. The general level of expert forecasts is thus different than that of the model forecasts. A potential reason for constant deviation might be that the expert has another loss function than used by the model (which is typically mean squared error loss). For example, the expert might believe that underpredicting is worse than overpredicting.

If $\alpha = 0$, but $0 < \beta < 1$, the relation between model forecasts and expert forecasts is less strong than when $\beta = 1$. A change in the next model forecast dampens the expert forecast in the same direction (on average). The expert feels that the model forecasts move in the right direction but not to the right extent and this results in $0 < \beta < 1$. At the same time, as $\alpha = 0$ and $\mathrm{E}[I_{t+h|t}] = 0$ and assuming the variable to be forecasted is always positive,[1] the expert forecasts are on average lower than the model forecasts.

If $\alpha = 0$ and $\beta > 1$, the expert reacts excessively to the model forecasts. On average, the expert forecasts move in the same direction as the model forecasts, but the expert has reasons to believe that the model generally underestimates the trend in the data. As $\alpha = 0$, and $\mathrm{E}[I_{t+h|t}] = 0$, and the variable to be forecasted is always positive, the expert forecasts are on average higher than the model forecasts.

Finally, an extreme variant of (2.2) appears when $\beta = 0$. Here, the expert does not

[1] For our SKU-level sales data this is in general the case.

consider the model forecast at all and the expert forecasts are determined by other factors. In this situation, expert forecasts do not entail judgmental adjustments to model-based forecasts, as the expert gives his or her own independently created forecasts. The expert forecast is equal to the intercept plus intuition.

Of course, there are other variants, like when $\alpha > 0$ and $0 < \beta < 1$. Here the expert forecasts do not necessarily deviate from the model forecasts (they might on average approximately be the same). The expert only partially follows the model forecasts, and uses corrections via the intercept.

In sum, expression (2.2) encompasses many of the possible expert forecasting practices and it is a good starting point for our analysis. It would now be interesting if there is any empirical evidence of the values of α and β. Recently, more data sets have become available containing statistical model forecasts and expert forecasts. Boulaksil and Franses (2009) showed with a questionnaire that 50% of the responding managers do not rely on the model forecasts when they create their final forecasts. This suggests that β is smaller than 1, or, stated differently, closer to 0. In Franses and Legerstee (2009) the parameters in model (2.2) are estimated using SKU-sales data and it is reported that β is close to 0.4, on average, and there is a large variety of potential estimated values.

Fildes et al. (2009) and Mathews and Diamantopoulos (1986) show that often the differences between expert forecasts and model forecasts are positive. Fildes et al. (2009) find for their one-step-ahead forecasts of SKU-sales more positive than negative adjustments and they also find that the upward adjustments tend to lead to final expert forecasts that overpredict. Franses and Legerstee (2011a) show that for forecasts with horizons ranging from one to twelve months there are more positive adjustments than negative adjustments. This might capture the preference of a manager to overpredict in order to prevent being out of stock and thus that managers may have a loss function different than that of the model forecasts. If we relate these findings to (2.2), we could state that for many experts α is larger than 0, that β is different from 1, or both.

If $\beta < 1$, as is frequently observed, then the observed upward adjustments imply

that α is often larger than 0. Even if there would not be an upward bias in the expert forecasts, a positive α makes sense in case of a β smaller than 1, in order to prevent a downward bias in the final forecasts, assuming that the model forecasts are unbiased. To summarize, we put forward the following two hypotheses

Hypothesis 1

a. For many experts $\beta \neq 1$ in (2.2).

b. When $\beta \neq 1$, for many experts $\beta < 1$ in (2.2).

Hypothesis 2

a. For many experts $\alpha \neq 0$ in (2.2).

b. When $\alpha \neq 0$, for many experts $\alpha > 0$ in (2.2).

2.2.2 What causes $\beta \neq 1$ and $\alpha \neq 0$?

Now that we have an idea about what it is that experts could do with model forecasts and what they might often do, we can look for factors that determine this behavior.

From the questionnaire results reported in Boulaksil and Franses (2009) we learn that managers are quite confident about their own ability to forecast and that they lack confidence in the model forecasts. As products with large sales volumes might be more important to a manager and as predictions for near-by sales are probably more important because of their urgency, the manager might put even less trust in the model in these situations. Boulaksil and Franses (2009) also find that recent volatile sales figures decreases the trust by managers in the model and they feel the need to make even more adjustments, which thus would result in an even lower value for β. Fildes et al. (2009) investigate if judgmental forecasts improve the forecast accuracy when sales volume volatility is high, but they find evidence of the opposite. These authors suggest that volatile series are more difficult to forecast, but with Boulaksil and Franses (2009) we would argue that it can also be due to excessive adjustment. We therefore hypothesize the following

Hypothesis 3 The probability that β in (2.2) deviates away from 1 towards 0 increases when

a. the mean of a target variable is higher;

b. a target variable fluctuates more;

c. the forecast horizon decreases.

When a manager wants to prevent being out of stock, then higher average sales volumes and more volatility increases the size of forecast adjustments. Furthermore, Franses and Legerstee (2011a) show that adjustments are more often upwards than downwards for all forecast horizons, but that this is most prominent for shorter horizons. Hence, we conjecture that

Hypothesis 4 The probability that α in (2.2) deviates away from 0 increases when

a. the mean of a target variable is higher;

b. a target variable fluctuates more;

c. the forecast horizon decreases.

In Section 2.4 we propose an econometric model with which can put these hypotheses to a test.

2.2.3 Experts' intuition

When the managers do not trust the model forecasts and make their own forecasts, it is quite likely that there are factors which influence both model forecasts and expert forecasts. Managers have stated in the questionnaire reported in Boulaksil and Franses (2009) that they include recent sales figures as input to their forecast adjustments, even though they know that recent sales figures are also covered by the statistical model forecasts. This is in accordance with the lab findings of Goodwin and Fildes (1999a), which is that experts do not only look at special events for their adjustments, but they also consider past data. As these past (sales) data are usually also the input for the

models used to create the model forecasts, the result would be a correlation between $MF_{t+h|t}$ and $I_{t+h|t}$ in (2.2), or stated differently $\mathrm{E}(MF_{t+h|t}I_{t+h|t}) \neq 0$. So, our final hypothesis about expert forecasting behavior is

> **Hypothesis 5** For many experts MF is endogenous in (2.2), meaning $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] \neq 0$.

Note that MF being endogenous (and thus not exogenous) has two important implications. First of all, it tells us something about what the experts do. It shows that experts use the same information as the model forecasts, possibly in the same way, but more likely in another way. The result could amount to double counting, or at least to an inefficient use of information, especially when the model forecasts are optimal in processing that same information.

The second implication of the endogeneity of MF in (2.2) has to do with parameter estimation. It is well known that Ordinary Least Squares (OLS) results in an inconsistent estimate of β if MF is endogenous, see Heij et al. (2004, p. 396-418). This may result in incorrect conclusions about what it is that experts do with model forecasts. For example, it may seem that there is a strong relation between EF and MF with $\beta \approx 1$, while in fact the expert does not look at the model forecasts at all, but simply uses the same factors as input for his or her forecasts as the statistical model used when creating the model forecasts. How to deal with this estimation issue is discussed in Section 2.4. Before we turn to our econometric model, we first discuss various implications of expert behavior on forecast accuracy.

2.3 Theoretical implications for accuracy

In this section we demonstrate the theoretical link between the behavior of the experts and their forecasting accuracy. To our knowledge this has never been done before in the literature.

To study the implications of deviating from the benchmark $\alpha = 0$ and $\beta = 1$,

we need to propose a loss function to evaluate forecast accuracy. We propose to consider a variant of the well-known and often used root mean squared prediction error ($RMSPE$), and this variant is the expected squared prediction error ($ESPE$) defined by

$$ESPE = \mathrm{E}[(R_{t+h} - EF_{t+h|t})^2], \tag{2.3}$$

where $EF_{t+h|t}$ is as defined before and where R_{t+h} is the realization at $t+h$. This loss function is chosen for convenience, and also because it gives implementable optimality results for α, β and I, as the managers only have expected values of sales instead of realized values when they create their forecasts. The conclusions obtained in this section with this loss function can be generalized to other loss functions, such as the mean squared prediction error ($MSPE$), the $RMSPE$ and the difference between the $(R)MSPE$ of the model and that of the expert ($D(R)MSPE$).

If (2.2) is substituted in (2.3) we obtain

$$ESPE = \mathrm{E}[(R_{t+h} - \alpha - \beta MF_{t+h|t})^2] + \mathrm{E}[I_{t+h|t}^2] - 2\mathrm{E}[((R_{t+h} - \beta MF_{t+h|t})I_{t+h|t}], \tag{2.4}$$

where we have used that $\mathrm{E}[I_{t+h|t}] = 0$. The expert can influence three factors of the $ESPE$, and these are α, β and $I_{t+h|t}$. For each of these we will discuss the optimal values of α, β and $I_{t+h|t}$ that minimize $ESPE$, and how deviations from the optimal values will influence this $ESPE$.

2.3.1 Optimal settings

For ease of derivation, at first we assume that MF is exogenous in (2.2) and thus that $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] = 0$. Later on we will relax this assumption.

$\frac{\partial ESPE}{\partial \alpha} = 0$ gives the value for α that minimizes $ESPE$, and that is the OLS estimate of the constant term in equation (2.2) given by

$$\alpha_{opt} = \mathrm{E}[R_{t+h}] - \beta \mathrm{E}[MF_{t+h|t}]. \tag{2.5}$$

$\frac{\partial ESPE}{\partial \beta} = 0$ and then substituting it with the optimal value for α in (2.5) gives the optimal value for β, that is,

$$\beta_{opt} = \frac{\mathrm{E}[MF_{t+h|t}R_{t+h}] - \mathrm{E}[MF_{t+h|t}]\mathrm{E}[R_{t+h}]}{\mathrm{E}[MF^2_{t+h|t}] - \mathrm{E}[MF_{t+h|t}]^2} = \frac{\mathrm{Cov}[MF_{t+h|t}, R_{t+h}]}{\mathrm{V}[MF_{t+h|t}]}, \qquad (2.6)$$

where Cov means covariance and V denotes variance. Under the condition that the model forecasts are unbiased relative to expected realizations, thus $\mathrm{E}[MF_{t+h|t}] = \mathrm{E}[R_{t+h|t}]$, we see that the more $\mathrm{E}[MF_{t+h|t}R_{t+h}]$ differs from $\mathrm{E}[MF^2_{t+h|t}]$, the more β_{opt} differs from 1. However, under the additional condition that $\mathrm{E}[MF_{t+h|t}R_{t+h}] = \mathrm{E}[MF^2_{t+h|t}]$, we obtain that $\beta_{opt} = 1$ and $\alpha_{opt} = 0$. We could call this additional condition the relative unbiasedness of the model forecasts. What this relative unbiasedness means is perhaps most easily understood by looking at the estimators of $\mathrm{E}[MF_{t+h|t}R_{t+h}]$ and $\mathrm{E}[MF^2_{t+h|t}]$, which are $\sum MF_{t+h|t}R_{t+h}$ and $\sum MF^2_{t+h|t}$, where the summations $\sum$ run over a sample of data. The condition is not met if $\sum MF_{t+h|t}R_{t+h} - \sum MF^2_{t+h|t} < 0$, which occurs when MF is larger than R especially for the larger MF, or if $\sum MF_{t+h|t}R_{t+h} - \sum MF^2_{t+h|t} > 0$, which occurs when MF is smaller than R especially for the larger MF.

To get more insight into this relative unbiasedness we consider an example. Suppose we have only two observations ($T = 2$), with realizations $R_2 = 5$ and $R_3 = 15$ and we have two different sets (marked with superscripts) of one-month-ahead model forecasts, namely $\{MF^1_{2|1} = 10, MF^1_{3|2} = 10\}$ and $\{MF^2_{2|1} = 11, MF^2_{3|2} = 9\}$. The first set of model forecasts is unbiased and relatively unbiased, as $\sum R_{t+h} = \sum MF_{t+h|t}$ and $\sum MF_{t+h|t}R_{t+h} = \sum MF^2_{t+h|t}$. The second set of model forecasts is unbiased, but not relatively unbiased, because $\sum MF_{t+h|t}R_{t+h} = 190$ and $\sum MF^2_{t+h|t} = 202$. We see now that deviations of MF from R have more weight for larger MF. If $\sum MF_{t+h|t}R_{t+h} - \sum MF^2_{t+h|t} < 0$, a value for β smaller than 1 is optimal and if $\sum MF_{t+h|t}R_{t+h} - \sum MF^2_{t+h|t} > 0$, a value for β larger than 1 is optimal (see (2.6)).

Finally, let us look at the influence of $I_{t+h|t}$ on $ESPE$. Remember that we restricted I and MF to be uncorrelated. Although it is impossible to derive for $I_{t+h|t}$

what its optimal value is, we can see from (2.4) that adding intuition is only beneficial for reducing the expected forecast error if R and I are positively correlated (see the negative sign before the third right-hand-side element). To be more precise, it should hold that

$$2\text{Cov}[R_{t+h}I_{t+h|t}] > \text{V}[I_{t+h|t}], \tag{2.7}$$

which means that the covariance between R and I should be larger than half the variance of I. However, we restricted I and MF to be uncorrelated and we might assume a strong correlation between R and MF. The stronger the last two are related, the harder it is for I and R to be correlated, while maintaining the exogeneity of MF in (2.2). Note that this conclusion supplements the conclusion of Blattberg and Hoch (1990, pp. 890-891), who state that combinations between model and expert forecasts will be more accurate than the model or expert forecasts separately if the intuition of the expert is related to the true values.

If we relax the exogeneity assumption that $\text{E}[MF_{t+h|t}I_{t+h|t}] = 0$, matters get more complicated. Working in the same way as for the case of exogenous model forecasts, we find the following value of α that minimizes $ESPE$:

$$\alpha_{opt} = \text{E}[R_{t+h}] - \beta\text{E}[MF_{t+h|t}], \tag{2.8}$$

which is the same as before, and the following value of β that minimizes $ESPE$:

$$\begin{aligned}\beta_{opt} &= \frac{\text{E}[MF_{t+h|t}R_{t+h}] - \text{E}[MF_{t+h|t}]\text{E}[R_{t+h}] - \text{E}[MF_{t+h|t}I_{t+h|t}]}{\text{E}[MF^2_{t+h|t}] - \text{E}[MF_{t+h|t}]^2} \\ &= \frac{\text{Cov}[MF_{t+h|t}, R_{t+h}] - \text{Cov}[MF_{t+h|t}I_{t+h|t}]}{\text{V}[MF_{t+h|t}]},\end{aligned} \tag{2.9}$$

which is different than before. If we assume the model forecasts to be unbiased and relatively unbiased we obtain

$$\alpha_{opt} = \frac{\text{Cov}[MF_{t+h|t}I_{t+h|t}]}{\text{V}[MF_{t+h|t}]}\text{E}[MF_{t+h|t}], \tag{2.10}$$

$$\beta_{opt} = 1 - \frac{\text{Cov}[MF_{t+h|t}I_{t+h|t}]}{\text{V}[MF_{t+h|t}]}. \tag{2.11}$$

We can see that the optimal value of β is now negatively correlated with the covariance between MF and I. The higher the correlation, the lower β_{opt} should be, and vice versa. This is intuitively understandable, as a high covariance between MF and I and a high β (equal to 1 or higher) would result in double counting. In that case the expert fully takes the model forecasts into account, but also lets the final forecasts be influenced by the same factors that determine the model forecasts.

At the same time, a higher covariance between MF and I should result in a higher value for α because of a lower value for β. As $\mathrm{E}[I_{t+h|t}] = 0$, α should in this case be different from 0 to make the expert forecasts unbiased.

The question now is: how beneficial is it for the expert to relate intuition to the model forecasts and to what extent? If we look at (2.4), our initial idea could be that a high correlation between R and I and a low, preferably negative, correlation between MF and I is best for expert forecast accuracy. Assuming unbiased and relatively unbiased model forecasts this would result in a β_{opt} larger than 1 and a negative α_{opt}. However, the gains in forecast accuracy achieved when I is positively related to R and when it is negatively related to MF are offset by the second term in (2.4), that is, a higher variance of I increases the forecast error. Furthermore, the more R and MF are related, the harder it is to let I be positively correlated with R and negatively correlated with MF.

If R and MF are not that strongly related, it might be best to choose $I_{t+h|t}$ in such a way that it corrects for the mistakes that the model forecasts make, thus to let factors that wrongly influence MF negatively influence I. This results in a negative correlation between I and MF and a positive correlation between I and R. In that case β should be larger than 1.

In short, we have to take a closer look at the last two terms in (2.4). We observe that adding intuition is only beneficial if

$$2\mathrm{E}[(R_{t+h} - \beta MF_{t+h|t})I_{t+h|t}] > \mathrm{V}[I_{t+h|t}]. \tag{2.12}$$

Hence, a necessary condition is that intuition is positively correlated with $(R_{t+h} -$

$\beta MF_{t+h|t}$), which implies that $\mathrm{E}[R_{t+h}I_{t+h|t}] > \beta\mathrm{E}[MF_{t+h|t}I_{t+h|t}]$. Thus for $\beta = 1$, the correlation between intuition and realization has to be larger than the correlation between intuition and model forecast.

2.3.2 Implications and hypotheses

Before we summarize the above in a set of statements we define the following conditions:

$$\mathrm{E}[R_{t+h}] = \mathrm{E}[MF_{t+h|t}], \tag{2.13}$$

$$\mathrm{E}[MF_{t+h|t}R_{t+h}] = \mathrm{E}[MF^2_{t+h|t}]. \tag{2.14}$$

Furthermore, we generalize the above results to the difference between the $ESPE$ of the model and that of the expert ($DESPE$), as usually the interest is in deterioration or improvement of the expert forecasts over the model forecasts. If we obtain a minimum value of $ESPE$ for particular values of α, β and $I_{t+h|t}$, we also obtain an optimal value of $DESPE$, meaning that (for given model forecasts) $DESPE$ is at its maximum value.

> **Statements** In order to have maximum improvement in expected forecast accuracy of EF over MF it has to hold in (2.2) that,
>
> **a.** $\alpha = 0$, $\beta = 1$, and (2.7) is met for $I_{t+h|t}$, assuming that (2.13) and (2.14) are met and that $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] = 0$;
>
> **b.** α is as in (2.10), β as in (2.11), and (2.12) is met for $I_{t+h|t}$, if (2.13) and (2.14) are met, but possibly $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] \neq 0$;
>
> **c.** α is as in (2.8), β as in (2.9), and (2.12) is met for $I_{t+h|t}$, if (2.13) and (2.14) are *not* met and possibly $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] \neq 0$.

Note that (2.7) and (2.12) are minimum requirements for intuition to be beneficial and for $DESPE$ to be optimal.

Any deviation from the optimal values for α and β and from (2.12) results in higher prediction errors for EF, where the amount of loss of precision depends on the inter-

action between α, β and $I_{t+h|t}$. For example, in case β is larger than 1, and the model forecasts are unbiased, relatively unbiased (conditions (2.13) and (2.14) are met) and exogenous in (2.2), it is optimal that α is smaller than 0. Furthermore, in that case, the correlation between the intuition of the expert and the realized values should be even larger than when β equals 1.

Although the described behavior is theoretically the behavior that generates the most accurate forecasts, it is questionable whether an expert can act according to the statements (a) to (c) in practice. The interactions between the various determinants of forecast accuracy, especially when taking into account the possibility that the conditions are not met, are quite complex. Furthermore, for a given set of actual model forecasts it might be assumed that conditions (2.13) and (2.14) are met approximately and that R and MF are strongly related in general. Therefore we put forward the following simpler hypothesis:

Hypothesis 6 The improvement in expected forecast accuracy of EF over that of MF increases when in (2.2)

a. α is 0 or α gets closer to 0;

b. β is 1 or β gets closer to 1;

c. the correlation between MF and I decreases;

d. the correlation between I and R increases.

For a given data set, for which we do not have reasons to doubt that the conditions as defined in (2.13) and (2.14) are met, it might be interesting to test Hypothesis 6.

2.4 Empirical models

In this section we will explain in detail how a (non-trivial) econometric model can be constructed to validate the components of Hypothesis 6. We first consider expert behavior and then its link with forecast accuracy.

2.4.1 Model of expert behavior

In this section we propose a model to estimate what the experts do with the model forecasts and which factors influence this behavior. It is a two-level Hierarchical Bayes model, for which the parameters can be estimated using panel data, consisting of model forecasts and expert forecasts for different products and for different time periods.

To meet the typical data format in practice, and also to reduce notational burden, we now introduce a slightly different notation. Let $EF_{i,t}$ denote the expert forecast created in period t for case i, where i covers products and forecast horizons. Furthermore, $MF_{i,t}$ is the model forecast created in that same period, for that same product and with the same forecast horizon. Let T_i be the number of observations for product and forecast horizon denoted with i, which can take a maximum value of T. There are N product-horizon combinations and thus time series. See Appendix 2.A.1 for a more detailed explanation of the data format. Using this notation we can then write (2.2) as

$$EF_{i,t} - MF_{i,t} = \alpha_i^* + \beta_i^* MF_{i,t} + \varepsilon_{i,t}, \tag{2.15}$$

with $\varepsilon_{i,t} \sim N(0, \sigma^2_{\varepsilon,i})$. Note that β_i^* in this model associates with $\beta - 1$ in (2.2) and α_i^* with α in (2.2). This expression constitutes the first level of our model.

To correctly estimate the parameters and to see which factors influence α_i^* and β_i^* over t, we add a second level to the model. As $\alpha_i^* = 0$ and $\beta_i^* = 0$ are the special benchmark cases in the behavior of experts and the forecast accuracy related to it, we take these as our starting point.

Let z_i be a vector containing explanatory variables such as mean and volatility of the variable being forecasted, we can expand the model with

$$\alpha_i^* = \begin{cases} 0 & \text{if } P_i = 1 \\ \alpha_i^\dagger = z_i'\gamma_\alpha + \xi_i & \text{if } P_i = 0, \end{cases} \tag{2.16}$$

and

$$\beta_i^* = \begin{cases} 0 & \text{if } S_i = 1 \\ \beta_i^\dagger = z_i'\gamma_\beta + \eta_i & \text{if } S_i = 0, \end{cases} \tag{2.17}$$

with $\xi_i \sim N(0, \sigma_\xi^2)$ and $\eta_i \sim N(0, \sigma_\eta^2)$. P_i and S_i are unobserved variables which can take values 1 and 0. With $Pr[P_i = 1] = \kappa_i$ and $Pr[S_i = 1] = \lambda_i$, we assume that there is an unconditional probability of size κ_i that $\alpha_i^* = 0$ and that there is an unconditional probability of λ_i that $\beta_i^* = 0$. Stated differently, with a probability of κ_i times λ_i the expert forecasts of case i follow the model forecasts closely and match with the benchmark situation as described in Section 2.2.1. If α_i^* differs from 0 it equals $\alpha_i^\dagger$ which is then conditional normally distributed and which depends linearly on the variables in z_i. If β_i^* differs from 0 it equals $\beta_i^\dagger$ which is also conditional normally distributed and which also depends linearly on the variables in z_i, but with other parameters (γ_β).

If we consider q_i and w_i to be unobserved random variables, we use the following conditional probabilities:

$$P_i = \begin{cases} 1 & \text{if } q_i = z_i'\psi_\alpha + \nu_i > 0 \\ 0 & \text{if } q_i = z_i'\psi_\alpha + \nu_i \leq 0, \end{cases} \tag{2.18}$$

and

$$S_i = \begin{cases} 1 & \text{if } w_i = z_i'\psi_\beta + \omega_i > 0 \\ 0 & \text{if } w_i = z_i'\psi_\beta + \omega_i \leq 0, \end{cases} \tag{2.19}$$

with $\nu_i \sim N(0,1)$ and $\omega_i \sim N(0,1)$. Stated differently, the probabilities that $P_i = 1$ ($\alpha_i^* = 0$) and that $S_i = 1$ ($\beta_i^* = 0$) are defined as a probit model with z_i as explanatory variables. We can also write this as

$$\kappa_i = \int_0^\infty \phi(q_i; z_i'\psi_\alpha, 1) dq_i, \tag{2.20}$$

and

$$\lambda_i = \int_0^\infty \phi(w_i; z_i'\psi_\beta, 1) dw_i, \tag{2.21}$$

where $\phi(\cdot; c1, c2)$ is the probability density function (pdf) of a normal distribution with mean $c1$ and variance $c2$. Thus, the variables in z_i are related to $\alpha_i^\dagger$ and $\beta_i^\dagger$, but also to the probabilities that $\alpha_i^* = 0$ and that $\beta_i^* = 0$. Although we use for all four relations the

same z_i here, it is of course also possible to use different sets of explanatory variables. Equations (2.16), (2.17), (2.20) and (2.21) constitute the second level of our model.

Sofar we have assumed that the error terms in our basic equation (2.15) are unrelated to the model forecasts, and thus that the model forecasts are exogenous. It is however very well possible that there is correlation between these two components, as explained in Section 2.2.3. If this problem is ignored we might find values for β_i^* that are inconsistent. To account for possible endogeneity in the first equation we therefore add the following component to the model, that is,

$$MF_{i,t} = \mu_i + \delta_i V_{i,t} + \zeta_{i,t}, \tag{2.22}$$

with $V_{i,t}$ an instrumental variable. Now we have $(\varepsilon_{i,t}, \zeta_{i,t})' \sim MN(0, \Omega_i)$, where $\varepsilon_{i,t}$ is from (2.15) and where $MN(0, \Omega_i)$ is the bivariate normal distribution with mean 0 for both variables and with covariance matrix Ω_i (which is a 2×2 matrix). If there is no correlation between $\varepsilon_{i,t}$ and $\zeta_{i,t}$, or, stated differently, $\Omega_i(1,2) = \Omega_i(2,1) = 0$, there is no endogeneity.

Taking everything together, the full final model now reads as

$$EF_{i,t} - MF_{i,t} = \alpha_i^* + \beta_i^* MF_{i,t} + \varepsilon_{i,t}, \tag{2.23}$$

$$MF_{i,t} = \mu_i + \delta_i V_{i,t} + \zeta_{i,t}, \tag{2.24}$$

$$\alpha_i^* = \begin{cases} 0 & \text{if } P_i = 1 \\ \alpha_i^\dagger = z_i'\gamma_\alpha + \xi_i & \text{if } P_i = 0 \end{cases} \tag{2.25}$$

$$\beta_i^* = \begin{cases} 0 & \text{if } S_i = 1 \\ \beta_i^\dagger = z_i'\gamma_\beta + \eta_i & \text{if } S_i = 0, \end{cases} \tag{2.26}$$

$$P_i = \begin{cases} 1 & \text{if } q_i = z_i'\psi_\alpha + \nu_i > 0 \\ 0 & \text{if } q_i = z_i'\psi_\alpha + \nu_i \leq 0, \end{cases} \tag{2.27}$$

$$S_i = \begin{cases} 1 & \text{if } w_i = z_i'\psi_\beta + \omega_i > 0 \\ 0 & \text{if } w_i = z_i'\psi_\beta + \omega_i \leq 0. \end{cases} \tag{2.28}$$

The first two equations are the first level of the model in which the difference between EF and MF is linked to MF and where possible endogeneity of MF is incorporated. The second level of the model is given by the other four equations, where the parameters of the first level are linked to potentially explanatory variables. The benchmark case $\alpha_i^* = 0$ and $\beta_i^* = 0$ has a key position in this model.

To estimate the posterior results of the parameters of this model, namely $\theta = (\{\beta_i^\dagger\}_{i=1}^N, \{\alpha_i^\dagger\}_{i=1}^N, \{\mu_i\}_{i=1}^N, \{\delta_i\}_{i=1}^N, \gamma'_\alpha, \gamma'_\beta, \psi'_\alpha, \psi'_\beta, \{\Omega_i\}_{i=1}^N, \sigma_\xi^2, \sigma_\eta^2)$, the Markov Chain Monte Carlo (MCMC) methodology, and in particular Gibbs sampling, is used. Technical details on this sampler are presented in Appendix 2.A.2. We are especially interested in the values of parameters $\{\beta_i^\dagger\}_{i=1}^N$, $\{\alpha_i^\dagger\}_{i=1}^N$, γ_α, γ_β, ψ_α, ψ_β and $\{\Omega_i\}_{i=1}^N$, as these represent the behavior of the experts and how this behavior is governed by other factors.

2.4.2 Evaluating forecasts

The estimated parameters of the model in the previous section can be used to test Hypotheses 1 to 5 about the behavior of experts. However, we are also interested in what the experts should do, which is the subject of the Statements and Hypothesis 6. As the Statements follow straightforwardly from optimization of the forecast accuracy target function there is no need to test it. However, the rules to follow according to these statements are quite complex and therefore Hypothesis 6 comprises a simpler set of rules to follow. To test the validity of Hypothesis 6 we need one additional model which we propose in this subsection. In this model we use a measure of the forecast precision of the expert as compared to the forecast precision of the model and relate this with variables as mentioned in Hypothesis 6.

Let $DRMSPE_i$ be the improvement in root mean squared prediction error of $EF_{i,t}$ over $MF_{i,t}$, thus

$$DRMSPE_i = \sqrt{\frac{1}{T_i}\sum(R_{i,t} - MF_{i,t})^2} - \sqrt{\frac{1}{T_i}\sum(R_{i,t} - EF_{i,t})^2}. \qquad (2.29)$$

We use this criterium instead of $DSPE$ to reduce variability. With the regression model

$$DRMSPE_i = r_i'\vartheta + \iota_i, \tag{2.30}$$

it is possible to test which factors influence forecast improvement.

First of all, we want to test if $\alpha^* = 0$ indeed increases forecast improvement, as compared to cases where $\alpha^* \neq 0$. This is the first part of Hypothesis 6a. We also want to test if, assuming that α^* is different from 0, a smaller value of α^* in absolute sense is beneficial to the forecast improvement (second part of Hypothesis 6a). Therefore, we consider the estimates of P_i and the estimates of $|\alpha_i^\dagger(1 - P_i)|$ as explanatory variables in (2.30), where we use the posterior means for P_i and $\alpha^\dagger$. We call the first variable in the remainder of this paper 'No intercept' and following Hypothesis 6a we expect this variable to have a positive effect. The second variable is called 'Size intercept' and following Hypothesis 6b we expect this variable to have a negative effect.

To test if $\beta^* = 0$ (or β in (2.2) equals 1) increases forecast improvement compared to $\beta^* \neq 0$ (first part of Hypothesis 6b), we add the posterior mean for S_i. The second part of Hypothesis 6b, namely that a larger absolute value of β^* decreases forecast improvement, is tested by using the estimates of $|\beta_i^\dagger(1-S_i)|$ as an explanatory variable, where we again use the posterior mean for S_i and we use the posterior mean for $\beta_i^\dagger$. These variables will carry the labels 'Relation MF' and 'Size relation MF' and we expect the first variable to have a positive effect and the second variable to have a negative effect.

Hypothesis 6c states that $DRMSPE$ increases if the correlation between MF and I decreases. To test this we use $\rho_{\Omega,i} = \Omega_i(1,2)/\sqrt{\Omega_i(1,1)\Omega_i(2,2)}$ as an explanatory variable, where we use the posterior mean for Ω_i, label $\rho_{\Omega,i}$ 'Endogeneity', and we expect a parameter with a negative value.

Finally, by including in (2.30) $\rho_{\varepsilon R,i} = corr(\varepsilon_{i,t}, R_{i,t})$ Hypothesis 6d is considered. That is, the correlation between the estimated errors of (2.15) and the realized values of the variable of interest is used to see if correlation between the expert intuition

and the true values increases the forecasts. The errors of (2.15), $\varepsilon_{i,t}$, are estimated as $EF_{i,t} - MF_{i,t} - \alpha_i^\dagger(1 - P_i) - \beta_i^\dagger(1 - S_i)MF_{i,t}$, using the posterior means for $\alpha_i^\dagger$, $\beta_i^\dagger$, P_i and S_i. The variable $\rho_{\varepsilon R,i}$ is labeled 'Intuition' in the remainder of the paper and following Hypothesis 6d we expect it to have a positive effect in (2.30).

Concluding, we have for (2.30) the set of six explanatory variables

$$r_i' = [1,\ P_i,\ |\alpha_i^\dagger(1 - P_i)|,\ S_i,\ |\beta_i^\dagger(1 - S_i)|,\ \rho_{\Omega,i},\ \rho_{\varepsilon R,i}]. \tag{2.31}$$

See Table 2.1 for an overview of the variables in r_i, the names of the variables and for the hypothetical sign of the parameters in (2.30) following Hypothesis 6.

Table 2.1: A summary of the variables in r_i in model (2.30) and their hypothetical effect on $DRMSPE$ as denoted in (2.29) according to Hypothesis 6.

Name	**Variable**	**Hypothetical effect**
No intercept	P_i	$+$
Size intercept	$\lvert\alpha_i^\dagger(1 - P_i)\rvert$	$-$
Relation MF	S_i	$+$
Size relation MF	$\lvert\beta_i^\dagger(1 - S_i)\rvert$	$-$
Endogeneity	$\rho_{\Omega,i}$	$-$
Intuition	$\rho_{\varepsilon R,i}$	$+$

2.5 Empirical results

To illustrate the usefulness of our two models we make use of an extensive panel data set. The data set covers SKU-level sales data and is described in detail in the next subsection. In Subsections 2.5.2 and 2.5.3 the results of our analysis are discussed.

2.5.1 Data set

For our case study we use monthly sales data of a large pharmaceutical company. The company has its headquarters in The Netherlands, and has local offices in various countries. The company uses an automated statistical package to create forecasts using lagged sales figures as the only input. Each month model selection and parameter estimation are updated, whereby the package uses techniques such as Box-Jenkins and Holt-Winters. These model forecasts are then sent to the managers in the local offices, after which they quote their own forecasts.

We have at our disposal model forecasts, manager forecasts and actual sales figures for November 2004 through November 2006, with for 1-step-ahead forecasts a maximum of 25 triplets per product (medicine), for 2-step-ahead forecasts a maximum of 24 triplets and so on. We have a total of 7250 time series for 1167 different products in 7 different categories, sold in 36 countries. For each series, two observations are lost, because of the instrumental variable we used (see below). Therefore, for each series we have a minimum of 10 observations, a maximum of 23 observations and the forecast horizon ranges from 1 to 7 months.

In the notation of Appendix 2.A.1 this means that we have $N = 7250$, $J = 1167$ and H_j for $j = 1, .., 1167$ is maximally 7. Because there is one manager per country responsible for the expert forecasts, we have $M = 36$. Furthermore, $t = 1$ corresponds with October 2004 and T corresponds with October 2006 (forecast origin).

As an instrumental variable in (2.22) we need a variable that correlates with the model forecasts, but not with the expert forecasts, see, for example, Heij et al. (2004, p. 396-418). The instrumental variable $V_{i,t}$ that we use is $R_{i,t-(h+1)} - MF_{i,t-(h+1)}$, where $R_{i,t-(h+1)}$ concerns case i in month $t - (h + 1)$ and $MF_{i,t-(h+1)}$ is the associated model forecast.[2] So, as instrumental variable we use the most recent forecast

[2]We use the same notation as in Appendix 2.A.1. Thus for MF the second subscript indicates in which period the forecasts are created. In case of R the second subscript indicates in which period the forecasts are created to which the realization belongs, thus it is the realization of period $t - 1$.

error of the model forecast that has the same forecast horizon and that is known at the moment of forecast creation. Franses and Legerstee (2009) show that this variable often does not correlate much with the difference between model forecasts and expert forecasts. Because we do think it correlates with model forecasts (because of the way model forecasts are created), we believe that expert forecasts and this instrument are not strongly correlated.

The variables that we use as explanatory variables in (2.16), (2.17), (2.20) and (2.21) and included in vector z_i are average sales volume, sales volatility and dummy variables for the forecast horizon. We also include dummy variables for the country (and by that for the manager responsible for forecasting) and dummy variables for the category of a product.

The optimal values for α and β depend on conditions (2.13) and (2.14) as defined in Section 2.3, which are conditions on the bias and relative bias of the model forecasts. Furthermore, the more the conditions are not met, the less likely it is that Hypothesis 6 is true. Therefore, it is first useful to find out to what extent these conditions are met for our data. To get insight into this we tested for each case i if there is a significant difference between the mean of MF and the mean of R (condition (2.13)) and if there is a significant difference between the mean of MF times R and the mean of MF^2 (condition (2.14)). For this, we used the common small-sample test for comparing two population means as described in Wackerly et al. (2002a). We find that condition (2.13) is rejected in about 17% of the cases and condition (2.14) in 6% of the cases, where we use a 5% significance level. The test requires the samples to be drawn from a normal distribution. According to the Jarque-Bera test, the hypotheses of normality are not rejected in only 61% of the cases. In again around 17% of these cases (for which both null hypotheses of normality are not rejected) condition (2.13) is rejected at the 5% significance level. To test the second condition, both the MF times R sample and the MF^2 sample need to be drawn from a normal distribution. Here, according to the Jarque-Bera test, the hypotheses of normality are not rejected in 53% of the cases and in around 7% of these cases (for which both null hypotheses of normality are not

rejected) condition (2.14) is rejected at the 5% significance level. Thus although the normality assumption does not always hold, we can state with fair confidence that condition (2.13) holds in about 83% of the cases and condition (2.14) holds in about 93% of the cases.

2.5.2 Expert Behavior

To estimate the parameters of the model described in Section 2.4.1 we generate 80,000 iterations of the Gibbs sampler as described in Appendix 2.A.2. The first 40,000 iterations are used as burn-in sample, and of the last 40,000 iterations every 10th draw is retained and used to calculate mean and standard deviation of the draws. Iteration plots are inspected to check for convergence and are available upon request.

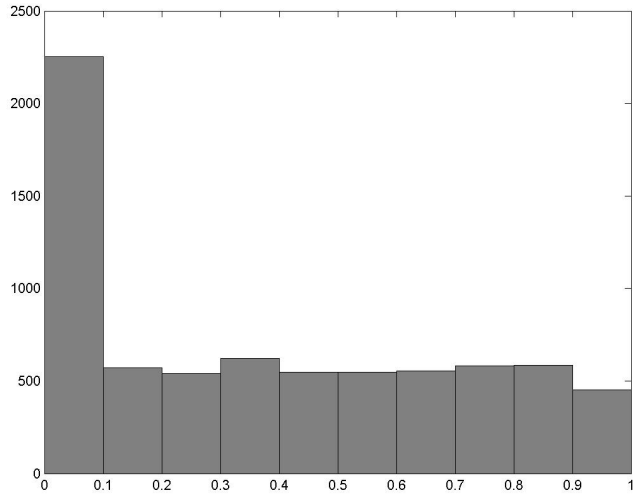

Figure 2.1: Histogram of posterior means for S_i in (2.19), for $i = 1, ..., N$.

The probability that $\beta^* = \beta - 1 = 0$ is varying, which can be seen from the histogram in Figure 2.1 showing the posterior means for S_i for $i = 1, ..., N$. The largest group of cases (2254) has a probability of less than 0.1 that $\beta_i^* = 0$. All the other cases have probabilities that are equally spread between 0.1 and 1. 2718 cases

have a probability higher than 0.5, indicating that in less than 40% of the cases β in (2.2) is likely to be close to 1.

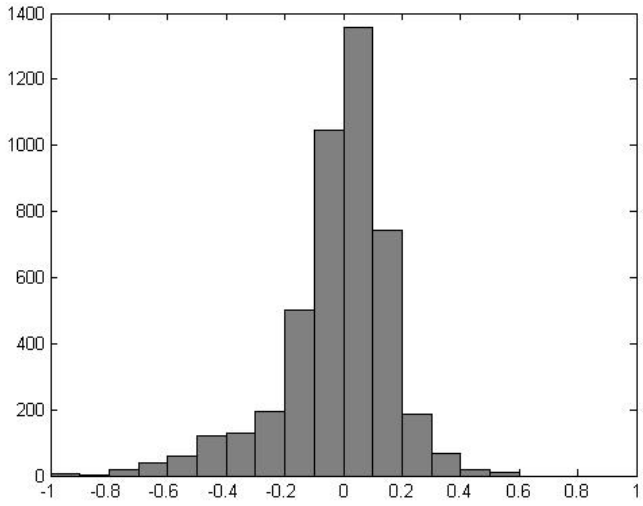

Figure 2.2: Histogram of posterior means for $\beta_i^\dagger$ in (2.17) for which the posterior mean for $S_i < 0.5$, the posterior mean for $\beta_i^\dagger > -1$ and the posterior mean for $\beta_i^\dagger < 1$, for $i = 1, ..., N$.

Figure 2.2 shows a histogram of the posterior means for $\beta_i^\dagger$ for which the posterior mean for $S_i < 0.5$ and for which the posterior mean for $-1 < \beta_i^\dagger < 1$. The smallest $\beta_i^\dagger$ is estimated as -1.14 and the largest is 1.5, but only 11 of the estimated $\beta_i^\dagger$ are below -1 and only 17 above 1. In the remainder of this section, we use $I[S_i < 0.5]\beta_i^\dagger$ as estimated β_i^* and $I[P_i < 0.5]\alpha_i^\dagger$ as estimated α_i^*, where $I[\cdot]$ is an indicator function which takes a value 1 if the expression between brackets is true and 0 otherwise and with posterior means for S_i, $\beta_i^\dagger$, P_i and $\alpha_i^\dagger$. We find that 2406 of the 4532 β_i^* values, that are estimated to be different from 0, are positive. Thus although part a of Hypothesis 1 seems to hold for this data set, part b of this Hypothesis is not supported: β is often different from 1, but when it is different from 1, it is just as likely smaller than 1 than it is larger than 1. However, we do see a fatter tail to the left than to the right: β is more often much lower than 1 than much higher than 1. Finally, note that β is not often close to 0, indicating that almost all managers producing forecasts in this data set look

at the model forecasts to some extent.

Figure 2.3 shows a histogram of posterior means for P_i for $i = 1, ..., N$. We see that the probability that $\alpha^* = \alpha = 0$ is often very high. In only 1030 of the 7250 cases the probability is lower than 0.5 and in 5469 cases it is higher than 0.9. Thus, part a of Hypothesis 2 does not seem to hold: not often is $\alpha \neq 0$ and is there a constant bias in the expert forecasts as compared to the model forecasts.

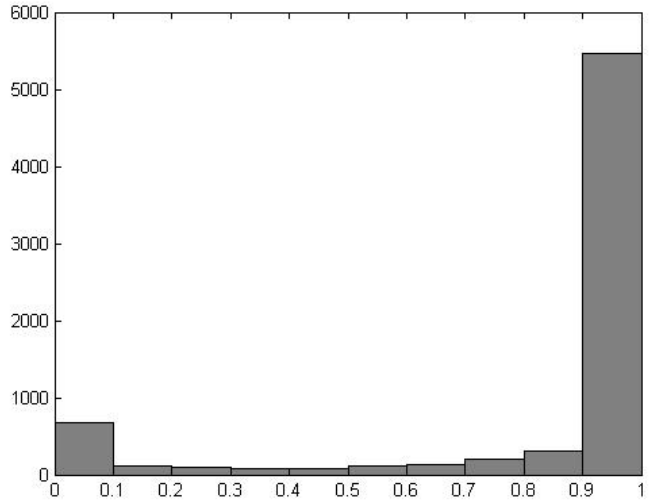

Figure 2.3: Histogram of posterior means for P_i in (2.18), for $i = 1, ..., N$.

Figure 2.4 shows a histogram of posterior means for $\alpha_i^\dagger$ for the cases for which the posterior mean for $P_i < 0.5$ and for which the posterior mean for $-200 < \alpha_i^\dagger < 4000$. The smallest estimated $\alpha_i^\dagger$ is -609.39 and the largest is 228587.73. Only 2 estimated $\alpha_i^\dagger$'s are smaller than -200, but still 135 are larger than 4000. Thus, looking at the histogram and at the values not included in the histogram, we can conclude that the estimated $\alpha_i^\dagger$'s are strongly positively skewed. Only in 44 of the cases is the estimated α negative, supporting part b of Hypothesis 2: when α is different from 0, it is often positive.

We observe that the first two hypotheses (1 and 2) are only partly validated. But to what extent are the expert forecasts positively biased, as is often found in previous

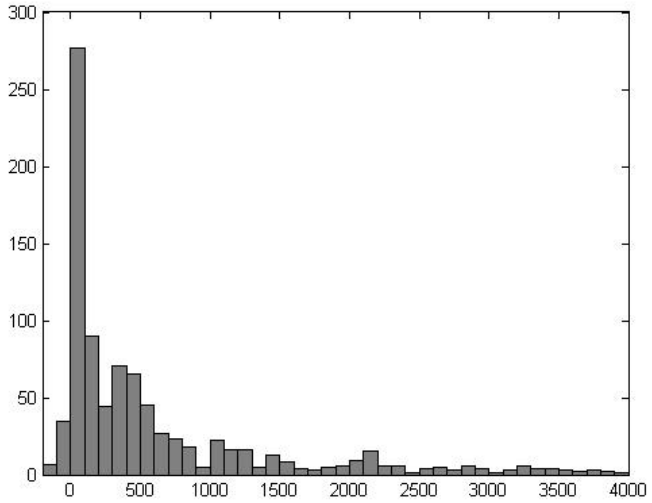

Figure 2.4: Histogram of posterior means for $\alpha_i^\dagger$ in (2.16) for which the posterior mean for $P_i < 0.5$ and the posterior mean for $-200 < \alpha_i^\dagger < 4000$, for $i = 1, ..., N$.

research (see Section 2.2.1)? This is the case when α^* is larger than 0, while β^* is 0 or also larger than 0, or when β^* is larger than 0, while α^* equals 0. We find that in only 2516 cases this seems to hold, which is a little over one third of the cases.

To see if the deviations of β^* from 0 follow the rules that hypothetically optimize the forecast improvement of EF over MF, we calculate the correlation between the posterior mean for β_i^* and $\beta_{i,opt} = \frac{\text{Cov}[MF_{i,t},R_{i,t}]-\text{Cov}[MF_{i,t}I_{i,t}]}{V[MF_{i,t}]}$ for $i = 1, ..., N$. In Section 2.3 we derived that the optimal value of β_i is given by this fraction in (2.9). We obtain a positive correlation of 0.11.

To get more insights, we also counted how often the posterior mean for β^* is positive while $(\text{Cov}[MF_{i,t}, R_{i,t}]-\text{Cov}[MF_{i,t}I_{i,t}]) > V[MF_{i,t}]$ plus how often the posterior mean for β^* is negative while $(\text{Cov}[MF_{i,t}, R_{i,t}] - \text{Cov}[MF_{i,t}I_{i,t}]) < V[MF_{i,t}]$. This appears to occur in 37% of the cases. The exact opposite is true in only 25% of the cases. Thus, according to (2.9), in 25% of the cases β^* has the wrong sign, while 37% has the correct sign. The remaining 2719 cases have a probability of 50% or higher that $\beta^* = 0$. For those cases, the difference between $\beta_{i,opt}$ and 1 is on average 0.69,

and $\beta_{i,opt}$ varies between -51.25 and 27.94 with a standard deviation of 1.80. For the complete data set these values are 0.77 (average difference from 1), -51.25 (minimum), 33.13 (maximum) and 1.89 (standard deviation). This all gives the impression that there are managers who recognize when β should be different from 1 and in which direction it should be different.

We also formulated hypotheses (3 and 4) about factors that might influence the value of α and β. To find out to what extent these hypotheses are valid for our data, we have to take a look at the posterior means for the parameters in the second level of the model, that is, γ_α, γ_β, ψ_α, ψ_β. Part of the estimated coefficients can be found in Table 2.2. First of all, we see support for part a of Hypothesis 4, that is, the average size of sales is positively related with α in (2.2). We find very strong posterior evidence that both the probability that α^* is different from 0 and the level of $\alpha^\dagger$ increase with the average size of sales.

We see that sales volatility has an opposite effect. The higher the volatility, the lower the probability that α^* differs from 0 and the lower the value of $\alpha^\dagger$. For both effects there is very strong posterior evidence. This contradicts part b of Hypothesis 4, as we expected that more volatile sales would make a manager to overpredict in order to prevent running out of stock.

Furthermore, we see that forecasts with a horizon of 2 to 7 months have on average a lower probability that α^* equals 0 as compared to forecasts with a horizon of just 1 month, with the horizon of 6 months having the lowest estimated coefficient. We also see a parabolic effect of the forecast horizon on $\alpha^\dagger$, with the highest $\alpha^\dagger$ for forecasts for 5 and 6 months ahead. Although this seems to contradict part c of Hypothesis 4, for this data these results are perfectly explainable. The management of the firm from which we use the forecasting and sales figures informed us that the 6-month horizon is an important planning horizon. This importance probably results in a suboptimal value for α.

Table 2.2: Posterior means (and standard deviations) for the parameters in the second level of the model about expert behavior, described in Subsection 2.4.1. Columns 2 to 5 contain the posterior means for part of γ_α, ψ_α, γ_β, and ψ_β, respectively.

Variable	κ	$\alpha^\dagger$	λ	$\beta^\dagger$
c	0.821	-35.361	-0.454	-0.061
	(0.124)	(4.501)	(0.154)	(0.019)
$\overline{R}$	-2.142e-05	0.592	-8.969e-06	-2.164e-06
	(3.868e-06)	(2.616e-04)	(2.897e-06)	(3.777e-07)
Vol(R)	2.077e-04	-0.354	4.477e-05	1.224e-05
	(2.997e-05)	(0.002)	(1.664e-05)	(2.188e-06)
Hor 2	-0.085	1.304	-0.030	-0.002
	(0.095)	(2.638)	(0.101)	(0.013)
Hor 3	-0.164	1.189	-0.061	-0.014
	(0.094)	(3.168)	(0.102)	(0.014)
Hor 4	-0.036	5.039	-0.131	-0.005
	(0.095)	(2.875)	(0.104)	(0.013)
Hor 5	-0.106	6.126	-0.181	-0.021
	(0.095)	(2.685)	(0.108)	(0.013)
Hor 6	-0.206	5.635	-0.368	-0.031
	(0.100)	(2.838)	(0.115)	(0.014)
Hor 7	-0.169	2.585	-0.557	-0.024
	(0.102)	(2.994)	(0.121)	(0.013)

For β we find a significantly negative effect of average sales volume on the probability that β^* is 0 and also a significantly negative relation between average sales volume and $\beta^\dagger$, both supporting Hypothesis 3a. However, we have to keep in mind that Hypothesis 3 was based on Hypothesis 1b stating that $\beta^\dagger$ would be smaller than 0, and that this hypothesis has already been shown to be incorrect: $\beta^\dagger$ is often larger

than 0. Thus, as long as $\beta^\dagger$ is smaller than 0, it moves in the expected direction when average sales volume increases, but when $\beta^\dagger$ is larger than 0, it moves in the same, but now unexpected direction. We calculated the average of $(\beta_i^\dagger)^2$ differentiated to each of the variables in z_i to see if the variables had an influence on $\beta_i^\dagger$ moving away or towards 0, but found only insignificant results. This confirms that the found relations are robust to a change of sign of $\beta_i^\dagger$.

An increase in the volatility of sales results in a higher probability that β^* equals 0 and in an increase in $\beta^\dagger$. As with the influence on α, this is not in line with what we hypothesized.

Finally, we see that the longer the forecast horizon the smaller the probability that $\beta^* = 0$ and that $\beta^\dagger$ is smallest for a forecast horizon of 6 months. This is in line with part c of Hypothesis 3, again modified for this data set, because the 6-month horizon is an important planning horizon.

The dummy variables for countries (and thus managers) and for medicine categories included in z_i are often significantly related to the four dependent variables.[3] Thus, on the basis of these results specific managers can be addressed when their α and/or β values are not optimal for (part of) their forecasts and can be given feedback.

We are also interested in the correlation between MF and I in (2.2). Hypothesis 5 stated that expert forecasts are often related to external factors which are also related to the model forecasts (endogeneity of MF in (2.2)). With Hypothesis 6 we stated that a lower or more negative correlation between MF and I in (2.2) might be beneficial to forecast accuracy. In order to evaluate the correlation between MF and I of the expert forecasts we first have to address two issues. First, we need to know if the instrument, which is the most recent model forecast error known at the moment of forecast creation, is a relevant instrument. We find that in more than 70% of the cases the posterior mean for δ in (2.22) is significantly different from 0, so we can conclude that we used a fairly relevant instrument.

[3]The estimated coefficients for these dummy variables are not shown here, but are available upon request.

Second, we need to know if the instrument is a valid instrument, that is, is it unrelated to expert forecasts? To that extent, we calculate the correlation between the estimated error terms in the first level of the model, $\varepsilon_{i,t}$, and the instrument. We find that the correlation in 2451 cases is < -0.3 and in 572 cases is > 0.3. Thus, the estimated $\beta_i^\dagger$ might be over- or underestimated and this might give a false impression on what it is the managers do. However, it is hard to find a better instrumental variable for this data set and the validity is certainly not completely rejected.

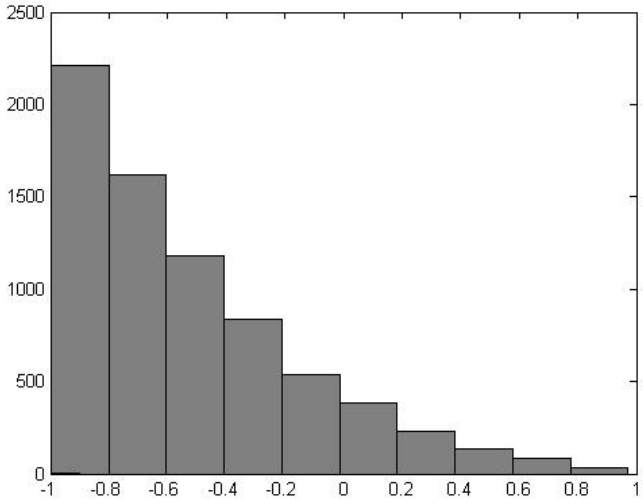

Figure 2.5: Histogram of posterior means for $\rho_{\Omega,i}$, correlation between $\varepsilon_{i,t}$ in (2.23) and $\zeta_{i,t}$ in (2.24), for $i = 1, ..., N$.

The endogeneity in (2.2) can now be measured by the correlation in the posterior mean for Ω_i, that is, by the posterior mean for $\rho_{\Omega,i} = \Omega_i(1,2)/\sqrt{\Omega_i(1,1)\Omega_i(2,2)}$ for all i. The estimated correlations are depicted in Figure 2.5. The result is surprising. We might expect positive correlations, indicating that factors influencing model forecasts influence the expert forecasts in the same way, resulting in double counting. However, we mainly find negative correlations (in almost 90% of the cases). This would mean that factors influencing the level of model forecasts have an opposite effect on expert forecasts. In Hypothesis 6c we stated that such a negative correlation would benefit the

forecast improvement of the expert forecasts over that of the model forecasts. In sum, it seems that the experts are properly adjusting model forecasts, but to what extent this is useful will be discussed in the next section.

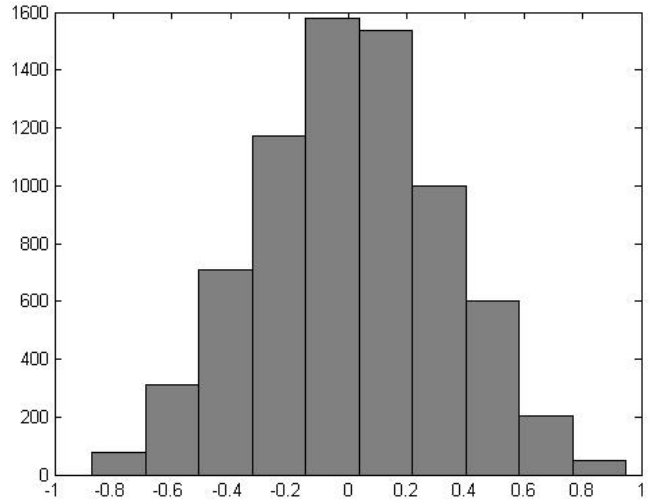

Figure 2.6: Histogram of the correlations between realized sales $R_{i,t}$ and the posterior mean for $\varepsilon_{i,t}$ from (2.15), for $i = 1, ..., N$.

Finally, it might be interesting to take a look at the correlation between the estimated error terms of the first level of the model, $\varepsilon_{i,t}$, and realized sales, as we have seen that this influences the forecast accuracy too. A histogram of these correlations can be found in Figure 2.6. The correlations are pretty much symmetrically centered around 0, with just a little more positive correlations than negative. This time, it would be preferred that the correlations is positive, see equation (2.4) and Hypothesis 6. However, the more model forecasts and realized values are related, the more difficult it is to add intuition to the model forecasts that is negatively related to model forecasts and positively related to the realized values. As we almost always see a negative endogeneity, this might explain why we also often see a negative relation between realized values and intuition. Probably the managers too often wrongly correct the model forecasts using factors also influencing these model forecasts, resulting in intuition I being

negatively correlated with the realized values R.

2.5.3 Forecast Evaluation

In Table 2.3 we give the estimated coefficients of model (2.30). First of all, we see that an expert who produces forecasts with $\alpha = 0$, $\beta = 1$ and no correlation between the residuals in (2.2) and the model forecasts and between the residuals in (2.2) and realized sales, performs on average better than the model. This can be seen from the sum of the estimated constant c and the estimated coefficients for the variables No intercept and Relation MF being positive. An expert who produces forecasts with α different from 0, β different from 1, but not larger than approximately 1.51 or smaller than approximately 0.49 and the correlations equal to 0, produces on average less accurate forecasts than the model. These values for the variables, that is, No intercept, Relation MF, Endogeneity and Intuition equal to 0, Size intercept positively valued and Size relation MF smaller than 0.51, multiplied by the estimated coefficients and summed up together with the estimated constant c, result in a negative $DRMSPE$.

Table 2.3: Estimated coefficients of the forecast evaluation model (2.30). Coefficients that are significantly different from 0 at the 5%-level are indicated by '*'.

Variable	**Estimated coefficient**
c	-454.136*
No intercept	46.071
Size intercept	-0.067*
Relation MF	459.968*
Size relation MF	887.618*
Endogeneity	-345.276*
Intuition	830.285*

A decrease in the probability that $\alpha = \alpha^* = 0$ or in the probability that $\beta^* = 0$ (equivalent to $\beta = 1$) both decrease on average the forecast accuracy of the expert forecasts as compared to the model forecasts. This confirms parts a and b of Hypothesis 6. Furthermore, we see a significantly negative coefficient for the size of the parameter $\alpha^\dagger$ which supports the second part of Hypothesis 6a.

The fifth estimated coefficient is not in line with Hypothesis 6b. According to this estimated coefficient, $\beta^\dagger$ moving away from 0 results on average in an increasing DRMSPE. Note however, that the variable 'Size relation MF' has to be larger than 0.518 in order to make up for the loss in accuracy due to $S \neq 1$. DRMSPE is on average approximately 460 higher for $S = 1$ than for $S = 0$, ceteris paribus, and only when $|\beta_i^\dagger(1 - S_i)| > 0.518$ is this same level of forecast accuracy improvement achieved. Of the 7250 cases, this happens only 139 times (looking at posterior means for the parameters), which is in less than 2% of the cases, thus in general it is still more beneficial to have $\beta = 1$ than $\beta \neq 1$.

The fact that values of β further away from 1 result in more accurate forecasts as compared to model forecasts than values of β closer to 1, has probably to do with the correlation between the optimal β and the bias and relative bias in model forecasts and the endogeneity of the model forecasts in (2.2). This is confirmed by the fact that we found a positive correlation between the optimal value of β_i and the estimated β_i^* in the previous section.

The next two estimated coefficients, corresponding to the correlation of intuition with model forecasts and of intuition with realized values, have the expected signs again. Hypotheses 6c and 6d get support as we find that a lower correlation between MF and I increases the forecast accuracy of expert forecasts and a higher correlation between intuition and realized sales increases the forecast accuracy of the expert forecasts.

Recall though from Section 2.3 that it is probably hard to achieve both a negative (or lower) correlation between intuition and model forecasts and a positive (or higher) correlation between intuition and realized values, as model forecasts and real-

ized values should be strongly related. Therefore we are interested to see how often the intuition of the expert increases the forecast accuracy relative to the model forecasts. According to the model this is the case when the sum of the variables Endogeneity and Intuition both multiplied by its estimated coefficient is positive. We find this to be true in 77% of the cases.

We can also look at (2.12), where we presented the theoretical condition under which intuition improves forecast accuracy. To test how often this is the case for our data we use $2[\text{Cov}(R_{i,t}, \varepsilon_{i,t}) - \beta\text{Cov}(MF_{i,t}, \varepsilon_{i,t})] > \text{Var}(\varepsilon_{i,t})$ for all i, with posterior means for $\varepsilon_{i,t}$. We find that only in 953 cases this inequality holds, and thus only in approximately 13% of the cases is intuition helpful in improving forecast accuracy.

We can conclude, at least for this data set, that the rules to follow for an expert formulated in Hypothesis 6 are a bit too simple and general. There seem to be experts who do recognize the situations in which the model forecasts are (relatively) biased and who are able to correct, at least partly, this bias. But, on average, an expert who does follow the rules formulated in Hypothesis 6 does perform better than the model and there are not many experts able to improve on the performance of this set of rules by choosing alternative values for α and β. Furthermore, it seems hard to improve the model forecasts by adding intuition.

2.6 Conclusions

Expert forecasts, created once statistical model forecasts are available, are quite often discussed in the literature, but still not much is known about how expert forecasts are created. Often the expert forecasts are analyzed on their forecasting performance without a proper analysis of what it is the experts actually did. In this paper we formulated hypotheses about the behavior of experts and about the impact of that behavior on forecast accuracy. We proposed a model to find out how expert forecasts are created in relation to model forecasts and to find out which factors influence this behavior. We proposed a novel and innovative two-level Hierarchical Bayes model in which we

also take into account that the model forecasts might be endogenous. The observed behavior could then be linked to forecasting performance.

We applied this model to a large data set consisting of model and expert forecasts and realizations of SKU-level sales data. The results for our data set were interesting and sometimes quite surprising. We found that in about one third of our expert forecasts there is a structural upward bias. There might be a bias in expert forecasts as compared to model forecasts, but at first it is unclear whether this is because the expert adds to the model forecasts or because the expert does not look at the model forecasts and creates own independent forecasts. We found that in approximately 37% of the cases there is a one-to-one relation between model forecasts and expert forecasts. In 50% of the remaining cases the expert reacts excessively to the model forecasts and in the other 50% of the remaining cases the expert only partially takes the model forecasts into account, if at all.

The intercept and the coefficient in the linear relation between expert forecasts and model forecasts were significantly influenced by factors such as average sales volume, sales volatility and forecasting horizon.

We furthermore found that the experts often take other factors into account that also influence the model forecasts. However, often this makes the expert forecasts to deviate in the opposite direction than that the model forecasts were influenced. Thus, we often find endogeneity of the model forecasts, or, to be more precise, a negative correlation between the model forecasts and the error terms in the linear relation between expert forecasts and model forecasts. Finally, we found different kinds of relations between the intuition of the experts (other factors than model forecasts influencing the expert forecasts) and the realized sales values.

Theoretically, when the model forecasts are unbiased and relative unbiased as compared to the realized values (see Section 2.3), then expert forecasts which are related to model forecasts in a linear relation with coefficient equal to 1 and intercept equal to 0, would be most accurate as long as intuition and model forecasts are unrelated. However, we find in our data set that the conditions for this (unbiasedness and rela-

tive unbiasedness of the model forecasts) are not always met, and that some experts are probably able to recognize this and correct for it. Furthermore, as soon as endogeneity of the model forecasts is introduced (correlation between intuition and model forecasts), things get more complicated and it is harder to draw straightforward conclusions about the optimal values of the coefficients and of the correlation between the residuals and model forecasts and of the correlation between the residuals and realized values. In general, experts who follow some simple rules, which optimize forecast performance under optimal circumstances, outperform the model forecasts in our data set. However, some experts who deviate from these rules, especially those for which β is further away from 1 and for which the error terms are negatively related to the model forecasts, also perform very well. We found that this has probably to do with the fact that these experts have to deal with poor model forecasts.

There are three main challenges in this area of research. The first is to apply the techniques described in this paper to other data sets. Our results are interesting and very informative, but are limited to the sales data of one company. It would be worthwhile using (sales) data from other companies or from other research areas, such as macroeconomics, to see if our results extend to other situations too.

The second challenge is to find and use appropriate instruments to deal with the endogeneity of model forecasts. Although we seemed to have done a pretty good job in our data set, the instrument we used is probably not perfect and this might influence the conclusions that we have drawn. In new research the most difficult task, besides finding a useful data set, is probably to find appropriate instruments.

Finally, in many forecasting situations only expert forecasts are available to the researcher and no model forecasts. It would be interesting to investigate ways to retrieve these model forecasts from the available data.

2.A Appendices

2.A.1 Typical data format

In this appendix we describe the data format as assumed in Section 2.4 and which is typical for forecast practices in which we have forecasts for multiple time periods and multiple variables. The data as described and used in Section 2.5 also follow this format. Let X be a general notation for the variables MF (model forecast), EF (expert forecast) and R (realized value). After cleaning up the data set (in which for example all forecasts for which no realizations are available are removed) the typical data format for X is as in Table 2.4.

The first four columns give the characteristics of X in the columns after that. The first column indicates which expert m receives the model forecasts and creates the expert forecasts. In case $X = R$ it indicates which expert created the expert forecasts for the realizations in that row. In total there are M experts.

The second column indicates for which variable (possibly product) the forecasts are created or to which variable (product) the realized values belong. Although different experts might produce forecasts, for example, for the same product in, for example, different geographical area's, we gave these variables a different index number j for the different experts and we analyze them as different variables. Thus, for a given forecast horizon (column 3) each variable number is unique and that variable is being forecasted by only one specific expert. Furthermore, the variables might be grouped into different (product) groups. This would result in an extra column with an index indicating to which group the variable belongs, but we did not depict such a column in Table 2.4. The first expert is responsible for J_1 variables and in total there are J variables being forecasted.

The third column shows for $X = MF$ and $X = EF$ for which forecast horizon the forecasts are created and for $X = R$ for which forecast horizon the belonging forecasts are created. H_j denotes the longest forecast horizon for product j. For different

Table 2.4: Typical data format for X, where X represents model forecasts MF, expert forecasts EF, realized values R or instrumental variable V.

Expert	**Product/ variable**	**Forecast horizon**	**Case**	**Time period in which forecast is created**				
m	j	h	i	$t=1$	$t=2$	$\ldots$	$t=T-1$	$t=T$
1	1	1	1	$X_{1,1}$	$X_{1,2}$	$\ldots$	$X_{1,T-1}$	$X_{1,T}$
1	1	2	2	$X_{2,1}$	$X_{2,2}$	$\ldots$	$X_{2,T-1}$	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	1	H_1	H_1	$X_{H_1,1}$	$X_{H_1,2}$	$\ldots$	NA	NA
1	2	1	H_1+1	$X_{H_1+1,1}$	$X_{H_1+1,2}$	$\ldots$	$X_{H_1+1,T-1}$	$X_{H_1+1,T}$
1	2	2	H_1+2	$X_{H_1+2,1}$	$X_{H_1+2,2}$	$\ldots$	$X_{H_1+2,T-1}$	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	2	H_2	H_1+H_2	$X_{H_1+H_2,1}$	$X_{H_1+H_2,2}$	$\ldots$	NA	NA
1	3	1	H_2+1	$X_{H_2+1,1}$	$X_{H_2+1,2}$	$\ldots$	$X_{H_2+1,T-1}$	$X_{H_2+1,T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	J_1	H_{J_1}	$\sum_{j=1}^{J_1} H_j$	$X_{\sum_{j=1}^{J_1} H_j,1}$	$X_{\sum_{j=1}^{J_1} H_j,2}$	$\ldots$	NA	NA
2	J_1+1	1	$\sum_{j=1}^{J_1} H_j+1$	$X_{\sum_{j=1}^{J_1} H_j+1,1}$	$X_{\sum_{j=1}^{J_1} H_j+1,2}$	$\ldots$	$X_{\sum_{j=1}^{J_1} H_j+1,T-1}$	$X_{\sum_{j=1}^{J_1} H_j+1,T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
M	J	H_J	N	$X_{N,1}$	$X_{N,2}$	$\ldots$	NA	NA

products, the largest forecast horizon might be different. Thus for the first product H_1 might be 7, while for the second product H_2 might be 5.

The fourth column sums up the information in the first three columns by a unique index number and indicates the cases. One case is one of the time series, thus one line in the table, and encompasses the forecasts and the realizations of those forecasts for which the expert forecasts are created by one and the same expert and for which the forecasts are created for one and the same product and over one and the same forecast horizon. The index i is an integer between 1 and N, the total number of cases.

Columns 5 to $T + 4$ give the forecasts as created in period t or the realizations belonging to those forecasts, thus the realizations in period $t + h$. Thus the first entry, $X_{1,1}$, gives the model forecast or expert forecast created in period $t = 1$ for period 2 or for $X = R$ it gives the realized value of period 2. The entry below that, $X_{2,1}$ gives the model forecast or expert forecast created in period $t = 1$ for period 3 or for $X = R$ it gives the realized value of period 3. For the table with $X = R$ the rows with different h, but the same j contain the same values, but in different columns. Thus the second row in Table 2.4 is the same as the first row, but the entries are shifted one column to the left and the third row is the same as the second row, but again the entries are shifted one column to the left and so on.

The maximum number of observations for a case is T, but as we have missing observations for some i, this results in T_i observations for case i.

Finally, note that the matrix with the values for the instrumental variable V as in (2.22) has the same format as for EF, MF and R. $V_{i,t}$ is the instrumental variable value for $MF_{i,t}$, where $V_{i,t}$ for our case study is as described in Section 2.5.1.

2.A.2 Parameter estimation

In this appendix we describe the method used to estimate the parameters of the two-level Hierarchical Bayes model described in Section 2.4.1. The Markov Chain Monte Carlo methodology is used, in particular, the Gibbs sampling technique in combination

with data augmentation.

Model parameters sampled are $\theta = (\{\beta_i^\dagger\}_{i=1}^N, \{\alpha_i^\dagger\}_{i=1}^N, \{\mu_i\}_{i=1}^N, \{\delta_i\}_{i=1}^N, \gamma_\alpha', \gamma_\beta', \psi_\alpha', \psi_\beta', \{\Omega_i\}_{i=1}^N, \sigma_\xi^2, \sigma_\eta^2)$. The latent variables P_i, S_i, q_i and w_i, $i = 1, ..., N$ are sampled alongside with the model parameters.

We apply in this appendix the more general notation $y_{i,t}$ for $EF_{i,t}$ and $x_{i,t}$ for $MF_{i,t}$. Furthermore, let y_i be a $T_i \times 1$ vector $(y_{i,1}, ..., y_{i,T_i})'$ with a similar definition for x_i and v_i.

To derive the likelihood function, we first consider the density function of the data $y_i = \{y_{i,t}\}_{t=1}^{T_i}$ and $x_i = \{x_{i,t}\}_{t=1}^{T_i}$ given P_i, S_i and θ:

$$f_{kl,i} = f(y_i, x_i | P_i = k, S_i = l, \theta) = \prod_{t=1}^{T_i} \Phi(y_{i,t}, x_{i,t} | m_{kl,i,t}, \Omega_i) \tag{2.32}$$

where Φ is the multivariate normal density function, k and l can take values 0 and 1, $m_{kl,i,t}$ is the mean vector when $P_i = k$ and $S_i = l$ and Ω_i is the covariance matrix. We have

$$\begin{aligned}
m_{11,i,t} &= (x_{i,t}, \mu_i + \delta_i v_{i,t})' \\
m_{01,i,t} &= (\alpha_i^\dagger + x_{i,t}, \mu_i + \delta_i v_{i,t})' \\
m_{10,i,t} &= (x_{i,t} + \beta_i^\dagger x_{i,t}, \mu_i + \delta_i v_{i,t})' \\
m_{00,i,t} &= (\alpha_i^\dagger + x_{i,t} + \beta_i^\dagger x_{i,t}, \mu_i + \delta_i v_{i,t})'.
\end{aligned} \tag{2.33}$$

The complete data likelihood is then

$$\begin{aligned}
f(\{y_i\}_{i=1}^N, \{x_i\}_{i=1}^N, \{P_i\}_{i=1}^N, \{S_i\}_{i=1}^N | \theta) = \prod_{i=1}^N & (f_{11,i}\kappa_i\lambda_i)^{P_i S_i} \\
& (f_{01,i}(1-\kappa_i)\lambda_i)^{(1-P_i)S_i} (f_{10,i}\kappa_i(1-\lambda_i))^{P_i(1-S_i)} \\
& (f_{00,i}(1-\kappa_i)(1-\lambda_i))^{(1-P_i)(1-S_i)} \phi(\alpha_i | z_i'\gamma_\alpha, \sigma_\xi^2) \phi(\beta_i | z_i'\gamma_\beta, \sigma_\eta^2),
\end{aligned} \tag{2.34}$$

with ϕ the normal density function.

We impose flat priors on most parameters. For the covariance of $\varepsilon_{i,t}$ and $\zeta_{i,t}$, thus for Ω_i, we use an inverted Wishart prior with $pr_{\Omega,sh} = 1$ degree of freedom

and scale parameter $pr_{\Omega,sc} = 100 * I_2$, where I_m denotes an m-dimensional identity matrix. For σ^2_ξ and σ^2_η, we use an inverted Gamma-2 prior with shape parameter $pr_{\sigma^2_\xi,sh} = pr_{\sigma^2_\eta,sh} = 1$ and scale parameters $pr_{\sigma^2_\xi,sc} = 0.001$ and $pr_{\sigma^2_\eta,sc} = 1$. Finally, for ψ_α and ψ_β we impose normal priors with mean 0 and covariance matrix $pr_{\psi_\alpha} = pr_{\psi_\alpha} = 4I_g$, where g is the number of variables in z_i. These priors are imposed to improve the performance of the algorithm and to reduce the number of iterations needed for convergence, but the influence of these priors on the posterior distribution is only marginal.

Sampling of P_i and S_i

The full conditional posterior distribution of P_i for $i = 1, ..., N$ is given by

$$Pr[P_i = 1|\theta, \text{data}] = \frac{\kappa_i(f_{11,i} + f_{10,i})}{\kappa_i(f_{11,i} + f_{10,i}) + (1 - \kappa_i)(f_{01,i} + f_{00,i})}, \tag{2.35}$$

and hence we can sample P_i from a Bernoulli distribution with parameters $n = 1$ and $p = Pr[P_i = 1|\theta, \text{data}]$. S_i can also be sampled from a Bernoulli distribution with $n = 1$, but with $p = Pr[S_i = 1|\theta, \text{data}]$, where

$$Pr[S_i = 1|\theta, \text{data}] = \frac{\lambda_i(f_{11,i} + f_{01,i})}{\lambda_i(f_{11,i} + f_{01,i}) + (1 - \lambda_i)(f_{10,i} + f_{00,i})}. \tag{2.36}$$

Sampling of q_i and w_i

The full conditional posterior distribution of q_i is

$$q_i|\theta, \text{data} \sim \begin{cases} N(z_i'\psi_\alpha, 1)I[q_i > 0] & \text{if } P_i = 1 \\ N(z_i'\psi_\alpha, 1)I[q_i \leq 0] & \text{if } P_i = 0, \end{cases} \tag{2.37}$$

which is in both cases the pdf of a truncated normal distribution. The inverse CDF technique is used to sample q_i. The sampling of w_i is analogous to the sampling of q_i, but then with ψ_β instead of ψ_α, with w_i instead of q_i and with S_i instead of P_i.

Sampling of ψ_α and ψ_β

To sample ψ_α, we notice that conditional on $\{z_i\}_{i=1}^N$ and on the sampled $\{q_i\}_{i=1}^N$ we have $q_i = z_i'\psi_\alpha + \nu_i$, with $\nu_i \sim N(0,1)$. Thus, ψ_α can be sampled from a multivariate normal distribution with mean $(\sum_{i=1}^N z_i'z_i' + pr_{\psi_\alpha}^{-1})^{-1}(\sum_{i=1}^N z_i q_i)$ and variance $(\sum_{i=1}^N z_i z_i' + pr_{\psi_\alpha}^{-1})^{-1}$. Following the same line of thought, ψ_β can be sampled from a multivariate normal distribution with mean $(\sum_{i=1}^N z_i'z_i' + pr_{\psi_\beta}^{-1})^{-1}(\sum_{i=1}^N z_i w_i)$ and variance $(\sum_{i=1}^N z_i z_i' + pr_{\psi_\beta}^{-1})^{-1}$.

Sampling of μ_i and δ_i

To derive the full conditional posterior of μ_i and δ_i, we need to take into account that $(\varepsilon_{i,t}, \zeta_{i,t})' \sim MN(0, \Omega_i)$. We therefore write,

$$x_{i,t} = \mu_i + \delta_i v_{i,t} + \rho(y_{i,t} - \alpha_i^\dagger(1-P_i) - x_{i,t} - \beta_i^\dagger x_{i,t}(1-S_i)) + e_{i,t}, \qquad (2.38)$$

with $\rho = \sigma_{\varepsilon\zeta,i}/\sigma_{\varepsilon,i}^2$ and $e_{i,t} \sim N(0, \sigma_{e,i}^2)$, where $\sigma_{e,i}^2 = \sigma_{\zeta,i}^2 - \sigma_{\varepsilon\zeta,i}^2/\sigma_{\varepsilon,i}^2$. Now μ_i and δ_i can be sampled from a multivariate normal distribution with mean $(\tilde{X}_i'\tilde{X}_i)^{-1}(\tilde{X}_i'\tilde{y}_i)$ and covariance $\sigma_{e,i}^2(\tilde{X}_i'\tilde{X}_i)^{-1}$, where $\tilde{X}_i$ is the $T_i \times 2$ matrix containing the constant and v_i and $\tilde{y}_i$ is the vector containing for every t in i $\tilde{y}_{i,t} = x_{i,t} - \rho(y_{i,t} - \alpha_i^\dagger(1-P_i) - x_{i,t} - \beta_i^\dagger x_{i,t}(1-S_i))$.

Sampling of Ω_i

Conditional on the other parameters, the covariance matrix Ω_i can be sampled from an inverted Wishart distribution with scale parameter $\sum_{t=1}^{T_i}(\varepsilon_{i,t}, \zeta_{i,t})'(\varepsilon_{i,t}, \zeta_{i,t}) + pr_{\Omega,sc}$ and degrees of freedom $T_i + pr_{\Omega_s h}$, with $\varepsilon_{i,t} = y_{i,t} - \alpha_i^\dagger(1-P_i) - x_{i,t} - \beta_i^\dagger x_{i,t}(1-S_i)$ and with $\zeta_{i,t} = x_{i,t} - \mu_i - \delta_i v_{i,t}$.

Sampling of γ_α and γ_β

We have $\alpha_i^\dagger = z_i'\gamma_\alpha + \xi_i$, $\forall P_i = 0$ and with $\xi_i \sim N(0, \sigma_\xi^2)$. Thus, γ_α can be sampled from a multivariate normal distribution with mean $(\sum_{i=1}^N z_i z_i'(1 -$

$P_i))^{-1}(\sum_{i=1}^{N} z_i\alpha_i^{\dagger}(1-P_i))$ and variance $\sigma_\xi^2(\sum_{i=1}^{N} z_i z_i'(1-P_i))^{-1}$. Similarly, we have $\beta_i^{\dagger} = z_i'\gamma_\beta + \eta_i$, $\forall S_i = 0$, with $\eta_i \sim N(0, \sigma_\eta^2)$. Thus, γ_β can be sampled from a multivariate normal distribution with mean $(\sum_{i=1}^{N} z_i z_i'(1-S_i))^{-1}(\sum_{i=1}^{N} z_i\beta_i^{\dagger}(1-S_i))$ and variance $\sigma_\eta^2(\sum_{i=1}^{N} z_i z_i'(1-S_i))^{-1}$.

Sampling of σ_ξ^2 and σ_η^2

Conditional on the data and the other parameters, σ_ξ^2 has an inverted Gamma-2 distribution with scale parameter $\sum_{t=1}^{N}\xi_i^2(1-P_i) + pr_{\sigma_\xi^2,sc}$ and degrees of freedom $\sum_{i=1}^{N}(1-P_i) + pr_{\sigma_\xi^2,sh}$, where we define $\xi_i = \alpha_i^{\dagger} - z_i'\gamma_\alpha$. To sample σ_ξ^2, we use that

$$\frac{\sum_{t=1}^{N}\xi_i^2(1-P_i) + pr_{\sigma_\xi^2,sc}}{\sigma_\xi^2} \sim \chi^2\left(\sum_{i=1}^{N}(1-P_i) + pr_{\sigma_\xi^2,sh}\right). \tag{2.39}$$

The sampling of σ_η^2 is analogous to the sampling of σ_ξ^2. Thus we have, conditional on the other parameters and data,

$$\frac{\sum_{i=1}^{N}\eta_i^2(1-S_i) + pr_{\sigma_\eta^2,sc}}{\sigma_\eta^2} \sim \chi^2\left(\sum_{i=1}^{N}(1-S_i) + pr_{\sigma_\eta^2,sh}\right), \tag{2.40}$$

where $\eta_i = \beta_i^{\dagger} - z_i'\gamma_\beta$.

Sampling of $\alpha_i^{\dagger}$

To sample $\alpha_i^{\dagger}$ we consider $\forall P_i = 0$,

$$y_{i,t} = \alpha_i^{\dagger} + x_{i,t} + \beta_i^{\dagger}x_{i,t}(1-S_i) + \rho(x_{i,t} - \mu_i - \delta_i v_{i,t}) + e_{i,t}, \tag{2.41}$$

with $\rho = \sigma_{\varepsilon\zeta,i}/\sigma_{\zeta,i}^2$ and $e_{i,t} \sim N(0, \sigma_{e,i}^2)$, where $\sigma_{e,i}^2 = \sigma_{\varepsilon,i}^2 - \sigma_{\varepsilon\zeta,i}^2/\sigma_{\zeta,i}^2$. Now we consider, again $\forall P_i = 0$,

$$\begin{aligned}\sigma_{e,i}^{-1}(y_{i,t} - x_{i,t} - \beta_i x_{i,t}(1-S_i) - \rho(x_{i,t} - \mu_i - \delta_i v_{i,t})) &= \sigma_{e,i}^{-1}\alpha_i^{\dagger} + \sigma_{e,i}^{-1}e_{i,t} \\ \sigma_\xi^{-1}z_i'\gamma_\alpha &= \sigma_\xi^{-1}\alpha_i^{\dagger} + \sigma_\xi^{-1}\xi_i \end{aligned} \tag{2.42}$$

Hence we have created a linear regression model with unit variances which can be written in vector notation

$$B = A\alpha_i^{\dagger} + d, \tag{2.43}$$

with $d \sim N(0, I)$ and where

$$B = (\sigma_{e,i}^{-1}(y_{i,1} - x_{i,1} - \beta_i x_{i,1}(1 - S_i) - \rho(x_{i,1} - \mu_i - \delta_i v_{i,1}),$$
$$\sigma_{e,i}^{-1}(y_{i,2} - x_{i,2} - \beta_i x_{i,2}(1 - S_i) - \rho(x_{i,2} - \mu_i - \delta_i v_{i,2}), ...,$$
$$\sigma_{e,i}^{-1}(y_{i,T_i} - x_{i,T_i} - \beta_i x_{i,T_i}(1 - S_i) - \rho(x_{i,T_i} - \mu_i - \delta_i v_{i,T_i}),$$
$$\sigma_{\xi}^{-1} z_i' \gamma_\alpha)'$$
$$A = (\sigma_{e,i}^{-1}, \sigma_{e,i}^{-1}, ..., \sigma_{e,i}^{-1}, \sigma_{\xi}^{-1})'. \tag{2.44}$$

Hence $\forall P_i = 0$, $\alpha_i^\dagger$ can be sampled from a normal distribution with mean $(A'A)^{-1}(A'B)$ and variance $(A'A)^{-1}$.
$\forall P_i = 1$ we sample $\alpha_i^\dagger$ from a normal distribution with mean $z_i' \gamma_\alpha$ and variance σ_ξ^2.

Sampling of $\beta_i^\dagger$

To sample $\beta_i^\dagger$ we consider $\forall S_i = 0$,

$$y_{i,t} = \alpha_i^\dagger(1 - P_i) + x_{i,t} + \beta_i^\dagger x_{i,t} + \rho(x_{i,t} - \mu_i - \delta_i v_{i,t}) + e_{i,t}, \tag{2.45}$$

with ρ and $e_{i,t}$ as defined above for the sampling of $\alpha_i^\dagger$. Now the sampling of $\beta_i^\dagger$ is analogous to the sampling of $\alpha_i^\dagger$. So we consider $\forall S_i = 0$,

$$\sigma_{e,i}^{-1}(y_{i,t} - \alpha_i^\dagger(1 - P_i) - x_{i,t} - \rho(x_{i,t} - \mu_i - \delta_i v_{i,t})) = \beta_i^\dagger(\sigma_{e,i}^{-1} x_{i,t}) + \sigma_{e,i}^{-1} e_{i,t}$$
$$\sigma_\eta^{-1} z_i' \gamma_\beta = \sigma_\eta^{-1} \beta_i^\dagger + \sigma_\eta^{-1} \eta_i, \tag{2.46}$$

and we have created a linear regression model with unit variances again and $\beta_i^\dagger$ can be sampled from a normal distribution.
Again, $\forall S_i = 1$ we sample $\beta_i^\dagger$ from a normal distribution with mean $z_i' \gamma_\beta$ and variance σ_η^2.

Chapter 3

Do experts SKU forecasts improve after feedback?

Joint work with Philip Hans Franses.

3.1 Introduction

Much empirical and experimental research is dedicated to the analysis of forecasts from experts who receive statistical model forecasts and then can quote their own forecasts by possibly adjusting the model forecasts. The main focus in this research is usually on forecast quality, that is, do the experts improve or deteriorate forecast accuracy? Theoretically, the experts should be able to improve the statistical model forecasts (see for example Goodwin, 2000) and in some instances they were found to do so (see for example Blattberg and Hoch, 1990; Mathews and Diamantopoulos, 1992; Fildes et al., 2009). Other empirical evidence, however, suggests that often the experts increased the forecast error (see for example Franses and Legerstee, 2010) and hence, more research is needed to understand what it is that the managers do and how this relates to forecast accuracy (see for example Sanders, 1992; Fildes and Goodwin, 2007b; Fildes et al., 2009; Franses and Legerstee, 2009, 2010; Legerstee et al., 2011).

The literature provides many recommendations to experts when they create their forecasts. Examples of such instructions range from making less adjustments (Fildes and Goodwin, 2007b; Franses and Legerstee, 2009), to making smaller-sized adjustments (Franses and Legerstee, 2010) or, in contrast, making larger-sized adjustments (Fildes et al., 2009; Trapero et al., 2010) and making less upward adjustments (Franses and Legerstee, 2009; Fildes et al., 2009). We believe that these instructions are all rather vague, hard to measure and to quantify and sometimes even contradictory. It is therefore questionable to what extent experts are able to improve their forecasts on the basis of those recommendations. For example, would telling the experts to adjust less often upwards result in improved forecast accuracy? On the other hand, quite some research exists on possible forecast improvement as a result of different kinds of feedback on judgmental forecasts, see Lawrence et al. (2006). However, these outcomes are usually based on laboratory experiments and do not include actual forecasters.

In this paper we aim to contribute to the literature by presenting the results of a natural experiment in which the actual experts, who are responsible for final SKU-level sales forecasts, receive information on the model used to create the statistical model forecasts and receive performance and cognitive process feedback. We have the statistical model forecasts, the final forecasts and the realized values of SKU-level sales for the period *before* and *after* the experts received that extra information and feedback. The feedback was based on the discussion and summary statistics presented in Franses and Legerstee (2009, 2010). By collecting the same kind of data for the period after the feedback we now have a unique opportunity to assess the impact of the feedback and model information.

In the next section we discuss the literature on feedback where we focus on forecasts by experts. After that we give the setting of our research and describe the data and the novelty of our research. In the third section we describe the results, where we first present the total results and after that the results per expert. This last part also shows how variations in adjustments based on the feedback result in variations in forecast improvements. The final section concludes with a summary of the main findings

and also discusses some limitations that provide challenges for further research.

3.2 Literature on feedback

Most of the literature on feedback dates back to the eighties and nineties and is based on laboratory experiments, see Lawrence et al. (2006). These authors provide an excellent overview of the literature on judgemental forecasting up to 2006, in which they separately discuss feedback and judgmental adjustments. The experiments consider the effects of different kinds of feedback on the accuracy of forecasts provided by different kinds of students. Focus of study are the forecast accuracy of point forecasts, probability forecasts and judgmental prediction interval forecasts before and after feedback. In our empirical work below we deal with point forecasts and hence we focus on the effect of feedback on such forecasts.

Although the labels and descriptions of the different kinds of feedback vary a little bit across the literature, it seems that a general distinction can be made between outcome feedback, performance feedback and cognitive process feedback. The first simply provides the forecaster with the realized values of the variable for which forecasts were generated. This type of feedback is most common in practice as forecasters are usually able to observe the actual data for the past few periods for which they created forecasts. However, as other types of feedback often show how to improve the forecast accuracy, outcome feedback is typically found to be the least effective, see also Lawrence et al. (2006). As stated in Goodwin and Fildes (1999b, p. 41) and Lawrence et al. (2006, p. 507), forecasters seem to be unable to filter the noise component from the realized values of the variable to be predicted and to assess systematic inadequacy in their forecasts.

The second type of feedback is performance feedback and it provides the forecaster with information on forecast accuracy with statistics such as the root mean squared prediction error. Remus et al. (1996) did not find evidence in their laboratory experiment that performance feedback improves forecasting practices as compared to outcome

feedback. In contrast, pertaining to judgmental interval predictions and pertaining to probability forecasts, Bolger and Önkal-Atay (2004) and Stone and Opel (2000) do find that performance feedback improves the forecasts. Furthermore, in the *Principles of Forecasting Handbook* (Armstrong, 2001b; Armstrong and Pagell, 2003) two of the principles, identified by 40 international researchers to increase forecast quality, state that forecasting methods should be compared on their past performance and feedback on forecasts should be sought (see also www.forecastingprinciples.com). Interestingly, Fildes and Goodwin (2007a) and Gönül et al. (2009) find in their surveys that these principles are not often followed. In fact, of the respondents only 75% and 35.5%, respectively, indicated to use performance feedback. Gönül et al. (2009) further investigate the reasons for adjusting externally acquired financial and economic forecasts. Getting performance feedback on the external forecasts is shown to result in more adjustments and in less reliance on other factors to determine whether to adjust (such as information on the source of the forecasts).

Another study on performance feedback worth noting is Athanasopoulos and Hyndman (2011). To our knowledge, this recent study on feedback is the only study that is not based on a laboratory experiment or a survey, as it is based on an online forecasting competition. This study shows that performance feedback significantly improves forecasting accuracy, although the setting is a bit different from most actual situations. After submission of the forecasts, the forecasters get performance feedback based on a random unknown portion of the forecasts and are able to resubmit a new set of forecasts. This is different from most laboratory settings and real-world situations in which a forecaster gives a forecast for time t, receives feedback on it at $t + 1$ and can then give a forecast for time $t + 1$.

The third kind of feedback is cognitive process feedback and it gives the forecaster information on his own forecasting practices. Such information can include how the forecaster reacts to certain cues or the behavior needed to improve the forecasts. Remus et al. (1996) found no evidence that cognitive process feedback might be helpful (in addition to task properties feedback, see below for a description) and this is a con-

firmation of the results found by Balzer et al. (1992) pertaining to probability forecasts. Lim et al. (2005) showed that the effectiveness of this type of feedback might be improved by the way the feedback is presented, that is multimedia messages might be more effective than textual messages.

A separate kind of feedback is task properties feedback, which is sometimes also called environmental feedback. It involves providing the forecaster with statistical information on the variable to be forecasted. It can encompass data characteristics or statistical model forecasts. Note that it might be argued that this is not genuine feedback as it is provided before the judgmental forecast is given and it is not feedback on the performance of the judgmental forecaster, see Björkman (1972). This task properties feedback has received most attention in research on feedback on judgmental forecasting, see Remus et al. (1996), Sanders (1997), Welch et al. (1998) and Goodwin and Fildes (1999b). In all cases it is found to improve forecast accuracy and in general it is found to be the most effective form of feedback (Lawrence et al., 2006).

Forecasters usually receive a statistical model forecast before stating their own judgmental forecasts in the case of judgmental adjustments, and as such this can be viewed as task properties feedback. Goodwin and Fildes (1999b) investigate if providing a statistical model forecast improves forecast accuracy and also if providing additional information on the statistical forecasts helps to further improve the judgmental forecasts. Both seems to be the case, although two remarks can be made. First, the statistical model forecasts appear not to be used efficiently, as is confirmed in Franses and Legerstee (2010). Second, providing information for trend-seasonal series did not improve the forecasting results, possibly due to problems of the subjects to comprehend this information.

In the next sections we study how various types of feedback can lead to different forecasts, where we study actual experts and actual feedback in a natural experiment. Although the results from previous research are sometimes contradictory, we expect that feedback in general results in more accurate judgmental forecasts. How the experts are expected to change their behavior in order to achieve higher accuracy is discussed

in the next section.

3.3 Setting

The natural experiment that we present is based on SKU-level sales data from a large pharmaceutical company. The data concern forecasts for monthly sales of pharmaceutical products in many countries and for various horizons. Final forecasts EF are delivered by experts who first receive statistical model forecasts MF created using (a version of) ForecastPro. The performance of the experts is assessed by their forecasting accuracy and part of their bonus depends on it. In Franses and Legerstee (2009, 2010) the behavior and effectiveness (in terms of accuracy) of the experts is analyzed. The analysis was performed using about two years of monthly data, covering September 2004 to September 2006.

The main conclusion from these studies is that the managers responsible for creating the final forecasts deviate too much from the statistical model forecasts. It is found that the difference between EF and MF is predictable, while it should not be, and that MF receives too small a weight in the final EF forecasts. The experts make frequent adjustments and these tend to be upwards. As a result the expert forecasts are either equally accurate as the model forecasts or much less accurate. When $EF - MF$ increases, that is, when the size of the upward adjustment becomes larger, it is found that forecast performance is deteriorated.

In August-September of 2007 the managers (experts) responsible for forecasting received feedback by way of a presentation at the headquarters' office. They received three kinds of feedback. First of all they received cognitive process feedback, as statistics were presented to the managers on their behavior in adjusting the model forecasts. Second, they received performance feedback, in the form of accuracy measures of their forecasts. Finally, they received more information and explanation on the statistical models used to create the forecasts. So, although they already received task properties feedback in the form of the statistical model forecasts, this type of feed-

back is extended by the extra information given at the headquarters' office. We have benchmark observations for the period in which the managers received outcome and simple task properties feedback. We also have new forecasts for the period after the presentation, in which the experts received cognitive process feedback, performance feedback and additional task properties feedback. We are now interested in studying the behavior and performance of the experts *before* and *after* the feedback session.

We use a data set that contains forecasts created in September 2006 to December 2007. In 2008 the pharmaceutical company was acquired by another company and many managers who were responsible for the forecasts left. So, data after December 2007 cannot be used for our purposes. We restrict our focus to 1-step-ahead forecasts and only the observations for products for which forecasts and realizations are available for all 16 ($t = 1, .., 16$) months are retained. We compare the data for September 2006 to September 2007 (first sample, 8411 observations) with the data for October 2007 to December 2007 (second sample, 1941 observations). The final forecasts are created by 21 managers located in as many countries.

We first address the behavior of experts. We consider judgmental adjustment, defined as

$$Adj_{i,t} = EF_{i,t} - MF_{i,t}, \tag{3.1}$$

and relative adjustment, defined as

$$AdjR_{i,t} = (EF_{i,t} - MF_{i,t})/MF_{i,t}, \tag{3.2}$$

where $EF_{i,t}$ is the SKU-level expert forecast created in month t for month $t+1$ and for product i. We present in the next section various statistics of these (relative) forecast adjustments for the periods before and after the feedback session and we test if any differences between these statistics are significant.

The second issue is whether any changes in behavior lead to changes of forecast accuracy. For that purpose we use the difference between absolute forecast errors of the expert forecasts and absolute forecast errors of the model forecasts, defined by

$$Err_{i,t} = |R_{i,t+1} - EF_{i,t}| - |R_{i,t+1} - MF_{i,t}|, \tag{3.3}$$

where $R_{i,t+1}$ is the realization of SKU-level sales of product i in month $t+1$ corresponding to the forecasts created the month before. Furthermore, we look at the relative difference in absolute forecast error, that is,

$$ErrR_{i,t} = (|R_{i,t+1} - EF_{i,t}| - |R_{i,t+1} - MF_{i,t}|)/R_{i,t+1}. \tag{3.4}$$

For this variable we also present various statistics and test results to see if before and after the feedback session performance has changed.

To compare the statistics in both samples we use the common large-sample test as described in Wackerly et al. (2002b).

3.4 Results

In this next section we first analyze the statistics and test results as they are computed for all the experts together. After that we consider the same statistics and test results but then computed per expert to see if there are significant differences and to see how possible changes in behavior of the experts influences forecast accuracy.

3.4.1 All experts

Behavior

In Table 3.1 we present statistics and test results for expert adjustments Adj (see equation (3.1)) and relative expert adjustments $AdjR$ (see equation (3.2)). The first observation that is noticeable is that there is a large and significant difference between average adjustments before the experts received feedback and after that session. As we also immediately see from the table that there is not a significant difference (p-value is 0.478) between average absolute adjustments, we might conclude that the experts make more negative adjustments in the second sample and that this causes the difference in average adjustments.

Table 3.1: Summary statistics for forecast adjustments Adj (see equation (3.1)) and relative forecast adjustments $AdjR$ (see equation (3.2)). The second column shows the statistics for the first sample (September 2006 - September 2007) and the third column those for the second sample (October 2007 - December 2007). The final column gives one-sided p-values of the test for the difference between the statistics in the two samples (if available).

	Sample 1	**Sample 2**	**p-value**
Mean Adj	212.000	44.041	0.006
Std Adj	2832.890	2634.060	.
Mean $\|Adj\|$	940.431	936.987	0.478
Std $\|Adj\|$	2680.616	2462.075	.
Mean $AdjR$	0.154	0.070	0.003
Std $AdjR$	2.699	0.399	.
Mean $\|AdjR\|$	0.275	0.202	0.008
Std $\|AdjR\|$	2.689	0.351	.
Mean $Adj(Adj \neq 0)$	212.606	44.756	0.007
Std $Adj(Adj \neq 0)$	2836.919	2655.355	.
Mean $\|Adj(Adj \neq 0)\|$	943.122	952.195	0.444
Std $\|Adj(Adj \neq 0)\|$	2683.976	2479.065	.
Mean $AdjR(Adj \neq 0)$	0.155	0.071	0.004
Std $AdjR(Adj \neq 0)$	2.703	0.402	.
Mean $\|AdjR(Adj \neq 0)\|$	0.276	0.205	0.010
Std $\|AdjR(Adj \neq 0)\|$	2.693	0.353	.

We also see that the standard deviation of adjustments and absolute adjustments is much smaller in the second sample. The averages of relative adjustments and absolute relative adjustments also get significantly smaller. This indicates that the forecast adjustments have smaller variation in the second sample than in the first and that the

adjustments in the second sample are on average relatively smaller than the adjustments in the first sample.[1]

Table 3.2: Distribution of relative forecast adjustments $AdjR$ (see equation (3.2)). The second column shows the distribution of the first sample and the third column the distribution of the second sample. The final column gives one-sided p-values of the test for the difference between the fractions in the two samples. $I[\cdot]$ is an indicator function which takes a value of 1 if the statement between brackets is true and is 0 otherwise.

	Sample 1	**Sample 2**	**p-value**
Mean $(AdjR > 0)$	0.571	0.536	0.003
Mean $(AdjR < 0)$	0.426	0.448	0.040
Mean $I[AdjR < -1]$	0.000	0.000	1.000
Mean $I[-1 \leq AdjR < -0.75]$	0.003	0.003	0.466
Mean $I[-0.75 \leq AdjR < -0.5]$	0.010	0.011	0.258
Mean $I[-0.5 \leq AdjR < -0.25]$	0.056	0.060	0.218
Mean $I[-0.25 \leq AdjR < 0]$	0.357	0.373	0.100
Mean $I[AdjR = 0]$	0.003	0.016	0.000
Mean $I[0 < AdjR < 0.25]$	0.404	0.366	0.001
Mean $I[0.25 \leq AdjR < 0.5]$	0.098	0.107	0.130
Mean $I[0.5 \leq AdjR < 0.75]$	0.032	0.032	0.482
Mean $I[0.75 \leq AdjR < 1]$	0.011	0.015	0.119
Mean $I[AdjR \geq 1]$	0.026	0.016	0.003

To get more insights, we also consider the fraction of zero-adjustments, that is,

[1] We used the variance test as described in Wackerly et al. (2002c) to test if the difference between the variances is significant for the variables Adj, $AdjR$, Err and $ErrR$. Test results showed highly significant differences for all four variables. However, this test requires that the variable for which the variance is being tested is normally distributed and this is never the case. Therefore, test results are not reliable and omitted.

how often is the model forecast unadjusted anyway? Table 3.2 shows that this fraction is less than 0.003 in the first sample, while in the second sample this fraction increased significantly to 0.016. Thus the feedback the managers received made them to adjust less often, although the model forecasts are still adjusted very frequently. Second, we see a significant decline in positive adjustments of 3.5% (from 0.571 to 0536) and a little bit smaller but also significant increase in negative adjustments of 2.2% (see Table 3.2 again). Hence, there is indeed a shift from positive adjustments to no adjustments and negative adjustments. Considering the fact that before the feedback at the headquarters' office the percentage of positive adjustments was around 57.1 and the percentage of negative adjustments was around 42.6, this results in more balance between positive and negative adjustments in the second sample, although there is still a clear difference between the two.

Is the decline in mean of adjustments, absolute adjustments, relative adjustments and absolute relative adjustments completely due to the increase in the number of no adjustments or are the adjustments that are made in the second sample also smaller than before? To answer that question we calculate these four statistics while leaving out the zero-adjustment observations and we test if they differ significantly across the two samples. In the second panel of Table 3.1 we see that the mean of adjustments, relative adjustments and absolute relative adjustments still decline significantly or almost significantly (largest p-value is 0.1) once we leave out the zero-adjustment observations. Hence, the adjustments that are made after the feedback are relatively smaller than before, but not in an absolute sense as the mean of absolute adjustments increases slightly.

If we take a closer look at the distribution of $AdjR$ before and after the feedback we see that the differences exist mainly in the relatively small adjustments (Table 3.2 fifth row from below). However, we also see a significant decline of approximately 1% in the amount of extremely large positive relative adjustments (larger than 100% of the size of the model forecast). Furthermore, note that the number of large forecast adjustments (larger than 25% of the size of the model forecast, but smaller than 100%

of the size of the model forecast) did not change significantly, both for negative and positive adjustments.

From Tables 3.1 and 3.2 we can conclude that the experts truly incorporated the feedback as they changed their forecasting behavior. They adjust less often and the adjustments that they still do are relatively smaller on average and there is more balance between positive and negative adjustments. However, one may feel that there is still room for improvement as adjustments still happen more often upward than downward and also around 25% of the forecasts still are associated with large or extremely large adjustments (larger than 25% of the size of the model forecast).

Forecast accuracy

In Table 3.3 we present statistics and test results for the differences in absolute forecast error Err (see equation (3.3)) and the relative differences in absolute forecast error $ErrR$ (see equation (3.4)). The first row shows a promising result. Where the experts perform worse than the model forecasts, after feedback the forecast accuracy increases substantially (p-value of 0.066). We furthermore see a decrease in the standard deviation of the differences in absolute forecast error. So not only is the difference on average lower, there is also less variation in the differences.

The average relative difference $ErrR$ also decreases, from EF being 1.8% of R less accurate than MF, to EF being 0.2% of R less accurate than MF, see the bottom two rows of Table 3.3. However, this improvement is not significant. Thus the forecast improvement as measured by the mean of Err is mainly achieved by improvements that are small relative to R.

In Table 3.4 we observe that the fraction of positive Err and $ErrR$, which is the fraction of forecasts where the managers deteriorate forecast accuracy, is lower in the second sample as compared to the first, with a p-value of 0.069. We also see that the fraction of negative Err and $ErrR$ increases after the feedback, but that this increase is not significant. As might be expected from the significant increase in no adjustments,

Table 3.3: Summery statistics for the differences between absolute forecast errors of the expert and absolute forecast errors of the model forecast, Err (see equation (3.3)), and the relative differences in absolute forecast error $ErrR$ (see equation (3.4)). The second column shows the statistics for the first sample and the third column shows the statistics for the second sample. The final column gives one-sided p-values of the test for the difference between the statistics in the two samples (if available).

	Sample 1	**Sample 2**	**p-value**
Mean Err	9.652	-69.891	0.066
Std Err	2218.144	2066.510	.
Mean $ErrR$	0.018	0.002	0.289
Std $ErrR$	1.304	1.028	.

we also find a significant increase in the number of forecasts with no difference in forecast accuracy between EF and MF.

In the remainder of Table 3.4 we see that the change in distribution of $ErrR$ largely follows the change in distribution of $AdjR$ in Table 3.2. The number of large deteriorations increases slightly (0.010 to 0.013) and the number of large improvements decreases slightly, possibly as a result of the slightly more large adjustments. The number of small deteriorations decreases significantly, as does the number of small positive adjustments. Finally, the number of forecasts with no difference and with small improvements in forecast accuracy increases, although the last one not significantly.

In sum, we can conclude that there is a small but significant improvement in forecast accuracy of EF over MF after feedback, and it seems to be related to the way the adjustments changed. It seems that the managers have partly changed their behavior as predicted by previous research and as a result of new feedback. Also, the changes have resulted in the expected improvements in forecast accuracy. There does seem to be room for further improvement though.

In the next subsection we analyze the behavior and forecast accuracy across the 21

Table 3.4: Distribution of the relative differences in absolute forecast error $ErrR$ (see equation (3.4)). The second column shows the distribution of the first sample and the third column shows the distribution of the second sample. The final column gives one-sided p-values of the test to see if there is a difference between the fractions in the previous two columns. $I[\cdot]$ is an indicator function which takes a value of 1 if the statement between brackets is true and is 0 otherwise.

	Sample 1	**Sample 2**	**p-value**
Mean $(ErrR > 0)$	0.502	0.483	0.069
Mean $(ErrR < 0)$	0.494	0.499	0.349
Mean $I[ErrR < -1]$	0.010	0.013	0.141
Mean $I[-1 \leq ErrR < -0.75]$	0.006	0.004	0.063
Mean $I[-0.75 \leq ErrR < -0.5]$	0.015	0.012	0.114
Mean $I[-0.5 \leq ErrR < -0.25]$	0.055	0.059	0.239
Mean $I[-0.25 \leq ErrR < 0]$	0.408	0.411	0.390
Mean $I[ErrR = 0]$	0.004	0.018	0.000
Mean $I[0 < ErrR < 0.25]$	0.409	0.374	0.002
Mean $I[0.25 \leq ErrR < 0.5]$	0.057	0.066	0.084
Mean $I[0.5 \leq ErrR < 0.75]$	0.018	0.023	0.075
Mean $I[0.75 \leq ErrR < 1]$	0.007	0.010	0.069
Mean $I[ErrR \geq 1]$	0.011	0.010	0.350

managers, to see if there exist large differences across the managers and whether these differences result in different forecast accuracy.

3.4.2 21 experts

In our data set there are 21 managers producing final forecasts, where each manager is responsible for the forecasts in a specific country. In this section we focus on relative forecast adjustment (equation (3.2)) and relative difference in absolute forecast error

(equation (3.4)) for each expert separately. We only look at $AdjR$ and $ErrR$, because the forecasts and sales figures substantially differ in size across the countries, so a comparison of Adj and Err is hard in this case.

Changes in behavior and accuracy

First we look at the distribution of some statistics concerning $AdjR$ and $ErrR$, see Table 3.5. In the first row of this table we see for the first sample that the mean of $AdjR$ per manager ranges between 0.008 and 0.921, so it is always positive. In the second sample this mean ranges between -0.078 and 0.314, so both minimum and maximum are lower than in the first sample, and hence the distribution has shifted. If we consider the 21 differences (one for each manager) between the mean of the second sample and the mean of the first sample, we see that these differences range between -0.812 and 0.096 with 14 of these differences being negative (see last column). Clearly, two-thirds of the managers decreased their average relative adjustments as a result of the received feedback.

The standard deviation of the mean of $AdjR$ has also decreased quite significantly, see the second row of Table 3.5. The maximum standard deviation changed from 9.309 to 0.934 and 15 of the 21 managers decreased the variation of their relative adjustments.

For absolute $AdjR$ we see the same patterns, see rows 3 and 4 of Table 3.5. Although the minimum of the mean and the minimum of the standard deviation of this variable have hardly changed, the maximum of the mean decreased from 1.060 to 0.422 and the maximum of the standard deviation decreased from 9.304 to 0.879, respectively. Furthermore, 14 of the managers decreased their average relative adjustments in absolute sense and 16 of the managers decreased the variation of absolute relative adjustments. Hence, the size of the adjustments is on average lower for 67% of the managers and is less extreme for 76% of the managers.

For the number of zero-adjustments and the number of positive adjustments we

Table 3.5: Summary statistics of the distributions of relative adjustment statistics and relative difference in error statistics across 21 managers. Columns 2 and 3 give the minimum and maximum of the statistics calculated over the first sample. Columns 4 and 5 give the minimum and maximum calculated over the second sample. Columns 6 and 7 give the minimum and maximum of the difference between the statistics (statistic second sample minus statistic first sample). The last column shows the number of times that the difference is negative, except for Mean $I[AdjR = 0]$ for which it shows the number of times that the difference is positive and for Mean $I[AdjR > 0]$ for which it shows the number of times that it approaches 0.5.

	Sample 1		**Sample 2**		**Diff.**		**Diff. Opt.**
	min	**max**	**min**	**max**	**min**	**max**	**nr.**
Mean $AdjR$	0.008	0.921	-0.078	0.314	-0.812	0.096	14
Std $AdjR$	0.156	9.309	0.171	0.942	-9.112	0.077	15
Mean $\|AdjR\|$	0.103	1.060	0.103	0.422	-0.790	0.120	14
Std $\|AdjR\|$	0.118	9.304	0.119	0.879	-9.157	0.090	16
Mean $I[AdjR = 0]$	0.000	0.031	0.000	0.301	-0.031	0.299	4
Mean $I[AdjR > 0]$	0.470	0.667	0.307	0.769	-0.189	0.237	11
Mean $ErrR$	-0.125	0.582	-0.388	0.253	-0.603	0.168	13
Std $ErrR$	0.160	5.846	0.115	3.275	-5.218	2.575	12
Mean $(ErrR > 0)$	0.423	0.677	0.323	0.795	-0.253	0.280	12

do not observe many changes. There are 8 managers who always adjusted before the feedback session at the headquarters' office and who always adjusted after that meeting. Only 4 managers increased the number of no adjustments relative to the number of forecasts. As the increase in no adjustments was significant over the complete group of forecasts (so for all managers together, see previous subsection), we might already suspect that those four managers substantially increased the number of no adjustments. We see indeed that the maximum of the fractions of no adjustments increased from 0.031 to 0.301.

What is most obvious from the fractions of positive adjustments is that the variation of these fractions increased. The minimum decreased (0.307), the maximum increased (0.769). Only 11 of the managers brought the fraction closer to the value of 0.5.

Both the minimum and maximum of the mean of relative differences in forecast accuracy decreased, as we would expect to see (the lower the $ErrR$, the more accurate is the manager as compared to the model). A little bit over 60% of the managers improved their forecasts as compared to the model forecast and relative to the size of the realization. Furthermore, both the minimum and maximum standard deviation of $ErrR$ decreased and 12 managers reduced the variation in $ErrR$. Although we would have expected the same for the fraction of $ErrR$ that is positive, in contrast we see an increase in the maximum of the fractions of $ErrR$ that is larger than zero (bottom row of Table 3.5). The minimum does decrease however and also 12 managers were able to decrease the fraction.

Relation between behavior and accuracy

We now turn to analyze whether the changes in adjustments relate to any changes in forecast accuracy. Do less adjustments and less positive adjustments result in better expert forecasts? To answer that question we use a linear regression with as dependent variable the difference between mean $ErrR$ in the second sample versus the mean $ErrR$ in the first sample. As independent variables we use the differences in the mean of $|AdjR|$, the differences in the standard deviation of $AdjR$ and the differences in how close the fraction of positive adjustments were to 0.5. Estimation results based on Ordinary Least Squares appear in Table 3.6.

The first observation from these estimation results is that there is a highly significant and positive relation between the change in mean absolute relative forecast adjustment and the change in mean relative difference in forecast accuracy. The more a manager decreased the relative adjustments in absolute sense on average, the more the manager was able to improve average relative forecast accuracy as compared to

Table 3.6: This table shows the estimated parameters with p-values, F-statistic with p-value and R-squared statistic of a linear regression of the difference in mean of relative differences in absolute forecast error (equation (3.4)) (dependent variable) and the differences in mean of absolute relative forecast adjustment (equation (3.2)), the differences in standard deviation of relative forecast adjustment and the differences in how close the fraction of positive adjustments is to 0.5. Differences are as measured between the second and first sample of the data and estimation is done for 21 observations using OLS.

Variable	**Coef.**	**Prob.**
Constant	-0.020	0.485
Diff. Mean $\|AdjR\|$	0.968	0.001
Diff. Std $AdjR$	-0.051	0.024
Diff. $\|$Mean $(AdjR > 0) - 0.5\|$	0.437	0.145
F-statistic	9.381	0.001
R-squared	0.623	.

model forecasts. So indeed, smaller adjustments do better.

The next independent variable shows an interesting result. Ceteris paribus, the more the standard deviation in relative forecast adjustments increased, the more the average relative forecast accuracy as compared to model forecasts improved. This result is significant at a 5% significance-level. Hence, although the adjustments should be smaller in size on average, an increase in variation resulted in a lower $ErrR$. Apparently, managers were better able to identify when and how to adjust, instead of careless adjustments of model forecasts.

The last variable shows the expected sign. If a manager made positive and negative adjustments closer to $50/50$, forecast accuracy improved. This parameter is significantly different from zero at 14.5%, which given the small sample size of only 21 could be considered as significant.

The fact that the first variable has a significantly positive parameter and the second variable a significantly negative parameter also indicates that replacing positive with

negative adjustments increases forecast accuracy as compared to making only positive or only negative adjustments. If average adjustment is positive, then the second parameter indicates that adjustments should fluctuate more around that mean. The first parameter indicates that the adjustments should be more close to zero. If the average adjustment is negative, then the second parameter also indicates that adjustments should fluctuate more around that mean, whereas the first parameter indicates that the adjustments should be closer to zero.

3.5 Conclusions

We analyzed forecast adjustments of experts, before and after giving these experts feedback and we examined if feedback improved subsequent forecast accuracy due to changes in behavior. We answered that question by analyzing the data from an actual natural experiment. In that experiment we considered the adjustments and forecast errors of SKU-level sales data of both before and after the managers who are responsible for the forecasts, received cognitive process feedback, performance feedback and extra information on task property feedback. The cognitive process feedback and performance feedback was based on the empirical results obtained in Franses and Legerstee (2009, 2010).

We clearly observe that the managers changed their forecast adjustment behavior significantly and in directions that could be beneficial. They adjust significantly less frequently, significantly less upwards and significantly more downwards. Furthermore, we have seen that the average of the adjustments, the average of the relative adjustments and the average of the absolute but relative adjustments decreased significantly, while the average of the absolute adjustments did not. This, together with decreased standard errors of adjustments and the changes in distributions of relative adjustments, shows that relatively extremely large positive adjustments and relatively small positive adjustments are replaced by zero-adjustments and relatively small negative adjustments. Even though many changes were significant, we concluded that

feedback could have resulted in even larger changes in behavior.

If we look at forecast accuracy, we are optimistic. Average forecast accuracy of the expert compared to that of the model increased significantly and changed from poorer performance to better performance. This increase can be largely ascribed to small improvements relative to the size of the realized sales. This can be seen from the fact that relative forecast accuracy of the expert compared to the model does increase, but not significantly. Next we observe less volatility in the forecast errors, significantly less forecast deteriorations by the experts, significantly more no improvements and also more forecast improvements.

We also compared the behavior and accuracy of the 21 different managers separately. We see large differences in the level of adaptation of the managers to the feedback. Furthermore, the way they changed their behavior influences the change in forecast accuracy significantly. Smaller adjustments in an absolute sense and more balance between the amount of positive and negative adjustments clearly increases forecast accuracy. However, ceteris paribus, more variation in the adjustments also improves the forecast accuracy. In sum, we can conclude that cognitive process feedback, performance feedback and extra information on the statistical model used to create the model forecasts results in more accurate expert forecasts than if the forecasters only receive outcome feedback and simple task performance feedback by way of statistical model forecasts.

Our study clearly shows that it is useful to examine what forecasters do and what the results are in terms of forecast accuracy. Presenting this information to the forecasters appears useful in practice. Of course, analyzing more data sets, analyzing forecasts from other companies and other forecasting areas, would result in even more reliable conclusions.

The fact that we analyze a natural experiment is the strength and novelty of this research, but of course also implies some limitations. We were not able to set the experimental design. Hence, we are not able to make a distinction between the individual effects of each of the feedback types, that is, of the information provision on the

statistical model, the performance feedback and the cognitive process feedback. We were also not able to have a control group. We hope to have a chance to run a natural experiment again in the future where we can accommodate these limitations.

Chapter 4

Estimating Loss Functions of Experts

Joint work with Philip Hans Franses and Richard Paap.

4.1 Introduction

Sales forecasts are often the outcome of a process in which an expert with domain-specific knowledge modifies a model-generated forecast. Typically, simple extrapolation models are used to create such model forecasts, and often they are generated by automated statistical software which gets fed by lagged sales and other possibly relevant variables.

There is a long tradition in the sales forecasting literature to examine the quality of these expert forecasts relative to model forecasts (if these are available). Key questions are whether the domain-specific knowledge translates into improved forecasts, or whether experts downplay the model forecasts too much, thereby quoting less accurate forecasts. Classical studies are Blattberg and Hoch (1990) and Mathews and Diamantopoulos (1986) where various case studies are examined.

Recently this literature has seen a revived interest with the advent of a range of large data sets that allow for more generalizing statements. For example, Fildes et al. (2009) study thousands of expert and model forecasts, and conclude that expert fore-

casts tend to be biased and that expert forecasts are not necessarily better than model forecasts. Franses and Legerstee (2010), using a database with over 30,000 forecasts and realizations, show that, on average, model forecasts and expert forecasts are about equally good, but when expert forecasts are worse they are much worse.

A common finding in these two recent studies is that expert forecasts tend to exceed model forecasts, or in other words, judgmental adjustment is often positive. A potential explanation for this finding is that the experts dislike underpredicting more than overpredicting, perhaps due to planning reasons. Hence, when creating forecasts their loss function may not be a mean squared error (MSE) loss function symmetric around zero, but some other, asymmetric loss function. If such an alternative loss function is used indeed, this may then also explain why expert forecasts seem less accurate than model forecasts, as typically forecasts are evaluated using criteria like the root mean squared prediction error (RMSPE).

The loss function of experts is usually not known in practice. Given available data, one may however try to estimate this loss function by evaluating theoretical properties of loss functions against actual data. Various forms of asymmetric loss functions have been proposed in the literature, like, for example, the lin-lin loss function, the quad-quad loss function and the linex function proposed by Varian (1975). These loss functions have been frequently analyzed, for example, by analyzing the optimal forecast under a specific asymmetric loss function, see Zellner (1986) and Christoffersen and Diebold (1996, 1997), among others.

In this paper we are interested in estimating the parameters of loss functions given the availability of expert forecasts, on which not much work exists. Clatworthy et al. (2011) investigate whether financial analysts' loss functions are asymmetric or not, but they do not estimate the loss function. A notable exception is Elliott et al. (2005). These authors propose a linear Instrumental Variable (IV) estimator for the shape parameter of a general class of loss functions which signals the degree of asymmetry in the loss function. The general class of loss functions nests four popular loss functions, and these are the absolute deviation loss function and its asymmetric counterpart the

lin-lin loss function, and the squared loss function and its asymmetric counterpart the quad-quad loss function. They use their methodology to estimate the asymmetry in forecasts of budget deficits for the G7 countries made by the IMF and OECD. Elliott et al. (2008) use the same methodology to estimate the asymmetry in survey forecasts of real output growth and inflation and to develop a more general method for testing forecast rationality jointly with asymmetric loss.

To estimate the loss function of experts in the sales forecasting industry we propose a methodology that differs from the methodology proposed by Elliott et al. (2005) in a number of ways. By making a normality assumption on the conditional distribution of the variable to be forecasted, and by that on the forecast distribution, we demonstrate that the estimation of the asymmetry parameter is simplified substantially. Elliott et al. (2005) need instrumental variables for their estimation method, but in our proposed methodology only simple linear regressions (OLS) are used, using panel data on expert forecasts and on the variable to be forecasted. If the normality assumption is valid, OLS is more efficient than using instrumental variables and the methodology can easily be extended to multiple-step ahead forecasts. Our proposed method can be used to estimate the key parameters of the well-known and useful linex loss function.

The outline of our paper is as follows. In Section 4.2 we show that for two well-defined loss functions, the lin-lin loss function and the linex loss function, simple regressions can be used to estimate the asymmetry parameter of the functions, provided the availability of the relevant data. In Section 4.3 we illustrate this methodology for a large database covering forecasts from a range of experts. We also consider statistical model forecasts to establish to what extent symmetric loss functions prevail. The robustness of our crucial assumption on the forecast distribution is tested in three ways. One way, for example, is to compare our estimates with those obtained with the methodology of Elliott et al. (2005). Upon estimating our two loss functions we find overwhelming support for the conjecture that experts may feel that negative forecast errors (meaning the forecasts are below actual sales) require more weight in the loss function than positive forecast errors. Section 4.4 concludes this paper with a summary

and suggestions for further research.

4.2 Loss Functions

Suppose that Y_{t+1} is the random variable to be forecasted with forecast density $f(y_{t+1}; \theta, \mathcal{Y}_t, \mathcal{X}_t)$ that may depend on parameters θ and lagged values $\mathcal{Y}_t = \{y_{t+1-j}\}_{j=1}^{J}$ and other exogenous variables summarized in $\mathcal{X}_t$. To simplify notation we write $f(y_{t+1}; \theta)$ instead of $f(y_{t+1}; \theta, \mathcal{Y}_t, \mathcal{X}_t)$. In this paper we confine our analysis to one-step ahead forecasts.

Given the forecast distribution, a point forecast p_{t+1} for Y_{t+1} can be obtained by specifying a loss function. For example, the quadratic loss function is given by

$$\text{QL}(Y_{t+1}, p_{t+1}) = (p_{t+1} - Y_{t+1})^2, \tag{4.1}$$

where we adopt the convention that a forecast error is the forecast minus the realization. The point forecast $\hat{p}_{t+1}$ results from minimizing expected quadratic loss $\text{E}[\text{QL}(Y_{t+1}, p_{t+1})]$ with respect to p_{t+1}, where E denotes the expectation operator. In case of quadratic loss, this results in $\hat{p}_{t+1} = \text{E}[Y_{T+1}|\theta]$. Hence, the optimal forecast is unbiased.

From a supply chain management point of view it can be necessary to put a higher penalty on negative forecast errors than on positive forecast errors. For example, if one forecasts sales, the consequences of a prediction which is lower than the realized demand may be worse than a prediction which is higher than the demand. In other words, being out of stock is worse than having a little too much stock. To allow for different penalties one may then consider an asymmetric loss function.

4.2.1 Asymmetric absolute loss function

An example of an asymmetric function is the lin-lin loss function, further also called the asymmetric absolute loss (AAL) function, which is given by

$$\text{AAL}(Y_{t+1}, p_{t+1}) = \begin{cases} \alpha_A |p_{t+1} - Y_{t+1}| & \text{if } p_{t+1} \leq Y_{t+1} \\ |p_{t+1} - Y_{t+1}| & \text{if } p_{t+1} > Y_{t+1}. \end{cases} \tag{4.2}$$

One sets $\alpha_A > 1$ if one wants to put more penalty on a forecast which is smaller than the true realization, see also Ferguson (1967). The optimal point forecast is obtained by minimizing expected loss, that is,

$$\mathrm{E}[\mathrm{AAL}(Y_{t+1}, p_{t+1})] = \int \mathrm{AAL}(y_{t+1}, p_{t+1}) f(y_{t+1}; \theta) dy_{t+1}. \tag{4.3}$$

The expected loss function $\mathrm{E}[\mathrm{AAL}(Y_{t+1}, p_{t+1})]$ can be written as

$$\int_{-\infty}^{p_{t+1}} (p_{t+1} - y_{t+1}) f(y_{t+1}; \theta) dy_{t+1} + \int_{p_{t+1}}^{\infty} \alpha_A (y_{t+1} - p_{t+1}) f(y_{t+1}; \theta) dy_{t+1}. \tag{4.4}$$

The first-order partial derivative is given by

$$\begin{aligned} \frac{\partial \mathrm{E}[\mathrm{AAL}(Y_{t+1}, p_{t+1})]}{\partial p_{t+1}} &= \int_{-\infty}^{p_{t+1}} f(y_{t+1}; \theta) dy_{t+1} - \int_{p_{t+1}}^{\infty} \alpha_A f(y_{t+1}; \theta) dy_{t+1} \\ &= F(p_{t+1}; \theta) - \alpha_A (1 - F(p_{t+1}; \theta)), \end{aligned} \tag{4.5}$$

where we used the Leibniz integral rule and where $F(\cdot; \theta)$ is the forecast distribution function of Y_{t+1} (with $f(\cdot; \theta)$ as its derivative). The optimal point forecast is obtained when this derivative is set equal to zero and solved for p_{t+1}, which results in

$$F(\hat{p}_{t+1}; \theta) = \frac{\alpha_A}{1 + \alpha_A}. \tag{4.6}$$

The point estimate corresponds to the $\alpha_A/(1+\alpha_A)$th percentile of the forecast distribution. Under symmetric loss ($\alpha_A = 1$) we obtain the median of the forecast distribution. For $\alpha_A > 1$ we have a forecast which is larger than the median, and for $\alpha_A < 1$ we obtain a forecast which is smaller than the median.

Hence, apparent biased forecasts of an expert may be due to the fact that an asymmetric loss function is used. Our main claim in this paper is that if we were to observe several forecasts of experts together with realizations of the forecasts, it is possible under some testable assumptions to estimate the value of α_A, see also Subsection 4.2.3 below.

Suppose that we have data with T forecasts where for each point forecast created at time $t = 1, ..., T$ the conditional forecast distribution is normal with mean m_t and variance s_t^2. Furthermore, assume that all forecasts are constructed using the same

asymmetric absolute loss function. Under these assumptions the forecasts are thus generated by

$$p_{t+1} = m_t + s_t \Phi^{-1}\left(\frac{\alpha_A}{1+\alpha_A}\right), \tag{4.7}$$

where Φ^{-1} is the inverse CDF of a standard normal distribution.

Further assume that the realizations y_{t+1} result from a normal distribution with mean μ_t and variance σ_t^2 for $t = 1, \ldots, T$ and hence $y_{t+1} = \mu_t + \sigma_t \eta_t$, where η_t is a realized draw from a standard normal distribution. If there is a systematic bias in the forecast distribution it holds that $m_t = \mu_t + b$ with $b \neq 0$. If we consider the difference between p_{t+1} and y_{t+1} we obtain

$$(p_{t+1} - y_{t+1}) = b + s_t \Phi^{-1}\left(\frac{\alpha_A}{1+\alpha_A}\right) - \sigma_t \eta_t. \tag{4.8}$$

If we can obtain a consistent estimate of s_t and σ_t, one can use the simple regression

$$\frac{(p_{t+1} - y_{t+1})}{\hat{\sigma}_t} = \frac{1}{\hat{\sigma}_t}\beta_0 + \frac{\hat{s}_t}{\hat{\sigma}_t}\beta_1 + \varepsilon_t \tag{4.9}$$

to provide the estimate for $\beta_0 = b$ and for $\beta_1 = \Phi^{-1}(\alpha_A/(1+\alpha_A))$. An estimate of α_A can easily be obtained by solving

$$\frac{\alpha_A}{1+\alpha_A} = \Phi(\beta_1) \Rightarrow \alpha_A = \frac{\Phi(\beta_1)}{1-\Phi(\beta_1)}.$$

In sum, in this scenario it is possible for a forecaster to have an asymmetric loss function and a systematic bias in its forecasting distribution. The expression in (4.9) shows that it is possible to calibrate the loss function and the bias.

4.2.2 Linex loss function

An alternative nonlinear asymmetric loss function is the linear-exponential function, also called the linex (LIN) loss function, see Varian (1975) and Zellner (1986). This function is given by

$$\text{LIN}(Y_{t+1}, p_{t+1}) = \exp(\alpha_L(p_{t+1} - Y_{t+1})) - \alpha_L(p_{t+1} - Y_{t+1}) - 1 \tag{4.10}$$

with $\alpha_L \neq 0$. A negative value of α_L implies that a p_{t+1} lower than Y_{t+1} is more costly than a p_{t+1} higher than Y_{t+1}. To be more precise, if $\alpha_L < 0$, the linex loss function shows an almost exponential increase in loss to the left of the origin $(p_{T+1} - Y_{t+1}) = 0$ and an almost linear increase in loss to the right of the origin. A positive value of α_L implies the opposite and a $\alpha_L \to 0$ implies symmetric loss. Zellner (1986) shows that the point forecast which minimizes expected loss is given by

$$\hat{p}_{t+1} = -\alpha_L^{-1} \log \mathrm{E}[\exp(-\alpha_L Y_{t+1})]. \tag{4.11}$$

Hence, if we assume that the forecast distribution of Y_{t+1} is normal with mean m_t and variance s_t^2, then the point forecast is given by

$$p_{t+1} = m_t - \frac{1}{2}\alpha_L s_t^2. \tag{4.12}$$

Again it is possible to estimate α_L in case we observe several forecasts of experts together with realized forecasts. Under the same conditions as above and using the same arguments, taking the difference between p_{t+1} and y_{t+1} and dividing by $\hat{\sigma}_t$ results in the simple regression

$$\frac{(p_{t+1} - y_{t+1})}{\hat{\sigma}_t} = \frac{1}{\hat{\sigma}_t}\beta_0 + \frac{\hat{s}_t^2}{2\hat{\sigma}_t}\hat{s}_t^2\beta_1 + \varepsilon_t. \tag{4.13}$$

OLS provides the estimate for the systematic bias $b = \beta_0$ and asymmetry parameter $\alpha_L = -\beta_1$.

4.2.3 Parameter estimation

To run the regressions (4.9) and (4.13) we need estimates of s_t^2 and σ_t^2. If we have the availability of unbiased model forecasts ($mf_{t+1} = \mathrm{E}[y_{t+1}|\mathcal{Y}_t, \mathcal{X}_t]$) and the realizations y_{t+1}, the variance of the data can be estimated using

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(y_{t+1} - mf_{t+1})^2 \tag{4.14}$$

under the assumption that $\sigma_t^2 = \sigma^2$.

The variance of the forecast distribution of the expert s_t^2, however, cannot be estimated from the variance of the available expert forecasts as these forecasts may be biased and/or result from an asymmetric loss function. To estimate the variance we assume that s_t^2 is constant ($s_t^2 = s^2$ for $t = 1, \ldots, T$) and that the variance of the expert is equal to the variance of the forecast distribution of an econometric model which fits the data at hand and produces unbiased forecasts.

Because σ_t and s_t are constant over t we need panel data with expert forecasts and realizations in order to estimate the parameters in (4.9) and (4.13). In other words, if we have point forecasts for variables $i = 1, ..., N$ over periods $t = 1, ..., T$, denoted by $p_{i,t+1}$ and $mf_{i,t+1}$, we are able to estimate $\hat{s}_i^2$ and $\hat{\sigma}_i^2$ for each i. In case of the lin-lin loss function we can now estimate the bias and asymmetry parameter with the regression

$$\frac{(p_{i,t+1} - y_{i,t+1})}{\hat{\sigma}_i} = \frac{1}{\hat{\sigma}_i}\beta_0 + \frac{\hat{s}_i}{\hat{\sigma}_i}\beta_1 + \varepsilon_{i,t}, \tag{4.15}$$

where $b = \beta_0$ and $\alpha_A = \Phi(\beta_1)/(1 - \Phi(\beta_1))$. In case of the linex function we can estimate the bias and asymmetry parameter with

$$\frac{(p_{i,t+1} - y_{i,t+1})}{\hat{\sigma}_i} = \frac{1}{\hat{\sigma}_i}\beta_0 + \frac{\hat{s}_i^2}{2\hat{\sigma}_i}\beta_1 + \varepsilon_{i,t}, \tag{4.16}$$

where $b = \beta_0$ and $\alpha_L = -\beta_1$.

4.2.4 Misspecification

Under our assumptions the error terms $\varepsilon_{i,t}$ for $i = 1, ..., N$ and $t = 1, ..., T$ should be normal with mean 0 and variance 1 in regressions (4.15) and (4.16). If this is not the case, (some of) the assumptions, such as the assumption of a normal forecast distribution, may not be valid or the loss function may not be adequate. It is therefore important to test if the estimated residuals are standard normally distributed.

If tests show that the error terms are not standard normally distributed or if there are other reasons to doubt whether the forecast density is normal, it is also possible to assume that the forecasts are lognormally distributed in case of the lin-lin loss function

(AAL). Under this distribution, the forecasts are generated by

$$\log(p_{t+1}) = m_t + s_t \Phi^{-1}\left(\frac{\alpha_A}{1+\alpha_A}\right), \tag{4.17}$$

where m_t is the mean and s_t^2 the variance of $\log(p_{t+1})$ and Φ^{-1} is the inverse CDF of a standard normal distribution. Assume now that the realizations y_{t+1} result from a lognormal distribution with parameters μ_t and σ_t^2 for $t = 1, \ldots, T$ and hence $\log(y_{t+1}) = \mu_t + \sigma_t \eta_t$ where ε_t is a realized draw from a standard normal distribution. We can now write

$$(\log(p_{t+1}) - \log(y_{t+1})) = b + s_t \Phi^{-1}\left(\frac{\alpha_A}{1+\alpha_A}\right) - \sigma_t \eta_t, \tag{4.18}$$

where b is again the systematic bias in the forecast distribution, thus $m_t = \mu_t + b$. Using the estimates of s_t and σ_t and using the relevant panel data the regression

$$\frac{(\log(p_{i,t+1}) - \log(y_{i,t+1}))}{\hat{\sigma}_i} = \frac{1}{\hat{\sigma}_i}\beta_0 + \frac{\hat{s}_i}{\hat{\sigma}_i}\beta_1 + \varepsilon_{i,t} \tag{4.19}$$

provides $\beta_0 = b$ and $\beta_1 = \Phi^{-1}(\alpha_A/(1+\alpha_A))$ and hence $\alpha_A = \Phi(\beta_1)/(1-\Phi(\beta_1))$. Again, if the assumptions are correct, including the assumption of lognormality of the forecast distribution, and the loss function is AAL, the error terms $\varepsilon_{i,t}$ for $i = 1, ..., N$ and $t = 1, ..., T$ should be normal with mean 0 and variance 1.

Another way to check if the assumptions are correct is to compare the results for the AAL loss function with the results as found with the method of Elliott et al. (2005). They use as a general loss function

$$L(Y_{t+1}, p_{t+1}) = [\alpha_E + (1 - 2\alpha_E) \cdot I(Y_{t+1} - p_{t+1} < 0)]|Y_{t+1} - p_{t+1}|^q, \tag{4.20}$$

where $I[\cdot]$ is an indicator function which takes a value of 1 if the statement between brackets is true and is 0 otherwise, where $\alpha_E \in (0, 1)$ and where they impose $q = 1$ or $q = 2$. By setting $q = 1$, the AAL loss function is obtained as defined above in (4.2), but with weight α_E for cases where $p_{t+1} \leq Y_{t+1}$ and with weight $1-\alpha_E$ for cases where $p_{t+1} > Y_{t+1}$. Stated differently, $\alpha_E/(1-\alpha_E) = \alpha_A$. Elliott et al. (2005) do not make assumptions on the distribution of the forecasts. Therefore, if the normality assumption

is valid, their methodology should result in an $\hat{\alpha}_E$ for which $\hat{\alpha}_E/(1-\hat{\alpha}_E) \approx \hat{\alpha}_A$, where $\hat{\alpha}_A$ is obtained from (4.15). Differences between $\hat{\alpha}_E/(1-\hat{\alpha}_E)$ and $\hat{\alpha}_A$ might be a result of the chosen instrumental variables for the estimation of α_E or the use of $\hat{s}$ and $\hat{\sigma}$ instead of s and σ for the estimation of α_A or both.

In the next section we will illustrate the techniques and robustness checks described in this section for a range of forecasts made by many experts.

4.3 Illustration

We apply our methodology to an extensive panel data set. The data set covers SKU-level sales data and is described in detail in the next subsection. In Subsections 4.3.2 and 4.3.3 the results of our analysis are discussed.

4.3.1 Data set

For our case study we use monthly sales data of a large pharmaceutical company. The company has its headquarters in The Netherlands, and has local offices in various countries worldwide. The company uses an automated statistical package to create forecasts using lagged sales data as the only input. The experts know that these data are the only input. Each month model selection and parameter estimation are updated, whereby the package uses techniques such as Box-Jenkins and Holt-Winters. These model forecasts are then sent to the managers/experts in the local offices, after which they quote their own forecasts.

The forecasts are available for the months November 2004 through November 2006. They are created for various horizons, but we only use the 1-step-ahead forecasts in the analysis presented in this paper. In each country, forecasts are created by a different expert and hence we have forecasts for 35 countries and thus 35 distinct individuals. For confidentiality reasons we denote the countries with roman numbers I to XXXV. Forecasts are created for 1038 different products. In the notation of the

previous section this means that i ranges from 1 to 1038. Per product we have a minimum of 15 and a maximum of 25 observations for which the model forecast, the expert forecast and realized sales are available to us. Thus, T depends on i and $15 \leq T_i \leq 25$ for $i = 1, ..., 1038$. All together, we have 24897 observations.

We denote the model forecasts as constructed by the statistical program of the company as MF, and the final forecasts from the experts are denoted as EF. The model that we use to estimate σ_i and s_i is for each i an AR(1) model for which the parameters are estimated over all available observations for i. For $mf_{i,t+1}$ $\forall i$ and $\forall t$ we consider the in-sample forecasts generated by these AR(1) models. Note that $MF_{i,t+1}$ and $mf_{i,t+1}$ are different forecasts, the first is the statistical model forecast as used by the company and the second is the forecast from the AR(1) model used to estimate σ_i and s_i.

The parameters in (4.15), (4.16) and (4.19) are estimated for each expert separately by multiplying the two variables in the regressions by dummy variables for the managers. We also estimate α_E per expert. Observations per expert range from 96 to 2132 with an approximate average of 710 observations.

4.3.2 Estimated asymmetry

We begin by analyzing the results as obtained under the assumption that the AAL function is used by the experts. Column 2 in Table 4.1 presents the estimated asymmetry parameter α_A per expert. We see that 26 of the 35 experts have an $\hat{\alpha}_A$ that is significantly different from 1 at a significance level of 10%. For 21 of these managers the difference is even significant at the 1% significance level. For all those 26 managers the $\hat{\alpha}_A$ exceeds 1, meaning that sales forecasts that are too low are penalized more than forecasts that are too high. On average, over 35 experts, $\hat{\alpha}_A$ has a value of 1.40, which indicates that too low forecasts are weighted 40% heavier than too high forecasts. To get some more insight into this value for $\hat{\alpha}_A$, see Figure 4.1.

Table 4.1: The estimated asymmetry parameter α_A of AAL and estimated systematic bias b following from regression (4.15), for each expert. Columns 2 and 3 show results for expert forecasts and Columns 4 and 5 for statistical model forecasts. The asterisks in the second and fourth column indicate if the $\hat{\alpha}_A$'s are significantly different from 1 and the asterisks in the third and fifth column indicate if the $\hat{b}$'s are significantly different from 0, where one is for the 10%, two are for the 5% and three are for the 1% significance level.

Country/	**EF**		**MF**	
expert	$\hat{\alpha}_A$	$\hat{b}$	$\hat{\alpha}_A$	$\hat{b}$
I	1.134*	10.513***	0.964	3.596
II	1.659***	-4.163**	1.081	-4.502**
III	1.617***	-5.918***	1.212***	-3.046
IV	1.310***	2.651	1.036	18.650***
V	1.295***	20.427***	1.019	-4.434
VI	1.215***	3.006	1.072	-8.601
VII	1.784***	4.274**	0.990	1.874
VIII	1.772**	112.332***	0.857	79.284***
IX	1.339***	-1.835	1.222	-13.143**
X	1.089	-0.263	1.037	-1.934
XI	1.489***	-1.431	1.026	0.710
XII	1.432***	-8.009***	1.018	-5.830***
XIII	1.856***	0.641	1.284***	-1.700
XIV	1.144*	0.522	1.250***	2.049
XV	1.146	160.758***	1.092	-53.567**
XVI	1.674***	-0.699	1.058	-12.795**
XVII	1.343***	-2.810	1.207***	2.301
XVIII	2.511***	-8.023	1.219	-2.008
XIX	1.095	0.423	1.094	67.514
XX	1.396***	24.837***	0.845	-4.500
XXI	1.170	29.717***	0.839	24.509**
XXII	0.964	-3.958	0.904	-3.199
XXIII	1.540***	10.496***	0.841**	3.136*
XXIV	1.161	26.693***	0.964	18.843**
XXV	1.018	0.143	0.968	0.020
XXVI	1.337***	-1.446	0.892	-1.068
XXVII	1.088	-20.161	0.782	-1.016
XXVIII	0.846	-6.465	0.366***	-190.657
XXIX	1.810***	-4.480	1.521***	-3.126
XXX	1.925***	2.411	0.854	5.099**
XXXI	1.267***	-10.693***	1.328***	-9.900***
XXXII	1.369***	-2.620	1.031	5.679
XXXIII	1.425*	120.716	0.968	-140.564
XXXIV	1.202*	-20.129	0.948	-27.534
XXXV	1.454***	-5.086	1.288***	-3.092

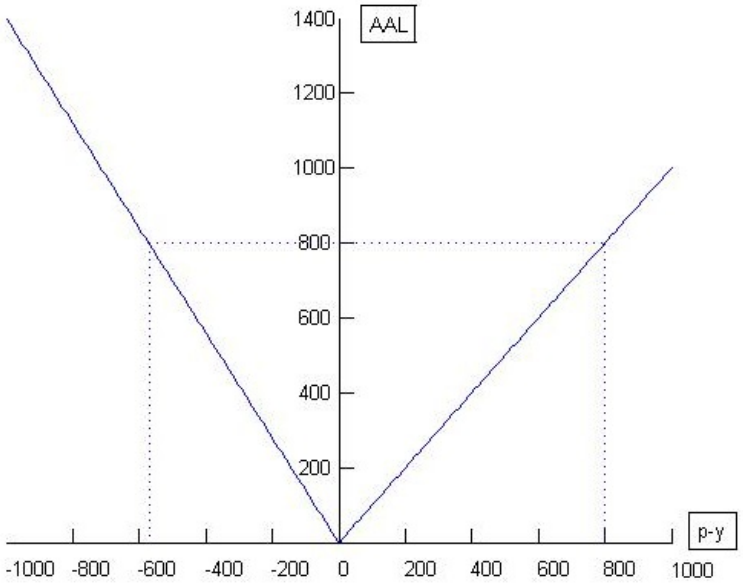

Figure 4.1: This figure shows the value of an AAL with $\alpha_A = 1.4$ for various values of the forecast error $p - y$.

The estimated systematic bias b for each expert can be found in Column 3 of Table 4.1. There are 11 experts with a significant systematic bias at the 1% significance level and another 2 experts with a significant systematic bias at the 5% significance level. Most of these are positive biases and most are linked to a significantly positive asymmetry parameter.

If we only take the 1% significance level into consideration, we can conclude that 15 experts have an asymmetric loss function, but no systematic bias. Another 6 experts have an asymmetric loss function and also a systematic bias. Only 5 experts have a systematic bias and no asymmetric loss function, and finally, only 9 experts seem to have a symmetric loss function and no systematic bias.

When we apply the test regression to the model forecasts MF, we obtain the results as reported in Columns 4 and 5 of Table 4.1. As we might expect from model forecasts based on techniques such as Box-Jenkins and Holt-Winters, we find much less evidence of asymmetry in the loss function and of systematic bias. For only 8 countries the $\hat{\alpha}_A$ are significantly different from 1 at the 1% significance level and in another 1 at the 5% significance level. The average of the 35 $\hat{\alpha}_A$ values is 1.03, which

is very close to 1. Some evidence of systematic bias is found in 12 countries, but at the 1% significance level only 4 of these cases remain. In sum, the model-based forecasts in general seem unbiased and have been created using a symmetric loss function.

Now we turn to the results when we assume that the linex loss function is used by the experts. See Table 4.2 for the estimated asymmetry parameters and systematic biases again for both EF and MF. In the second column of this table we find $\hat{\alpha}_L$ for each expert. For 18 experts we find an $\hat{\alpha}_L$ significantly different from 0 (thus asymmetry) at a significance level of 10%. For 12 of these is the difference also significant at 1%. So this is almost half of the cases where we found asymmetry for the AAL function. All except 1 (which is only significant at the 10% level) have a negative asymmetry parameter, indicating that again negative forecast errors weigh more heavily than positive forecast errors. All except 2 (which are both again only significant at the 10% level) were also found to have an asymmetric loss function under AAL. On average, $\hat{\alpha}_L$ has a value of -0.0002. See Figure 4.2 for the shape of LIN with an α_L equal to this average estimate.

However, we do find more often a significant systematic bias under the linex loss function than under the lin-lin loss function, see Column 3 of Table 4.2. 22 experts have a $\hat{b}$ significantly different from 0 at the 10% significance level and for 16 of them is this difference also significant at the 1% level. In some instances the linear asymmetry as found under AAL seems to be replaced by a (more profound) systematic bias, see for example the experts denoted with IV, XX and XXX. In general, the bias is positive again.

In sum, we find that at the 1% significance level there are far more experts with a symmetric loss function (23) than with an asymmetric loss function (12) if we assume the linex loss function. 12 of the experts with a symmetric loss function also do not have a systematic bias, although 16 experts have a systematic bias. Results are also a bit more ambiguous, because there are more countries for which significant asymmetry and/or bias is found with the 5% or 10% significance level and not with the 1% significance level, as compared to the AAL situation.

Table 4.2: The estimated asymmetry parameter α_L of LIN and estimated systematic bias b following from regression (4.16), for each expert. Columns 2 and 3 show results for expert forecasts and Columns 4 and 5 for statistical model forecasts. The asterisks in the second and fourth column indicate if the $\hat{\alpha}_L$'s are significantly different from 0 and the asterisks in the third and fifth column indicate if the $\hat{b}$'s are significantly different from 0, where one is for the 10%, two are for the 5% and three are for the 1% significance level.

Country/	**EF**		**MF**	
expert	$\hat{\alpha}_L$	$\hat{b}$	$\hat{\alpha}_L$	$\hat{b}$
I	4.4e-05	15.009***	7.1e-05*	2.585
II	-3.7e-04***	2.528	-1.2e-04***	-3.548*
III	-2.1e-04***	3.775**	-8.2e-05***	0.852
IV	1.1e-05	11.290***	-1.7e-05	19.688***
V	-9.2e-06	25.171***	5.4e-06	-4.066
VI	-6.4e-05***	13.438***	-2.1e-05	-4.848
VII	-5.8e-04***	13.489***	-2.2e-04	1.522
VIII	-4.2e-04**	136.130***	-2.3e-05	72.009***
IX	-2.3e-04	10.143**	-2.0e-04	-5.096
X	8.7e-05	1.611	2.1e-04	-0.888
XI	-6.2e-04***	2.742**	-7.8e-05	0.954
XII	-6.4e-04***	-4.652***	-1.3e-04	-5.689***
XIII	-6.3e-04***	8.567***	-4.3e-04**	1.414
XIV	-6.2e-06	1.560	-3.1e-05*	3.755***
XV	-8.5e-04*	160.967***	-1.8e-04	-48.658***
XVI	-1.5e-03***	9.614**	-1.1e-04	-11.557**
XVII	-8.4e-05***	2.979	-3.9e-05**	6.029**
XVIII	-3.0e-04***	15.927***	1.4e-04	3.585
XIX	-4.4e-05	18.512	-5.5e-06	105.082
XX	-7.6e-05	33.160***	3.0e-04**	-8.123**
XXI	-4.0e-05	36.277***	-1.4e-06	16.901**
XXII	-7.2e-05	-4.624	4.6e-04	-4.377
XXIII	-8.9e-05	15.560***	8.4e-05	1.126
XXIV	-1.9e-05	35.699***	1.5e-05	16.767**
XXV	2.9e-05	1.036	2.6e-05	-1.126
XXVI	1.1e-05	-0.221	4.2e-05	-1.543
XXVII	1.8e-05	-10.861	3.3e-04	-21.134
XXVIII	1.9e-05*	-77.468	5.3e-05***	-652.308***
XXIX	-9.7e-06**	6.806	2.1e-06	4.913
XXX	1.2e-05	7.557***	4.7e-06	3.859**
XXXI	-4.1e-05*	-6.420**	-6.3e-05**	-4.799*
XXXII	-1.3e-04**	20.429**	2.7e-05	8.526
XXXIII	-2.5e-06	221.010***	-7.2e-07	-150.069**
XXXIV	-4.5e-05***	-5.803	-2.7e-05	-33.433*
XXXV	-3.0e-04***	3.617	-1.4e-04*	2.981

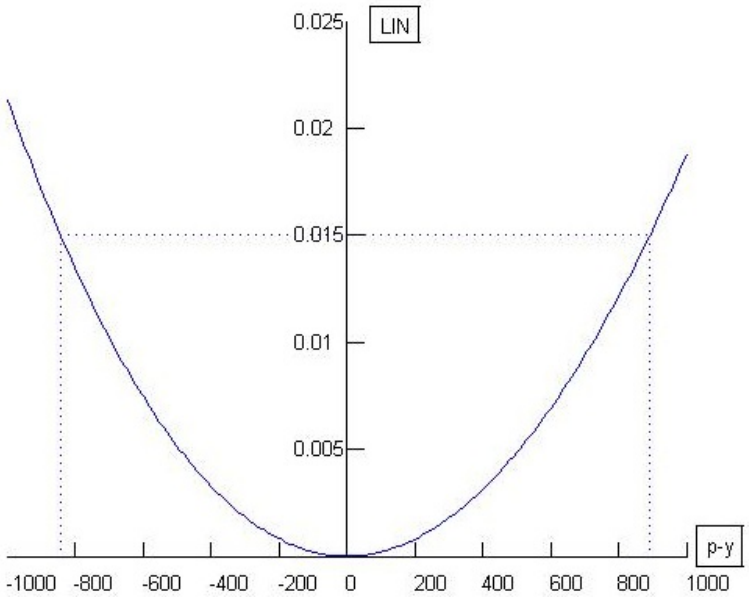

Figure 4.2: This figure shows the value of a LIN with $\alpha_L = -0.0002$ for various values of the forecast error $p - y$.

Finally, we also compare these linex results for EF with the linex results for MF, see Columns 4 and 5 of Table 4.2. Again we do not find much evidence for asymmetry and systematic bias in the model forecasts. $\hat{\alpha}_L$ is on average -4.13e-06, so much closer to 0 than the average $\hat{\alpha}_L$ of -0.0002 found for EF. For only 10 countries is the asymmetry parameter significantly different from 0 at the 10% level and in only 3 countries at the 1% level. The number of significant systematic biases is 16 at the 10% level and 6 at the 1% level. So again these results confirm that statistical model forecasts are unbiased and derived from a symmetric loss function.

4.3.3 Specification checks

So far, we have analyzed the results given the assumptions underlying the analysis. To test these assumptions we now follow the strategy as outlined in Section 4.2.4.

The first step is to check if the error terms of the regression models (4.15) for AAL and (4.16) for LIN are standard normally distributed. To that end, we use the Kolmogorov-Smirnov test, see D'Agostino and Stephens (1986). The test is performed

on the error terms of each country separately, so we have 35 test results. In the second and third column of Table 4.3 we see how often these 35 tests reject the null hypothesis of standard normally distributed error terms at the 1% significance level. For the asymmetric absolute loss function we see fairly low figures. For EF we see that in a little bit over one-third of the tests the null hypothesis is rejected and for MF this is a little bit over one-fifth. Note that the number of observations on which the normality test is performed is quite large (see Subsection 4.3.1), which means that the test has high power against tiny deviations of standard normality. For countries with many observations we may therefore choose an even lower significance level which implies that the number of rejections may even be lower.

Table 4.3: This table shows the number of times out of 35 that the hypothesis that the error terms of the regressions (4.15) (Column 2), (4.16) (Column 3) and (4.19) (Column 4) are standard normally distributed is rejected. We use the Kolmogorov-Smirnov test with a significance level of 1%.

	AAL	**LIN**	**AAL log**
EF	13	23	35
MF	8	12	35

For the linex loss function we find much higher numbers of rejection, namely 23 (66%) for EF and 12 (34%) for MF. As the numbers for AAL are much lower, this might indicate that we should not reject the assumption of a normal forecast distribution at this point, but that the assumption of a linex loss function is perhaps not an appropriate assumption. The AAL function seems to be the loss function that is more likely to be used by the managers creating the forecasts in this data set.

As we deal with sales forecasts in this application, which are always positive, it might be resonable to assume that the forecasts are lognormally distributed instead of normally. Therefore, we also estimate (4.19), again with separate coefficients for each country, and again we test if the error terms are standard normally distributed. We find overwhelming evidence that the forecast distribution is not lognormal, see Column 4

of Table 4.3. Both for EF and MF the null hypothesis of standard normal error terms is rejected for all 35 countries. This again indicates that assuming a normal forecast distribution seems acceptable for our data.

Our final specification check involves a comparison of our AAL results with those upon using the method of Elliott et al. (2005). Table 4.4 and Figures 4.3 and 4.4 give the results. Note that Columns 2 and 4 of Table 4.4 are the same as Columns 2 and 4 in Table 4.1, but are repeated for ease of comparison. Columns 3 and 5 present the results as obtained using the method of Elliott et al. (2005), where we used as instrumental variables a constant and one-month lagged sales. Remember that we expect $\hat{\alpha}_A$ and $\hat{\alpha_E}/(1-\hat{\alpha_E})$ to be approximately the same if the assumptions for our method are correct.

First note, from Table 4.4, that whenever $\hat{\alpha_A}$ is significantly larger than 1 at each significance level, $\hat{\alpha_E}/(1-\hat{\alpha_E})$ is never significantly smaller than 1 at each significance level. Furthermore, whenever $\hat{\alpha_A}$ is significantly smaller than 1 at the 1%, 5% or 10% significance level (happens only twice for MF), $\hat{\alpha_E}/(1-\hat{\alpha_E})$ is never significantly larger than 1 at the 1%, 5% or 10% significance level. Both statements also hold true when $\hat{\alpha_E}/(1-\hat{\alpha_E})$ is evaluated against $\hat{\alpha_A}$. These results indicate that we never find fully conflicting results with the two alternative methods.

The largest difference in results appears when we find a significant asymmetry with one method and no significant asymmetry with the other method. If we focus on the 1% significance level, this happens 8 times for EF and 2 times for MF, but in most of these cases (7) the other method also shows asymmetry at the 5% or 10% level. Hence, we find that both methods may differ in terms of the amount of asymmetry, but not in the sign of the asymmetry and hardly in the existence of the asymmetry.

To get a more precise idea of the size of the differences in estimated asymmetry parameters, we can take a look at the histograms in Figures 4.3 and 4.4. Here the differences between $\hat{\alpha}_A$ and $\hat{\alpha_E}/(1-\hat{\alpha_E})$ are depicted, for EF in the first figure and for MF in the second. Multiplying the differences with 100% shows the differences in percentages. Thus for example, a value of 0.1 indicates that the difference in weight

Table 4.4: The estimated α_A of AAL and estimated $\alpha_E/(1-\alpha_E)$ of general loss function (4.20) with $q = 1$ using the estimation method of Elliott et al. (2005), for each country. Columns 2 and 3 show results for expert forecasts and Columns 4 and 5 for statistical model forecasts. The asterisks in the second and fourth column indicate if the $\hat{\alpha}_A$'s are significantly different from 1 and the asterisks in the third and fifth column indicate if the $\hat{\alpha}_E$'s are significantly different from 0.5, where one is for the 10%, two are for the 5% and three are for the 1% significance level.

Country/ expert	**EF**		**MF**	
	$\hat{\alpha}_A$	$\hat{\alpha}_E/(1-\hat{\alpha}_E)$	$\hat{\alpha}_A$	$\hat{\alpha}_E/(1-\hat{\alpha}_E)$
I	1.134*	1.321***	0.964	1.091
II	1.659***	1.387***	1.081	1.001
III	1.617***	1.579***	1.212***	1.170***
IV	1.310***	1.170***	1.036	1.148**
V	1.295***	1.468***	1.019	0.851**
VI	1.215***	1.291***	1.072	1.085
VII	1.784***	1.618***	0.990	1.036
VIII	1.772**	2.253***	0.857	1.365
IX	1.339***	1.505***	1.222	1.057
X	1.089	1.045	1.037	0.888
XI	1.489***	1.506***	1.026	1.152**
XII	1.432***	1.180**	1.018	0.918
XIII	1.856***	1.778***	1.284***	1.078
XIV	1.144*	1.171**	1.250***	1.211***
XV	1.146	2.118***	1.092	0.780*
XVI	1.674***	1.464***	1.058	0.924
XVII	1.343***	1.455***	1.207***	1.360***
XVIII	2.511***	1.655***	1.219	0.929*
XIX	1.095	1.090	1.094	1.423*
XX	1.396***	1.296**	0.845	0.821**
XXI	1.170	1.802***	0.839	1.268
XXII	0.964	0.970	0.904	0.990
XXIII	1.540***	1.846***	0.841**	1.043
XXIV	1.161	1.442***	0.964	1.127
XXV	1.018	1.019	0.968	1.019
XXVI	1.337***	1.415***	0.892	1.105
XXVII	1.088	1.027	0.782	0.798**
XXVIII	0.846	1.136	0.366***	0.598***
XXIX	1.810***	1.651***	1.521***	1.007
XXX	1.925***	1.532***	0.854	1.087
XXXI	1.267***	1.586***	1.328***	1.421***
XXXII	1.369***	1.476***	1.031	1.124
XXXIII	1.425*	2.228***	0.968	0.916
XXXIV	1.202*	1.084	0.948	0.864
XXXV	1.454***	1.397***	1.288***	1.290***

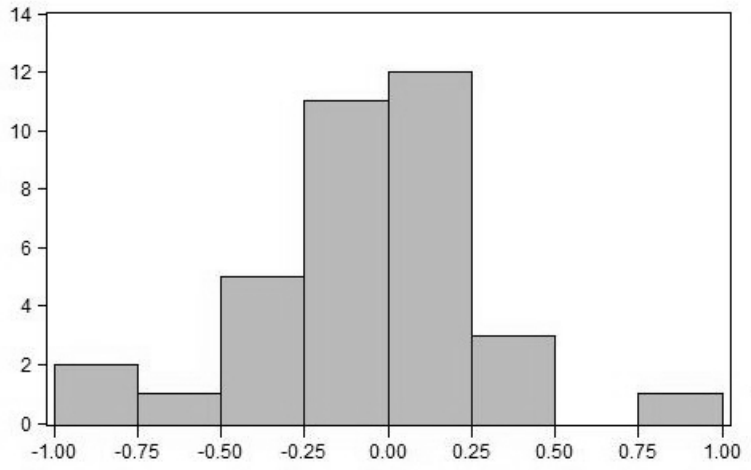

Figure 4.3: Histogram of differences between $\hat{\alpha}_A$ and $\hat{\alpha_E}/(1-\hat{\alpha_E})$ for 35 experts estimated over EF.

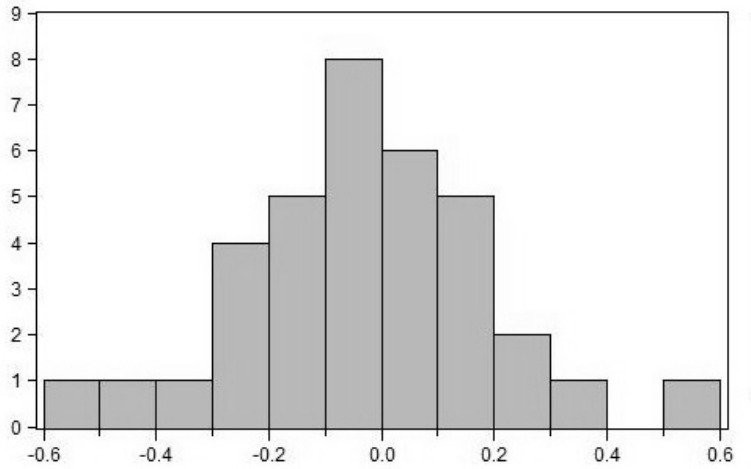

Figure 4.4: Histogram of differences between $\hat{\alpha}_A$ and $\hat{\alpha_E}/(1-\hat{\alpha_E})$ for 35 countries estimated over MF.

between too low forecasts and too high forecasts is 10% higher according to $\hat{\alpha}_A$ than according to $\hat{\alpha}_E$.

Although we see some outliers in the graphs, the largest one being that $\hat{\alpha_E}$ is 97% larger than $1-\hat{\alpha_E}$ than that $\hat{\alpha}_A$ is larger than 1, on average the difference is around to be equal to 6% (0.06 in the figure). Furthermore, in 23 of the 35 countries the difference is smaller than 25% in absolute sense and in 31 of the 35 countries the difference is

smaller than 50%. For MF, see Figure 4.4, these differences are even smaller, with an average difference of around 2.5% and a maximum difference of around 51%, both in absolute terms.

The larger differences do not necessarily seem to be related with the rejection of the standard normality of the residuals of the regression. The correlation between the absolute difference and the p-value of the Kolmogorov-Smirnov test is -0.14 for EF and -0.06 for MF. If we look at the test results for some countries with large differences in estimation results, we sometimes find rejection of the null hypothesis and sometimes we do not.

To conclude, we do not find large differences in the results of both methods and we take this as a final indication that the assumptions underlying our analysis do not need to be rejected, at least, for our data set at hand.

4.4 Conclusions

There is much available research on asymmetric loss functions for forecasters, but most of it is focussed on the theoretical discussion of possible shapes of those loss functions and on resulting optimal forecasts. Very little is known about which loss function is actually exercised by experts when they create their forecasts and rarely it is quantified to what extent the loss functions are asymmetric. We are aware of one study only, and this is presented in Elliott et al. (2005).

In the present paper we propose a new and simple methodology to deduce the asymmetry parameter of the asymmetric absolute loss function and of the linex loss function. The derivation is based on some simplifying assumptions which can be held against actual data in a number of ways. The derivations were shown to lead to simple linear regressions.

We applied our methodology to a large data set of SKU-level sales forecasts where model forecasts are received by experts, after which they provided their final forecasts. We documented substantial evidence that the experts use an asymmetric loss function,

where we diagnosed that most likely it is the asymmetric absolute loss function. Forecasts that are too low have a weight in the loss function that is on average 40% higher than forecasts that are too high.

The methodology proposed in this paper results in similar results as found with the methodology of Elliott et al. (2005) and in general we find no obvious indications that the assumptions underlying our analysis should be rejected. To what extent this is true for other data sets remains to be analyzed.

Further research on loss functions of experts could focus on multi-step-ahead forecasts. As forecasts errors might be correlated in such situations, the methodology might be a bit more complicated than the one presented here. Finally, forecast updates, that is, sequential forecasts for the same event, are also interesting to analyze.

Chapter 5

Does Disagreement amongst Forecasters have Predictive Value?

Joint work with Philip Hans Franses.

5.1 Introduction

There are many situations in macroeconomics in which there is not just one single expert forecast available. Indeed, quite often, forecasts from surveys are available, consisting of forecasts from various experts who make predictions for the same variable. A well-known example is the Survey of Professional Forecasters (SPF) conducted by the Federal Reserve Bank of Philadelphia. Most studies that analyze SPF-type data focus on the predictive value of the mean or median of these SPF forecasts. Theoretical and empirical research has shown the relevance of these two statistics (see e.g. Einhorn and Hogharth, 1975; Clemen, 1989; Armstrong, 2001a). Often these mean or median values are combined with forecasts from time series models, see for example recent work from Elliott and Timmermann (2005).

Recently, a growing literature discusses other features of experts-based survey forecasts. For example, disagreement amongst experts is described (Dovern et al., 2009),

an explanation is sought for this disagreement (Capistrán and Timmermann, 2009; Mankiw et al., 2003), and its effects on decision makers are investigated (Baillon and Cabantous, 2009). Besides disagreement amongst experts, various other features are described and evaluated. Survey forecasts typically involve much dependence between forecasts from the same expert (Cooke, 1991). There is also positive serial correlation in the forecast errors (Mankiw et al., 2003; Capistrán and Timmermann, 2009). Expert opinions are biased (Laster et al., 1999) and characteristics of the forecasters as age and experience are related to forecast performance (Lamont, 2002).

Interestingly, although much information on features of survey forecasts is available, there is no research on the predictive power of these features. In this paper, we therefore look at this predictive power where we focus on the disagreement amongst the forecasters. Our conjecture, which we outline in more detail below, is that the degree of disagreement could signal upcoming structural and temporal changes in an economic process and in the predictive power of the survey forecasts.

In our empirical work, we examine a variety of US-specific macroeconomic variables, and we consider different ways to measure the degree of disagreement. The models include measure for the degree of disagreement and also measures for location of the survey data and autoregressive components. Forecasts from simple linear models and forecasts from Markov regime-switching models with constant and with time-varying transition probabilities are constructed in real-time and compared on forecast accuracy. The survey forecasts are from the Survey of Professional Forecasters. Our main finding is that disagreement can indeed have predictive value.

The remainder of the paper is structured as follows. Section 2 discusses in more detail the literature and it explains which variables and models are used. Section 3 describes the data, the models and the methods to evaluate the forecasts. Section 4 gives a summary of the results and Section 5 concludes.

5.2 Background

In this section, we first discuss the features of expert forecasts that could be relevant for forecasting, and next, we explain how such features can appear in forecasting models.

5.2.1 Disagreement

Without doubting the usefulness of the mean and median of survey forecasts, there are many reasons to also look at other statistics of survey forecasts. Researchers have been puzzled by some (seemingly) irrational characteristics of expert forecasts and explanations have been sought in multiple directions. For example, most theoretical macroeconomic models do not endogenously generate disagreement, while disagreement is prevalent in every survey of (professional) forecasters (Dovern et al., 2009; Laster et al., 1999; Capistrán and Timmermann, 2009). Other characteristics unexplained by full rationality are autocorrelated forecast errors and insufficient sensitivity to recent macroeconomic news (Capistrán and Timmermann, 2009; Mankiw et al., 2003). Note that explanations given in these studies for these features might also explain why these features could have predictive power, and that is what we will consider next.

The new statistics that we consider here are the standard deviation, the 5th percentile and the 95th percentile of the survey forecasts and the number of forecasts collected. The 5th percentile and the 95th percentile of survey forecasts can be seen as measures of disagreement amongst forecasters when used in combination with the mean or median of the survey forecasts. We begin by explaining why we address the first three of these statistics, which can be seen as direct measures of disagreement amongst forecasters, and we return to the number of forecasts later on.

Laster et al. (1999) describe why it is questionable that professional forecasts are rational in the sense of being efficient and unbiased. They construct a model in which a forecaster is driven by two conflicting incentives, and these are (1) to forecast as accurate as possible and (2) to generate publicity for their firms. Most ideally, a prediction

is accurate and all other predictions are very inaccurate. Indeed, being accurate while all others are too does not generate much publicity. At the same time, being wrong once in a while and being the only one close to the true value at other times, might be better for a firm than always following consensus. Therefore, professional forecasters may behave as strategically rational.

Related to this model is the work of Lamont (2002). He finds that the age and experience of forecasters are related to forecast accuracy. The older and more established forecasters, the more radical are their forecasts and the more inaccurate. This can again be explained by reputational factors.

Following this line of thought, it might be beneficial for a firm or an individual expert to give extreme forecasts at times in which change may come up, even if they are not sure what kind of change it will be or how it will look like. Some forecasters, who may be more dependent on publicity, might react more extreme to certain information than other forecasters would do. In these periods some forecasters might take an additional risk, because correctly forecasting extreme future observations generates positive publicity, while being wrong about it is not that bad.

In our empirical work below, we rely on survey forecasts from anonymous sources. It is possible to follow one and the same forecaster by way of a code and it is known what kind of firm provides the forecast, but the names of the forecasters and the firms are unknown. The question is now, to what extent the forecasters behave strategically as described above. Without having any information about this, it is very well possible that the forecasts provided in the anonymous survey are used in other contexts too, where reputational factors do play a role. Furthermore, within the firms the personal reputation of a forecaster might also be important and factors described by Lamont (2002) might still influence the practice of forecasting.

Forecasters who react in different ways to specific information also fit the arguments in Capistrán and Timmermann (2009). The forecasters are presumed to have asymmetric loss functions. There is heterogeneity in agents' loss functions. And, a constant loss component can explain how dispersion in inflation beliefs evolves over

time and why it is correlated with the level and volatility of inflation. Without discussing their model in detail, it is intuitively clear that if forecasters have asymmetric and differing loss functions (possibly also convex), they update their forecasts in very different ways and particular information might cause dispersion in forecasts.

Mankiw et al. (2003) propose a model that can reproduce the distribution of inflation expectations. They use a sticky-information model in which agents only update their forecasts periodically because of costs involved in gathering information and adjusting projections. Each period, only a fraction of the forecasters gets new information and processes it. Mankiw et al. (2003) find that this model is capable of matching the disagreement and its evolution in inflation expectations.[1]

If we were to consider their model to represent how expert opinions are adjusted, then disagreement in expert expectations might very well signal upcoming changes in an economic process. If only some of the experts receive information about this and process it, then the standard deviation, the 95th percentile or the 5th percentile might contain this information, while the mean and the median might do so only partially or perhaps not at all.

A final argument why we focus on the predictive content of disagreement in forecasts originates from the work of Zarnowitz and Lambros (1987) and Lahiri and Sheng (2010), amongst others. Zarnowitz and Lambros (1987) show, also by using the SPF data, that measures of consensus, measured by the degree of agreement amongst point predictions, and uncertainty, which is a lack of confidence, are positively correlated. Lahiri and Sheng (2010) also argue that disagreement amongst experts, measured as the standard deviation of expectations, is a good proxy for forecast uncertainty, assuming that the variability of future aggregate shocks is stable.

When disagreement in expectations increases while the variability of future aggregate shocks does so too or stays the same, the forecast uncertainty of expert opinions

[1]Carroll (2003) uses a very similar model to explain the evolution of variation in inflation expectations, namely an epidemiological model in which information goes from one person to another in the same manner as diseases go from one person to another.

also increases and the predictive power of the survey forecasts gets reduced. It might be better in such a situation to rely more on statistical model forecasts and less on expert opinion. However, if the level of disagreement and the variability of future aggregate shocks move in opposite directions, it is unclear what to do.

Hence, the balance between fully relying on model forecasts and also relying on experts depends on the situation at hand.

Finally, the fourth explanatory variable that we consider, which is the number of forecasts collected, is a bit different from the previous three. This variable is related to disagreement, because with more forecasts there is more room for differing statements. As more forecasts do not necessarily result in a larger variation in forecasts, we discuss this variable separately.

Firms might go bankrupt and therefore they no longer provide forecasts. It is also known that participation in the survey might depend on efforts of the organization collecting the data.[2] But firms may also strategically decide to stop or to start forecasting. If they are unsure about the future they may decide not to provide a forecast, so they cannot be wrong. Or, they provide forecasts if they have information they believe is exclusive. So, the number of forecasts might itself be informative about what might happen next.

Another reason why this might be the case, probably more convincing in case of anonymous data, is that the economic situation can influence the necessity for and the ability of firms to forecast. If firms are in trouble because of the economic situation, research departments may be closed. On the other hand, economic volatile times can increase the need to get insight into what the future will bring and thus the need to create forecasts. So, although it is not clear how the number of forecasts and the predictability of macroeconomic variables are linked exactly, it is clear that there are

[2]In case of the Survey of Professional Forecasters: initially the survey was conducted by the American Statistical Association (ASA), together with the National Bureau of Economic Research (NBER), but was taken over by the Federal Reserve Bank of Philadelphia in 1990. They revived the survey by inviting new forecasters to the survey as participation rates had dropped in the years before.

good reasons why they might be linked.

Finally, the number of forecasts might indicate how reliable the mean or median of those forecasts is. One can imagine that if the number of forecasts is low that extreme erroneous forecasts are not cancelled out and that the weight put on it in models and forecast combinations should better be small in that case.

5.2.2 Including disagreement in forecasting models

We will use the above mentioned four explanatory variables in three different forecasting models. The first is a simple ARX model in which autoregressive terms and each time one of the new statistics of the survey data is included. The second is a Markov regime-switching model with constant transition probabilities (MRScons) where one of the explanatory variables each time is one of the new survey statistics. The third is a Markov regime-switching model with time-varying transition probabilities (MRSvar). In this model, one of the four new statistics is used to predict if a regime switch will happen. The models are explained in more detail in the next section, but here we will give some arguments why we use these MRS models.

First of all, Markov regime-switching models are an often used and popular tool to describe and forecast macroeconomic variables, ever since the publication of the influential paper of Hamilton (1989). In Granger (2001) some of the many applications are discussed, showing its suitability in a variety of forms for macroeconomic variables such as output growth, inflation and interest rates. Timmermann (2000) showed the flexibility of these models by deriving the moments for a range of Markov switching models. MRS models showed to be capable of accounting for specific features of macroeconomic time series such as volatility clustering, asymmetry, and fat-tail behavior.

Second, Elliott and Timmermann (2005) showed that Markov regime-switching models might be useful when combining forecasts. They indicate that the weights of the combination of AR model forecasts and expert forecasts could be driven by a

Markov regime-switching process. For three of their six macroeconomic series, that is, unemployment rate, inflation and nominal GDP growth, this combination method performs better than a range of alternative combination methods.

A third and final reason to use MRS models follows from our arguments above why we focus on statistics of disagreement as explanatory variable. This variable might signal upcoming structural and temporal changes in an economic process and in the predictive power of the survey forecasts. A good way to model structural changes is by incorporating these variables as explanatory variables for the transition probabilities in a MRSvar model.

Our models include lags as explanatory variables, because AR models and AR-MRS models have a proven track record. Further, as many studies have shown that combinations of expert forecasts with statistical model forecasts outperform both individual forecasts, we include both components. That is, we include the mean or median of the survey forecasts in some of the models. Most studies that analyze survey type forecasting data restrictively focus on the predictive value of the mean or median. Clemen (1989) found in a broad study on forecast combinations that simple arithmetic averages of forecasts are accurate for many types of forecasts. Einhorn and Hogharth (1975) argue that equal weights produce precise forecasts because there is no estimation error, because no degrees of freedom are lost, and because no mistakes can be made with the 'true' relative weights, potentially giving the wrong forecast the largest weight. According to Armstrong (2001a), who refers to multiple other studies, there is evidence that the median is even more accurate. As the mean is an often used and praised statistic (Zarnowitz and Braun, 1993; Laster et al., 1999; Elliott and Timmermann, 2005), we will incorporate both variables in our empirical study.

5.3 Methodology

In this section we will discuss the data, the models and the evaluation methods in more detail.

5.3.1 Data

The survey forecasts we use are from The Survey of Professional Forecasters (SPF), which is a quarterly survey of macroeconomic forecasts in the United States of America. The survey began in the fourth quarter of 1968 and was conducted by the American Statistical Association and the National Bureau of Economic Research, so it is often called the NBER-ASA survey. The Federal Reserve Bank of Philadelphia took over the survey in the second quarter of 1990.

The respondents to the survey are forecasting professionals. Participants include, amongst others, financial firms, banks, consultancy firms and university research centers. For each variable, 'forecasts' are given for the previous quarter (for which the first release is already available), for the current quarter (for which no realized data is available yet) and for the four following quarters. We consider one-quarter-ahead predictions, where one quarter ahead is the first quarter for which no data is available at the moment of estimating the model parameters and creating forecasts. So, we use the survey forecasts given for the current quarter.

At present, the survey encompasses 31 macroeconomic variables. We use five of these variables in our analysis[3]: the index of industrial production (INPROD), Nominal Gross Domestic Product (GNP before 1992) (NGDP), inflation as measured through the GDP chain-weighted price index (PGDP), the unemployment rate (UNEMP) and private housing starts (HOUSING). Our main focus is on INPROD, for which we will give the most detailed results and we compare the results for INPROD with the results for the other variables. For INPROD and HOUSING, realized monthly figures are available, so averages are taken to obtain quarterly figures. Except for the unemployment rate, we look at growth rates, measured as the first differences in natural logs of the current value and the previous quarter's value.

The SPF forecasts, used for the explanatory variables in the forecasting models, are

[3]Except for corporate profits, we use the same variables as Elliott and Timmermann (2005) do in their study.

also transformed into growth rates. We include the rate of change in the forecast for the current quarter over the 'forecast' of that same forecaster for the previous quarter. So, growth is measured as $100 * (\ln(ySPF_{i,t}) - \ln(ySPF_{i,t-1}))$, where $ySPF_{i,t}$ is the forecast of forecaster i for variable y for the current quarter and $ySPF_{i,t-1}$ is the 'forecast' of that same forecaster for the previous quarter. We use the forecasters' own stated value for the previous quarter instead of the first release, because we think it gives a better representation of the forecasted growth. Most of the time, these two figures are the same and sometimes they differ because of mistakes in for example the base year used to construct some of the variables. These mistakes cancel out by using forecasts instead of released values.

At the start of our analysis, SPF forecasts were available for the last quarter of 1968 to and including the third quarter of 2009. Realized data were also available up until the third quarter of 2009. So we work with $n = 164$ data points. Each quarter there are f_t forecasts available from the SPF for the current quarter and the previous quarter.

5.3.2 The general model

The Markov regime-switching model with time-varying transition probabilities nests all models we consider. This MRSvar for the variable to be explained y_t, with m regimes and p lags, is

$$y_t = \alpha_{s_t} + \phi'_{s_t}(L)y_{t-1} + \beta'_{s_t}X_t + \varepsilon_{t,s_t}, \tag{5.1}$$

or stated differently:

$$y_t = \begin{cases} \alpha_1 + \phi_1(1)y_{t-1} + ... + \phi_1(p)y_{t-p} + \beta'_1X_t + \varepsilon_{t,1} & \text{if in state 1} \\ \alpha_2 + \phi_2(1)y_{t-1} + ... + \phi_2(p)y_{t-p} + \beta'_2X_t + \varepsilon_{t,2} & \text{if in state 2} \\ \vdots \quad \vdots \quad \vdots & \vdots \\ \alpha_m + \phi_m(1)y_{t-1} + ... + \phi_m(p)y_{t-p} + \beta'_mX_t + \varepsilon_{t,m} & \text{if in state m,} \end{cases} \tag{5.2}$$

with $\phi_{s_t}(L)$ a polynomial lag of order p, $s_t \in [1, m]$ the regime state in period t, X_t is a vector of k explanatory variables and $\varepsilon_{t,s_t} \sim N(0, \omega_{s_t})$. The model can assume the

ω_{s_t} to vary per regime, or ω_{s_t} to be constant across the different regimes. The variable s_t is unobserved and it is assumed to develop according to a first-order Markov chain with transition probabilities

$$p_{ijt} = Pr[s_{t+1} = j | s_t = i, Z_t] = \frac{\exp(\delta_{ij} + \gamma'_{ij} Z_t)}{\sum_{j=1}^{m} \exp(\delta_{ij} + \gamma'_{ij} Z_t)}, \tag{5.3}$$

with Z_t a vector of r explanatory variables for the regime switching and δ_{i1} and γ_{i1} are set to zero for identification purposes.

If Z_t in (5.3) is empty (or if all γ_{ij} are set to zero), the model reduces to a Markov regime-switching model with constant transition probabilities (MRScons). With all δ_{ij} and γ_{ij} set to zero the regimes have an equal probability of occurring.

If $m = 1$ the resulting model is a linear model with p lags and additional explanatory variables X (ARX). If $m = 1$ and X is empty (or if β is set to zero) we get a simple AR model.

5.3.3 Models considered

X_t and Z_t are vectors of zero or of one or more variables related to the SPF forecasts (transformed into growth rates). The variables used for Z_t are standardized to facilitate the estimation process. The variables that we consider for X_t and Z_t are divided into two sets. The first set consists of two variables, that is the mean and median of the forecasts, and therefore these are called the location variables. The second set consists of the standard deviation (std), the 0.05 quantile (5p) and the 0.95 quantile (95p) of the forecasts and the number of forecasts f_t (nr), and each of these four is a different measure for the degree of disagreement amongst forecasters.

For each of the five macroeconomic variables four groups of models are put forward. These are AR models, ARX models, MRScons models and MRSvar models. The first group consists of linear models with lags of the dependent variable as the explanatory variables. Models with zero, one, two, three and four lags are considered, so in total this group consists of five different models.

The second group contains the same linear models as the first group, only now with one or two additional explanatory variables. So, this group consists of models with zero, one or more lags and one of the six explanatory variables described above and it consists of models with zero, one or more lags and two explanatory variables, one being a location variable and one being a variable measuring disagreement. In total, this group encompasses 70 models.

The third group is the group with MRScons models, consisting of 150 models. For each of the models in the first two groups two models are estimated in this group, each with two regimes ($m = 2$) and one with common variance and one with varying variance per regime.

The final and fourth group contains 24 MRSvar models. These models are estimated with m set to two, with one lag, X_t containing zero or one of the variables of the first group of explanatory variables (that is, the mean or the median of the SPF forecasts), Z_t containing one of the variables of the second group of explanatory variables and with a common or varying variance per regime.

All model parameters are estimated 40 times. The first time with 124 data points, leaving out the last 40 observations, the second time again with 124 data points, leaving out the first observation and the last 39 observations and so on. Stated differently, parameter estimation adopts a rolling window of data. Each model estimated with data up till date t will be estimated with the vintage of date t (that is, the last data release available at date t). Only for INPROD, which is the focal variable, the model parameters are also estimated 60 times, using a rolling estimation window of 104 data points and leaving out 60 observations.

Estimation of the parameters proceeds by optimizing the likelihood function associated with the Markov regime-switching model or with the linear model. As the underlying state variable s_t is assumed to be unobserved, it is treated as a latent variable and the EM algorithm described in Hamilton (1994) is used for the estimation of the MRScons models. The EM algorithm developed by Diebold et al. (1994) is used for estimation of the MRSvar models.

For the MRS models, the estimation results might depend on the starting values of the parameters used in the estimation procedure, as the likelihood function has multiple local optima. Therefore, we use a grid of different starting values for δ_{ij} and γ_{ij} the first time that the model parameters are estimated and every fifth time after that and we select the model with the maximum log-likelihood. For the remaining estimation rounds, the estimated parameters of the previous step are used as starting values.

5.3.4 Forecast evaluation

For each model and each macroeconomic variable a set of one-step-ahead forecasts is created. The forecasts are created making use of the parameters estimated with the most recent available data if the forecasts were created in real-time while making use of a rolling estimation window. As we use 124 data points to estimate the model parameters, we obtain $P = 40$ forecasts to evaluate.

For INPROD we also look at two other sets of forecasts per model. The first set is created with the models estimated in the first round with the first 124 data points, so with a fixed estimation window. We also obtain $P = 40$ forecasts in this case. The second set is created making use of a rolling estimation window again, but now the models are estimated over 104 data points, so here we obtain $P = 60$ forecasts.

Forecasts for the MRS models are constructed as in Hamilton (1994). This means that with two regimes the one-step-ahead forecast is

$$\hat{y}_{t+1|t} = E[y_{t+1}|s_{t+1} = 1, \Omega_t] \cdot P(s_{t+1} = 1|\Omega_t; \theta) + E[y_{t+1}|s_{t+1} = 2, \Omega_t] \cdot P(s_{t+1} = 2|\Omega_t; \theta), \quad (5.4)$$

where θ denotes the estimated parameters, Ω_t is all the data available up to date t and $P(s_{t+1} = j|\Omega_t; \theta)$ are the one-step-ahead state probabilities computed from the filtered state probabilities $P(s_t = j|\Omega_t; \theta)$ (which are obtained from the estimation procedure) and multiplied by the transition probabilities in (5.3). For the fixed estimation window these one-step-ahead state probabilities obtained from the first estimation round are

multiplied by the transition probabilities in (5.3) to obtain forecasts in the second round and so on.

One way to evaluate the forecasts is to select in each step the number of lags for the models in the first three groups using an information criterion. In the same way, a selection between constant and varying variance in the MRS models could be made. However, we decide to only analyze and compare the forecasts of the models estimated with one lag and to look at the models with varying and constant variance without selecting one of these with an information criterion. This decision is based on our finding that results do not necessarily improve by working with information criteria to select the number of lags and to make a selection between constant and varying variance. Furthermore, it is not that clear which information criterion to use.[4] This way, we focus completely on the predictive value of the disagreement variables without the results being flawed because of a possible inappropriate use of information criteria.

The forecasts are analyzed using two kinds of data realizations. These are the first release and last release data, as they are known in the third quarter of 2009. Root mean squared prediction errors (RMSPE) are constructed and the RMSPE's of our models (including one of the four SPF variables introduced in this study) are compared with 9 benchmarks. As benchmark forecasts we use the mean and median of SPF forecasts and forecasts from 7 different benchmark models, which are the AR model with one lag, the ARX model with one lag and with the mean or median of SPF forecasts and the MRS model with one lag and with mean or median of SPF forecasts and with varying or constant variance of the error terms (as inspired by Elliott and Timmermann (2005)). In the literature, these 9 benchmarks are approved for their accuracy, their simplicity or both and especially the first 5 are often used. To prevent drawing the wrong conclusions, because the wrong benchmark is used and in order to give a com-

[4]See for example Psaradakis and Spagnolo (2006), Smith et al. (2006) and Awirothananon and Cheung (2009). They investigate which information criterion to use to choose between different MRS models, but also focus on the decision on the number of regimes and for example not on the choice to use common or varying variance. Furthermore, their results are conflicting.

prehensive view of how our proposed models perform compared to other models, we use these 9 benchmarks.

We use two different tests to see if the differences in RMSPE's are significant. The first is the well-known test of Diebold and Mariano (1995) (DM). McCracken (2000) and Clark and McCracken (2001) have shown that this test is not valid if the models are nested models, because the asymptotic distribution of the test statistic is not standard in this case. However, Giacomini and White (2006) showed that the DM-test remains valid for nested models when the estimation sample size remains finite, or stated differently, when a rolling or fixed estimation sample forecasting scheme is used. Thus, although we work with nested models, by using rolling and fixed estimation sample forecasting schemes it is possible to use the DM-test in a standard way.

It is found that the DM-statistic tends to be over-sized in small samples. As we have a rather small sample of forecasts we adjust the DM-test in a way proposed by Harvey et al. (1997) to meet this problem. To that end we adjust the DM-statistic by multiplying it with the square root of $(P-1)/P$. We also compare this adjusted statistic with critical values obtained from a Student's t-distribution with $P-1$ degrees of freedom, instead of using the standard normal distribution.

The second test we use is proposed by Van Dijk and Franses (2003). This test is put forward to specifically compare the forecasting performance of linear and nonlinear time series models. Van Dijk and Franses (2003) argue that nonlinear time series models often do not outperform linear models in out-of-sample forecasting, despite their superior in-sample fit. They suggest that this might be due to the use of inappropriate evaluation criteria and suggest using a criterium that weights the forecasted observations. Therefore, they use the DM-statistic and adjust it by using different weight functions in such a way that more weight is placed on the relevant observations which are most associated with non-linearity (for example, turning points). Weight functions that they propose focus on one or two tails (LT and RT) of the distribution of the dependent variable. We look at the same three weight functions, being: $w_T(y_t) = 1 - \phi(y_t)/\max(\phi(y_t))$, $w_{LT}(y_t) = 1 - \Phi(y_t)$ and $w_{RT}(y_t) = \Phi(y_t)$. Here,

$\phi(\cdot)$ is the density function of y_t and $\Phi(\cdot)$ is the cumulative distribution function of y_t. The density function of y_t is estimated using the relevant in-sample observations and using a normal kernel function with automatic bandwidth selection. The empirical cumulative density function is used as an estimate of $\Phi(y_t)$.

5.4 Results

As announced in the previous section, we save space by mainly focussing on the variable INPROD. We discuss the models with every new SPF variable to use in forecasting (std, 5p, 95p and nr) one by one, first focussing on the results for INPROD obtained using a rolling estimation window of 124 observations and then analyzing the robustness of the results. We check the robustness of these results in three different ways. First of all, we look at fixed estimation window forecasts for INPROD. Second, we look at forecasts for INPROD obtained using a rolling estimation window of 104 observations. Finally, we analyze forecasts for the other four macroeconomic variables obtained using a rolling estimation window of 124 observations.

We perform these robustness checks for several reasons. The first reason is that we want to know how our results hold in different forecasting situations. Especially the MRS-models with time varying transition probabilities are difficult to estimate and the estimation procedure is very time-consuming. Therefore, it would be easier to just estimate the parameters of the optimal model once and use this in the subsequent periods to forecast (fixed estimation window), but we need to know if this works properly. Furthermore, to give conclusions to what extent the same model(s) can be used for different macroeconomic variables, we analyze the other four selected variables.

The second reason is that we are not sure beforehand which method works best. For example, in general one might believe that a rolling estimation window gives better results than a fixed estimation window (as it allows for slowly changing parameters), but we find evidence that the opposite seems to be true for INPROD (see Table 5.1). Furthermore, we have a limited data set and we both need enough observations to

Table 5.1: RMSPE's for models estimated for growth of INPROD. Forecasts are created using a rolling estimation window ('roll') or a fixed estimation window ('fix') and RMSPE's are calculated over 40 or 60 forecasts. All models include one lag as explanatory variable and the SPF variables as indicated below. The 'c' or 'v' indicates if a constant or varying variance per regime is used for the error terms in the MRS models. The bold RMSPE's are the ten smallest RMSPE's in that column.

INPROD	**Last release**			**First release**		
	Roll 40	**Fix 40**	**Roll 60**	**Roll 40**	**Fix 40**	**Roll 60**
AR	1.247	1.226	1.126	1.198	1.186	1.064
ARX-mean	0.967	0.918	0.893	0.864	0.817	0.773
ARX-med	0.966	**0.911**	0.896	0.861	**0.804**	0.773
MRScons-mean-c	0.959	0.959	0.902	0.844	0.848	0.776
MRScons-mean-v	0.994	0.969	0.899	0.880	0.877	0.784
MRScons-med-c	0.997	0.954	0.894	0.896	0.845	0.769
MRScons-med-v	0.999	0.938	0.911	0.887	0.835	0.790
SPF-mean	**0.916**	0.916	**0.857**	**0.837**	0.837	**0.740**
SPF-med	**0.909**	**0.909**	**0.853**	**0.823**	0.823	**0.729**
ARX-mean+std	0.977	0.939	0.918	0.876	0.832	0.793
ARX-mean+5p	0.960	0.917	0.912	0.859	0.816	0.784
ARX-mean+95p	0.978	0.969	0.907	0.876	0.857	0.787
ARX-mean+nr	0.965	0.915	0.901	0.858	0.814	0.776
ARX-med+std	0.978	0.936	0.920	0.874	0.823	0.793
ARX-med+5p	0.982	0.941	0.900	0.875	0.828	0.783
ARX-med+95p	0.952	0.917	0.910	0.852	0.809	0.780
ARX-med+nr	0.972	**0.912**	0.908	0.861	0.805	0.780
MRScons-mean+std-c	0.983	0.928	0.956	0.861	0.811	0.828
MRScons-mean+std-v	0.965	0.931	0.931	0.845	0.822	0.803
MRScons-mean+5p-c	**0.917**	0.943	0.910	**0.810**	0.837	0.771
MRScons-mean+5p-v	0.981	0.918	0.903	0.870	0.827	0.761
MRScons-mean+95p-c	0.966	0.938	0.927	0.844	0.816	0.805
MRScons-mean+95p-v	0.993	0.968	0.932	0.879	0.850	0.814
MRScons-mean+nr-c	**0.917**	0.921	0.907	**0.799**	**0.795**	0.786
MRScons-mean+nr-v	**0.935**	0.953	0.914	**0.811**	0.831	0.788
MRScons-med+std-c	0.990	0.959	0.942	0.873	0.847	0.814
MRScons-med+std-v	1.008	0.956	0.913	0.891	0.842	0.776
MRScons-med+5p-c	0.978	0.974	0.892	0.865	0.861	0.774
MRScons-med+5p-v	0.985	1.006	0.897	0.870	0.894	0.781
MRScons-med+95p-c	0.976	0.914	0.911	0.861	**0.798**	0.782
MRScons-med+95p-v	0.960	0.924	0.896	0.844	0.816	0.754
MRScons-med+nr-c	0.947	0.965	0.909	**0.832**	0.842	0.782
MRScons-med+nr-v	**0.920**	**0.907**	0.901	**0.806**	**0.794**	0.777
MRSvar-mean+std-c	1.029	**0.908**	0.884	0.934	**0.803**	**0.741**
MRSvar-mean+std-v	**0.945**	0.923	**0.844**	0.863	0.821	**0.712**
MRSvar-mean+5p-c	1.035	0.932	0.978	0.943	0.832	0.844
MRSvar-mean+5p-v	0.976	0.931	**0.823**	0.889	0.827	**0.694**
MRSvar-mean+95p-c	0.978	0.982	1.027	0.872	0.904	0.928
MRSvar-mean+95p-v	0.975	0.913	0.946	0.844	**0.793**	0.828
MRSvar-mean+nr-c	0.952	**0.880**	**0.873**	0.840	**0.792**	0.750
MRSvar-mean+nr-v	**0.882**	**0.891**	**0.870**	**0.773**	**0.799**	0.752
MRSvar-med+std-c	1.007	0.949	**0.873**	0.926	0.848	**0.736**
MRSvar-med+std-v	1.028	0.919	**0.827**	0.961	0.814	**0.697**
MRSvar-med+5p-c	1.038	0.923	0.885	0.942	0.819	0.765
MRSvar-med+5p-v	1.030	0.961	**0.854**	0.946	0.855	**0.728**
MRSvar-med+95p-c	**0.933**	1.147	0.934	**0.810**	1.071	0.810
MRSvar-med+95p-v	**0.936**	**0.892**	0.986	**0.810**	**0.773**	0.873
MRSvar-med+nr-c	1.071	**0.885**	0.874	0.976	**0.783**	**0.745**
MRSvar-med+nr-v	1.089	**0.875**	**0.865**	0.976	**0.779**	**0.746**

Table 5.2: RMSPE's for models estimated for UNEMP and growth of NGDP, PGDP and HOUSING. The RMSPE's are calculated over 40 forecasts, which are created using a rolling estimation window. For further information, see the caption of Table 5.1.

Roll 40	**Last release**				**First release**			
	NGDP	**PGDP**	**UNEM**	**HOUS**	**NGDP**	**PGDP**	**UNEM**	**HOUS**
AR	0.776	0.331	0.392	7.390	0.702	0.349	0.391	7.683
ARX-mean	0.443	0.279	0.121	5.338	0.358	0.264	0.113	5.425
ARX-med	0.433	0.278	0.120	**5.089**	0.349	0.266	0.115	**5.176**
MRScons-mean-c	0.443	0.271	0.120	5.348	0.354	0.256	0.116	5.392
MRScons-mean-v	0.441	0.278	0.123	5.369	0.348	0.270	0.114	5.415
MRScons-med-c	**0.429**	0.277	0.120	**5.096**	**0.341**	0.264	0.120	**5.139**
MRScons-med-v	0.437	0.278	0.122	**5.091**	0.346	0.267	0.118	**5.142**
SPF-mean	0.458	**0.237**	0.135	6.010	0.363	**0.230**	0.129	6.126
SPF-med	0.450	**0.241**	0.127	5.624	0.354	**0.239**	0.123	5.735
ARX-mean+std	0.441	0.278	0.120	5.676	0.355	0.263	**0.111**	5.754
ARX-mean+5p	0.440	0.281	0.122	5.593	0.353	0.265	0.114	5.679
ARX-mean+95p	0.443	0.277	**0.114**	5.594	0.357	0.263	**0.104**	5.666
ARX-mean+nr	0.441	0.282	0.119	5.626	0.359	0.266	0.111	5.709
ARX-med+std	0.433	0.277	0.119	5.362	**0.346**	0.265	0.114	5.442
ARX-med+5p	**0.433**	0.279	0.120	**5.137**	0.349	0.266	0.115	**5.227**
ARX-med+95p	**0.432**	0.280	0.117	5.401	**0.345**	0.268	**0.113**	5.465
ARX-med+nr	**0.425**	0.280	**0.117**	5.430	**0.342**	0.267	0.113	5.513
MRScons-mean+std-c	0.451	**0.264**	0.124	5.301	0.355	**0.253**	0.119	5.311
MRScons-mean+std-v	0.447	**0.267**	0.123	5.550	0.352	0.261	0.117	5.567
MRScons-mean+5p-c	0.447	**0.257**	0.130	5.438	0.358	**0.250**	0.124	5.481
MRScons-mean+5p-v	0.453	0.276	0.123	5.453	0.352	0.274	0.116	5.483
MRScons-mean+95p-c	0.450	**0.263**	**0.112**	5.312	0.364	**0.251**	**0.107**	5.315
MRScons-mean+95p-v	0.446	0.269	0.119	5.289	0.355	0.260	**0.110**	5.348
MRScons-mean+nr-c	0.442	0.271	0.129	5.690	0.356	**0.251**	0.124	5.752
MRScons-mean+nr-v	0.433	0.275	**0.115**	5.623	0.354	0.261	**0.108**	5.695
MRScons-med+std-c	0.443	0.270	0.117	5.512	0.351	0.265	0.114	5.505
MRScons-med+std-v	0.449	0.270	0.120	5.394	**0.345**	0.269	0.118	5.414
MRScons-med+5p-c	**0.427**	**0.268**	**0.116**	**5.031**	**0.343**	0.259	0.114	**5.082**
MRScons-med+5p-v	0.473	0.272	0.121	**5.049**	0.379	0.265	0.117	**5.094**
MRScons-med+95p-c	0.445	0.268	**0.110**	5.389	**0.344**	0.260	**0.111**	5.381
MRScons-med+95p-v	**0.430**	0.276	**0.117**	5.274	**0.334**	0.269	0.114	5.294
MRScons-med+nr-c	**0.423**	0.280	0.120	5.406	**0.339**	0.267	0.120	5.485
MRScons-med+nr-v	**0.420**	0.272	0.120	5.303	**0.338**	0.264	0.116	5.367
MRSvar-mean+std-c	0.494	0.269	0.129	5.826	0.473	0.265	0.121	5.922
MRSvar-mean+std-v	0.491	0.277	0.131	5.302	0.477	0.266	0.123	5.390
MRSvar-mean+5p-c	0.517	0.273	0.124	**4.945**	0.389	0.264	0.119	**5.061**
MRSvar-mean+5p-v	0.506	0.277	0.124	5.559	0.397	0.270	0.114	5.553
MRSvar-mean+95p-c	0.535	0.281	0.120	5.798	0.434	0.261	0.115	5.886
MRSvar-mean+95p-v	0.514	0.275	0.124	5.741	0.399	0.262	0.119	5.708
MRSvar-mean+nr-c	0.491	0.275	**0.114**	5.353	0.385	**0.255**	**0.105**	5.478
MRSvar-mean+nr-v	0.460	0.276	**0.108**	5.358	0.384	**0.254**	**0.100**	5.487
MRSvar-med+std-c	**0.424**	**0.262**	0.127	**5.153**	0.353	0.257	0.123	**5.214**
MRSvar-med+std-v	**0.423**	0.275	0.127	5.391	0.355	0.267	0.120	5.510
MRSvar-med+5p-c	0.502	**0.266**	0.125	5.356	0.373	**0.255**	0.123	5.334
MRSvar-med+5p-v	0.473	**0.266**	0.123	5.453	0.369	0.257	0.121	5.462
MRSvar-med+95p-c	0.504	0.280	0.123	5.644	0.407	0.263	0.121	5.742
MRSvar-med+95p-v	0.499	0.273	0.122	5.585	0.402	0.259	0.120	5.690
MRSvar-med+nr-c	0.475	0.271	0.119	**5.107**	0.374	**0.254**	0.116	**5.238**
MRSvar-med+nr-v	0.451	0.276	**0.115**	**5.136**	0.365	0.262	**0.111**	**5.249**

estimate the model parameters and to get enough forecast errors to test if forecasting differences are significant. As we do not know what is enough in both cases, we use

the 124-40 as well as the 104-60 proportion, where 124 (104) indicates the number of observations used for the estimations, and 40 (60) the number of created forecasts.

The RMSPE's of a large part of the estimated models can be found in Tables 5.1 and 5.2. Although models without the mean or median of SPF forecasts are estimated too, results from these models are omitted from the tables, because these models appeared to have a very poor forecasting performance in all cases.

5.4.1 Standard deviation of SPF

Columns 2 and 5 of Table 5.1 show the RMSPE's of the sets of 40 INPROD forecasts created by using a rolling estimation window.

What is remarkable is that the simple mean and median of the SPF forecasts perform very well over this period (RMSPE's of 0.916 and 0.909). It is obviously more precise than any of other 7 benchmarks. Clearly, it is also difficult for the alternative models to outperform these SPF forecasts.

If we look at the models were std is included we see that the MRS model with one lag and the mean of the SPF forecasts as explanatory variables, the standard deviation of SPF forecasts used to model regime switches and a varying variance of the error terms per regime (MRSvar-mean+std-v) is amongst the ten models with the lowest RMSPE's. Before we discuss the forecasting performance in detail it might be interesting to see what the estimated model parameters look like. To that extent we estimate the model parameters over the complete data set and compared the results with the models estimated over parts of the data set. The results look quite the same, so we discuss the estimates for the complete data set here.

Figure 5.1 shows the standard deviation of SPF forecasts, the estimated smoothed probabilities of regime two and recessions as officially declared by the NBER. It can be seen that this model estimates one regime that occurs most of the time. This happens when the standard deviation of SPF forecasts is not too high. When the standard deviation of SPF forecasts rises above approximately 1.1 the process switches to regime

two. In panel 1 of Table 5.3, the estimated parameters are given with their significance. In regime one, we see a negative constant, a significantly (5%) positive coefficient for the lag of INPROD growth, a significantly positive coefficient for the mean of SPF forecasts and a variance of the error terms of around 0.6. In regime two, the estimated intercept is much more negative, the coefficient for the lag is not significantly different from zero anymore, the coefficient for the mean of SPF forecasts is a little bit higher and still significant and the variance of the error terms is with a value of around 2 much higher than in regime one.

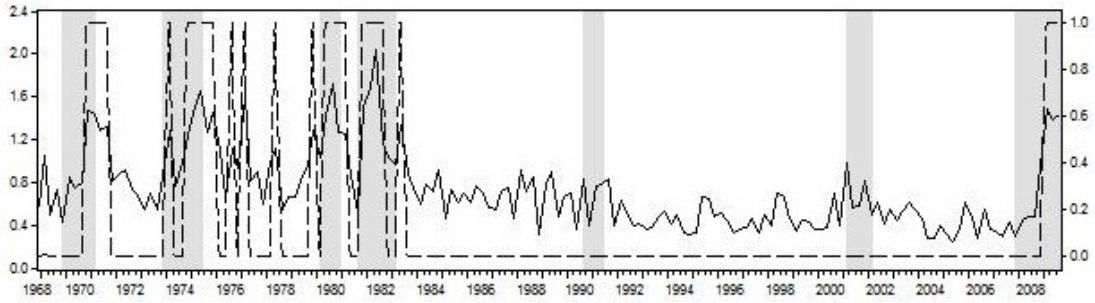

Figure 5.1: This figure shows the probability of regime two as estimated by the MRSvar-mean+std-v model for growth of INPROD in combination with the standard deviation of SPF forecasts. The solid line, with its scale on the left, is the standard deviation of SPF forecasts. The dashed line, fluctuating between 0 and 1 and with its scale on the right, is the probability of regime two. The shaded areas are recessions as indicated by the NBER.

In Figure 5.2, growth of INPROD, the estimated smoothed probabilities of regime two and recessions as officially declared by the NBER are shown. During regime two as indicated by the model, growth of INPROD is on average negative and during regime one it is positive. Furthermore, in regime two growth of INPROD is much more volatile than in regime one and regime two covers more extreme values than regime one.

We can conclude from these results that the standard deviation of SPF forecasts might predict volatile periods in which a relatively higher weight must be put on the mean of SPF forecasts and a lower weight on the lag of INPROD growth in combining

Table 5.3: Coefficients of models estimated for INRPOD data from the fourth quarter of 1968 to the third quarter of 2009. If the coefficients are significantly different from 0 at the 5%-level is indicated by '*'.

	Regime 1	**Regime2**
	MRSvar-mean+std-v	
c	-0.064	-0.240
lag	0.274*	-0.000
mean	0.808*	1.038*
var	0.563*	2.170*
	MRScons-mean+5p-c	
c	-0.576	-0.217
lag	0.180*	0.110
mean	0.902*	1.744*
5p	0.183*	-0.619*
var	0.378*	0.378*
	MRSvar-med+95p-c	
c	-0.100	1.011
lag	0.244*	-0.127
median	0.905*	0.514*
var	0.787*	0.787*
	MRSvar-mean+nr-v	
c	-0.165	0.367
lag	0.100*	0.626*
mean	1.106*	-0.180
var	0.812*	0.236*

forecasting models.

If we take a closer look at the forecasting performance of this model, we see that the RMSPE of this model is lower than all the benchmark models (last release data), but it is not lower than the mean or median of the SPF forecasts. In the last column of Table 5.4, we see that the MRSvar-mean+std-v model does produce significantly (10%) more accurate forecasts than a few of the benchmark models, but not all.

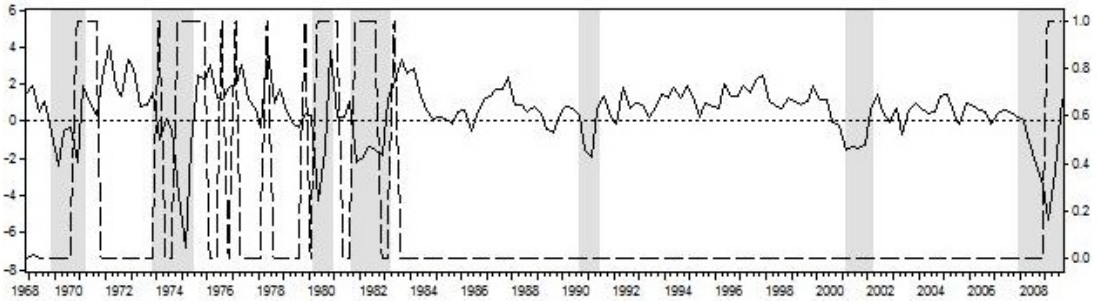

Figure 5.2: This figure shows the probability of regime two as estimated by the MRSvar-mean+std-v model for growth of INPROD in combination with growth of INPROD. The solid line, with its scale on the left, is growth of INPROD. The dashed line, fluctuating between 0 and 1 and with its scale on the right, is the probability of regime two. The shaded areas are recessions as indicated by the NBER.

How do these results carry over to different forecasting circumstances? Let us first look at forecasts created using a fixed estimation window. RMSPE's can be found in the third and sixth column of Table 5.1.

The RMSPE's of the two MRSvar-mean+std models are quite low and the RMSPE's of the model with a constant variance for the error terms is amongst the ten models with the lowest RMSPE's and has lower RMSPE's than all the benchmark models and than mean and median of SPF (see Table 5.1). However, according to both the unweighted and weighted tests there is no evidence that these models produce significantly more accurate forecasts than the benchmark models and SPF mean and median.

We expect that using a shorter estimation window might result in more significant improvements in forecast accuracy for the following reason. The models with std that perform best are MRSvar models. These models need enough observations in both regimes in order for the parameters to be estimated properly. We can see from Figures 5.1 and 5.2 that most regime switches occur in the beginning of the data set. Therefore, a shorter estimation window might be good enough. On the other hand, we do get low RMSPE's with the MRSvar models estimated over a longer horizon, but it is hard to find significant differences. This can be due to the low amount of forecasts over which

Table 5.4: This table shows if the models in the header of the column forecast significantly more accurate than the models in the header of the row, according to the tests as described in Section 5.3.4. The models are estimated for INPROD data, a rolling estimation window is used and 40 forecasts are created. '++' indicates that the model in the column header produces more accurate forecasts than the model mentioned in the row header according to the unweighted test. '+' indicates that not the unweighted test, but at least one of the weighted tests shows a significant difference. A significance level of 10% is used.

INPROD Roll 40	**MRSvar-mean+nr-v**	**MRScons-med+nr-v**	**MRScons-mean+nr-c**	**MRScons mean+5p-c**	**MRSvar med+95p-c**	**MRSvar mean+std-v**
	Last release					
AR	++	++	++	++	++	++
ARX-mean	++	+		+		
ARX-med	++	+	+	++		
MRScons-mean-c	++	++	++	++		
MRScons-mean-v	++	++	++	++	++	++
MRScons-med-c	++	++	++	++	++	++
MRScons-med-v	++	++	++	++	++	
SPF-mean	+					
SPF-med	+					
	First release					
AR	++	++	++	++	++	++
ARX-mean	++			+		
ARX-med	++	++	++	++	++	
MRScons-mean-c	++	++	++	++		
MRScons-mean-v	++	++	++	++	++	
MRScons-med-c	++	++	++	++	++	
MRScons-med-v	++	++	++	++	++	
SPF-mean	+			+		
SPF-med	+	+	+	+	+	

the tests are performed, and more forecasts might be beneficial in finding significant results. Therefore, in this case, the 104-60 proportion might give better results than the 124-40 proportion.

Columns 4 and 7 of Table 5.1 demonstrate that the SPF mean and SPF median give again lower RMSPE's than all the benchmark models in the last 60 quarters. There are, however, two MRSvar models using std which create RMSPE's lower than these SPF mean and SPF median, namely MRSvar-mean+std-v and MRSvar-median+std-v. Here we find that MRSvar-mean+std-v significantly improves on almost all the bench-

mark models and has a lower RMSPE than the mean and median of SPF forecasts, although not significantly (see column 5 of Table 5.5). The MRSvar-median+std-v is even more accurate (see Table 5.1 and column 4 of Table 5.5). Interpretation of the models is again similar to the interpretation of the MRSvar model with std estimated over the complete data set.

Table 5.5: This table shows if the models in the headers of the columns forecast significantly more accurate than the models in the headers of the rows, according to the tests as described in Section 5.3.4. The models are estimated for INPROD data, a rolling estimation window is used and 60 forecasts are created. See for more information the caption of Table 5.4.

INPROD Roll 60	**MRSvar-mean+5p-v**	**MRSvar-med+5p-v**	**MRSvar-med+std-v**	**MRSvar-mean+std-v**	**MRSvar-med+nr-v**	**MRSvar-mean+nr-v**
	Last release					
AR	++	++	++	++	++	++
ARX-mean	+		++	+		
ARX-med	+		++	++		
MRScons-mean-c			++	++		
MRScons-mean-v	+		++	+		
MRScons-med-c			++			
MRScons-med-v	++		++	++	++	
SPF-mean	+					
SPF-med	+					
	First release					
AR	++	++	++	++	++	++
ARX-mean			++	++		
ARX-med	+		++	++		
MRScons-mean-c			++	++		
MRScons-mean-v	++		++	++		
MRScons-med-c			++			
MRScons-med-v	++		++	++		
SPF-mean	+					
SPF-med						

Finally, we look at four other macroeconomic variables for which SPF forecasts are available. These are NGDP, PGDP, UNEMP and HOUSING. See Table 5.2 for RMSPE's and see Table 5.6 for the test results on forecast accuracy of some of the models.

Before we look at std, note from the third and seventh column of Table 5.2 that the mean and median of SPF forecasts for PGDP are more accurate than any of the

benchmark models. For this data set these two simple forecasts seem even harder to beat than for INPROD, because also not one of the alternative models has a lower RMSPE. Therefore, it is maybe needless to say that none of the models improves (significantly) on SPF mean and median.

Table 5.6: This table shows if the models in the headers of the columns forecast significantly more accurate than the models in the headers of the rows, according to the tests as described in Section 5.3.4. The models are estimated for NGDP, PGDP, UNEMP and HOUSING data, a rolling estimation window is used and 40 forecasts are created. See for more information the caption of Table 5.4.

Roll 40	**PGDP MRScons-mean+std-c**	**HOUSING MRSvar-mean+5p-c**	**UNEMP ARX-mean+95p**	**UNEMP MRScons-mean+95p-c**	**NGDP MRScons-med+nr-c**	**HOUSING MRSvar-med+nr-c**
	Last release					
AR	++	++	++	++	++	++
ARX-mean	++	++	++	++	+	++
ARX-med	++				+	+
MRScons-mean-c	++	++		++	++	+
MRScons-mean-v	++	++	++	++		+
MRScons-med-c	++	+		+	++	+
MRScons-med-v	++	+	++	++	++	+
SPF-mean		++	++	++	++	++
SPF-med		++	++	++	+	++
	First release					
AR	++	++	++	++	++	++
ARX-mean	++	++	++	+	++	++
ARX-med	++		++	+		+
MRScons-mean-c	+	++	++	++	++	+
MRScons-mean-v	++	++	++	+		+
MRScons-med-c	++		++	++		+
MRScons-med-v	++		++	++		+
SPF-mean		++	++	++	+	++
SPF-med		++	++	++		++

There are four models interesting if we look at std (not making a distinction here between a constant and a varying variance for the error terms): MRSvar-med+std estimated over NGDP data, MRSvar-med+std estimated over HOUSING data and MRSvar-med+std and MRScons-mean+std estimated over PGDP data.

The first three models are very similar to the estimated MRSvar-mean+std model for INPROD data. In these estimated models one regime occurs when std rises and

that is around periods declared as recessions by the NBER. For NGDP the intercept is much lower and negative in this regime, the coefficient of the lag is negative as opposed to around 0 in the first regime and the coefficient of the median of SPF is around 1.7 as opposed to 1 in the first regime. For HOUSING we see a much higher positive intercept in the 'recessionary' regime, a coefficient of the lag that does not differ much between the regimes and a coefficient of median of SPF that is around one in the first regime and around 1.2 in the second. For PGDP the differences between the regimes in coefficient estimation differ greatly over the 40 models estimated.

The fourth model that performs quite well, is MRScons-mean+std estimated for PGDP. It is however hard to define where this success comes from exactly, as the 40 estimated models differ very much. Some only show one switch between the regimes, whereby regime one occurs the first half of the estimation period and the second regime the second half, where the median of SPF receives relatively more weight. Other models show multiple switches and different parameter estimates.

Although all these models are in the top ten of models with the lowest RMSPE's (see Table 5.2), they in general do not improve significantly on the benchmarks. As for INPROD this might be due to the relatively small number of forecasts over which the tests are performed.

In general we can conclude that the variable std of survey forecasts has predictive value in macroeconomic forecasting, especially when used in MRSvar models. However, as these models need enough observations in both regimes for the parameters to be estimated properly and as these models are quite hard to estimate, more data is needed to get potentially significant results across multiple macroeconomic variables.

5.4.2 5th percentile

The MRS model with constant transition probabilities, constant variance of the error terms and as independent variables one lag and SPF mean and 5p (MRScons-mean+5p-c) has even lower RMSPE's than the best model with std we found for INPROD (basic

forecasting scheme). Although there are differences between the models estimated to create forecasts and the model estimated for the complete data set, some general features of the full data model are prevalent in the smaller models. Therefore we will again look at this full data model.

This model distinguishes two regimes that occur about equally often, see Figure 5.3. The smoothed probability of regime two follows the fluctuations in growth of INPROD very closely. On average this growth is higher in regime two, with also a higher 5th and 95th percentile. In regime one, growth is slightly negative on average. In panel 2 of Table 5.3, we present the estimated coefficients in both regimes. We see that in regime one where growth of INPROD is most of the time declining, the coefficients of the mean and 5th percentile of survey forecasts have the same sign. The mean of survey forecasts has a coefficient close to 1 and when the most 'negative' forecasters predict a negative growth in this regime the final growth forecast should be lowered, ceteris paribus. In regime two, where growth of INPROD is most of the time increasing, the coefficients of the mean and 5th percentile of survey forecasts have opposite signs. The coefficient of the mean of survey forecasts indicates that the average forecast is too modest and should be inflated, ceteris paribus, and when the most 'negative' forecasters predict a negative growth in this regime the final forecast should be increased and visa versa.

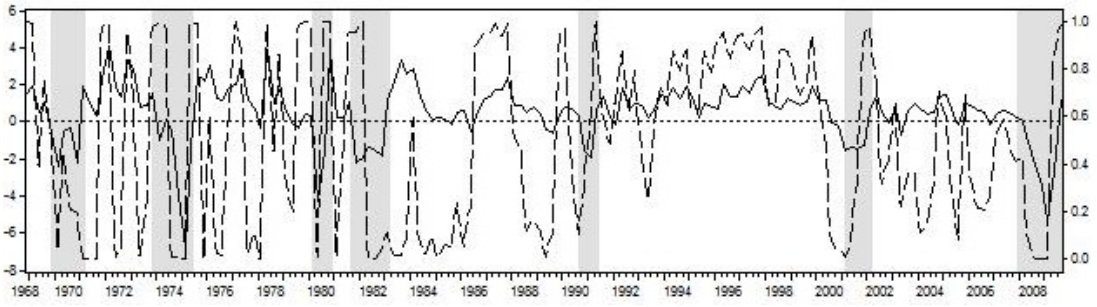

Figure 5.3: This figure shows the probability of regime two as estimated by the MRScons-mean+5p-c model for growth of INPROD in combination with growth of INPROD. For more information, see the caption of Figure 5.2.

The forecasting performance of this model (MRScons-mean+5p-c) is very good. Its RMSPE's are lower than the RMSPE's of all benchmark models and even lower than the RMSPE's of SPF mean and median if we look at first release data, see columns 2 and 5 of Table 5.1. It produces significantly more accurate forecasts according to the unweighted test than all benchmark models except one, both compared with first release and last release data. The remaining benchmark model is beaten according to at least one of the weighted tests, again both compared to first and last release data, and SPF mean and median are beaten significantly according to at least one weighted test if we look at first release data, see Table 5.4.

For 5p we find no evidence in the fixed estimation window situation that the variable has predictive value. None of the models with 5p is capable of outperforming the benchmarks.

Using a rolling estimation window again, but now of size 104, this is different. The models MRSvar-mean+5p-v and MRSvar-median+5p-v are amongst the ten models with the lowest RMSPE's. Especially the first one is interesting. It has the lowest RMSPE based on first and last release data, also lower than mean and median of SPF forecasts and the differences in RMSPE's are often significant (see Table 5.5). The interpretation is quite similar to the same model with std to predict regime switches. There is one regime that occurs less often than the other regime and that occurs around the same quarters as the second regime occurs in the model with std. In that model the probability of a regime switch increases if std increases, but here this probability increases if 5p declines. The coefficient for the mean of SPF is different in both regimes and the standard deviation of the error terms is much higher in the second regime than in the first. So besides a higher std of survey forecasts, also a lower 5p is a signal that a different regime is coming up in which INPROD growth is more volatile and thus different weights should be used in forecasting models or forecasting combinations.

Finally for the explanatory variable 5p, let us look at the other macroeconomic variables. The most interesting models in which 5p is used are the MRScons-median+5p-c and MRSvar-mean+5p-c models for HOUSING. The first of these two shows much

resemblance with the MRScons-mean-5p-c model estimated for INPROD, as the parameters have approximately the same estimated values (except the intercepts). The second shows much resemblance with the MRSvar-mean+5p-v estimated over a small horizon for INPROD, discussed above, and with the MRSvar-mean+std-v estimated for INPROD, discussed in Section 5.4.1. It estimates a recession regime when 5p decreases sharply with a large negative intercept, a negative lag parameter and a coefficient of 0.55 for mean and a non-recession regime with a positive intercept, a positive lag parameter and a coefficient of 1.15 for mean.

Both models have a lower first release and last release RMSPE than all the benchmarks. In most of the cases this difference is significant, see Table 5.6.

Also for the other three variables there are models in which 5p is included that perform well in forecasting. These are mainly MRScons models. However, these models are in general not capable of significantly outperforming the benchmarks. Only for PGDP are these models capable of outperforming all benchmark models, but SPF mean and median remain superior here.

So the main conclusions about the 5th percentile of SPF-forecasts as predictive variable is twofold. On the one hand we can see it as a disagreement variable that is slightly different from std and with which we find similar optimal models as with std, namely MRSvar models where 5p indicates if a recessionary regime is coming up. On the other hand we also find well performing MRScons models with this variable as explanatory variable. Here we see that the 5p variable has significant forecasting information additional to a location variable of SPF, but this information is different in different regimes. In the slow or negative growth regime the parameter of 5p is positive and has the same sign as the parameter of the location variable, indicating that the pessimists should be followed in this regime. In the growth regime the parameter of 5p is negative and has the opposite sign of the parameter of the location variable, indicating that now the pessimists should be followed in the opposite direction.

5.4.3 95th percentile

The third explanatory variable of interest, 95p, seems to have predictive value in the MRS model with one lag and the median of survey forecasts as explanatory variables, with 95p as explanatory variable for regime switches and with a constant variance of the error terms (MRSvar-med+95p-c). If we estimate this model over the complete data set it again partly resembles the estimated models used to create the forecasts. We find one regime that occurs the most and that is when the 95th percentile of growth forecasts is lower than approximately 3%, see Figure 5.4. In this regime (see the third panel of Table 5.3) the constant is slightly negative, the coefficient of the lag is significantly positive and the coefficient of the median is significantly different from zero with a value of around 0.9. When 95p rises above 3 we find a regime with an intercept around 1, a coefficient of the lag that is not significantly different from 0 and a coefficient of the median which is significantly different from 0 with a value of around 0.5. In both regimes the variance of the error terms is approximately 0.8. In Figure 5.5 we see that regime two often occurs right after a recession or otherwise right after a dip in INPROD growth. During regime two growth of INPROD is on average much higher than during regime one and always positive.

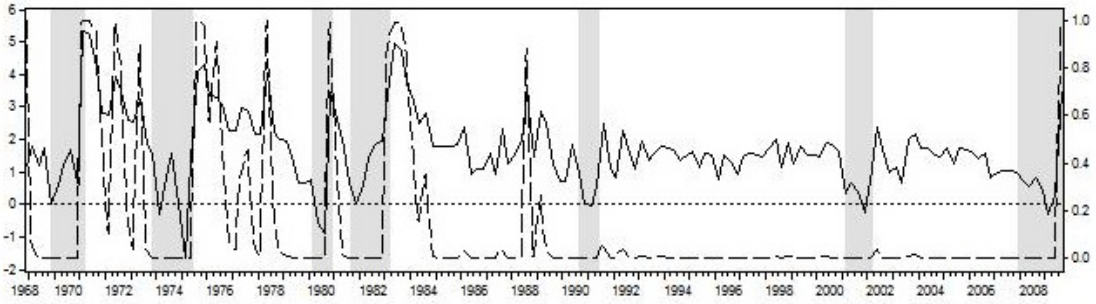

Figure 5.4: This figure shows the probability of regime two as estimated by the MRSvar-med+95p-c model for growth of INPROD in combination with the 95th percentile of SPF forecasts. For more information, see the caption of Figure 5.1.

This all indicates that when the 95th percentile of survey forecasts increases it is likely that a recovery period is coming up with on average high growth of industrial

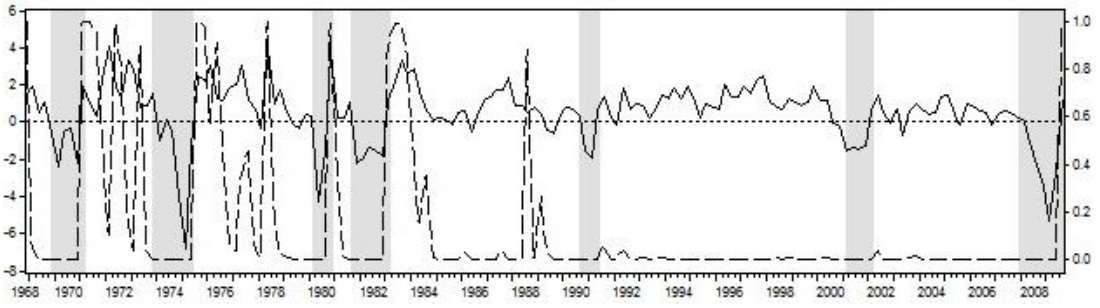

Figure 5.5: This figure shows the probability of regime two as estimated by the MRSvar-med+95p-c model for growth of INPROD in combination with growth of INPROD. For more information, see the caption of Figure 5.2.

production. This is a plausible result. In this period the lag of INPROD should not receive much weight in the forecast, while the constant should be higher and the coefficient of SPF median should be lower, as forecasters might be too optimistic in general.

The forecasting performance of the MRSvar-med+95p-c model is quite good, see again Table 5.1, columns 2 and 5. The RMSPE is clearly lower than the RMSPE's of all benchmark models, both calculated over first release and last release data. However, the difference is not always significant (see Table 5.4) and the SPF mean and SPF median are more precise predictors.

The same model that performs well using a rolling estimation window also performs well using a fixed estimation model, namely a MRSvar-med+95p model. Only here the model that performs well has varying instead of constant variance for the error terms. However, the differences between these two models are very small in case of the rolling estimation window. Furthermore, the estimation and interpretation of this model is much the same. The variance of the error terms is in the first regime around 0.9 and in the second regime around 1, so this is essentially the same as a MRSvar-med+95p-c model. The MRSvar-med+95p-c model did not perform well using a fixed estimation window, because the first estimation of the 40 rolling window estimations (used to create the 40 fixed estimation window forecasts) did not converge to parame-

ters similar to the complete data estimation and similar to most of the other 39 estimations, as opposed to the first set of MRSvar-med+95p-v parameters estimated.

Now we see (Table 1, columns 3 and 6) that the RMSPE of this model is lower than all benchmark models and than the simple mean and median of SPF, both for first and last release data. In the penultimate column of Table 5.7 we see that the difference is almost always significant.

Table 5.7: This table shows if the models in the headers of the columns forecast significantly more accurate than the models in the headers of the rows, according to the tests as described in Section 5.3.4. The models are estimated for INPROD data, a fixed estimation window is used and 40 forecasts are created. See for more information the caption of Table 5.4.

INPROD Fix 40	**MRSvar-med+nr-v**	**MRSvar-mean+nr-c**	**MRScons-med+nr-v**	**MRScons-mean+nr-c**	**MRSvar-med+95p-v**	**MRScons-med-95p-c**
	Last release					
AR	++	++	++	++	++	++
ARX-mean		+		+		
ARX-med			+		++	
MRScons-mean-c	++	++	++	++	++	++
MRScons-mean-v	++	++	++	+	++	++
MRScons-med-c	++	++	++	++	++	++
MRScons-med-v	++	++	++	+	++	+
SPF-mean		+				+
SPF-med		+	+	+	+	+
	First release					
AR	++	++	++	++	++	++
ARX-mean				+		
ARX-med			++	+	++	
MRScons-mean-c			++	++	++	++
MRScons-mean-v			++	++	++	++
MRScons-med-c	+		++	++	++	++
MRScons-med-v			++	++	++	++
SPF-mean				+		+
SPF-med			+	+	++	+

Another model that performs well, especially in terms of first release data, is the MRS model with constant transition probabilities, with as explanatory variables the median and 95p and with a constant variance for the error terms. However, the interpretation for this model is difficult, as the probabilities at the different regimes are often not close to zero or one and the coefficients in both regimes are often insignificantly

different from zero.

Surprisingly, we do not find models containing 95p that perform well in the rolling 60 forecasting practice, probably due to the fact that this time the model parameters did not converge to the best performing estimations.

Finally, let us look at the performance of 95p in models for other macroeconomic variables. There are again multiple models worth analyzing. These are all ARX or MRScons models estimated for NGDP, PGDP or UNEMP data. The most interesting models are probably the models estimated for UNEMP, because these models are capable of outperforming the benchmarks significantly. We find for example an ARX model where the mean has a coefficient of approximately 1.5 and 95p a coefficient of -0.35. In words, the higher the 'pessimists' predict unemployment, the lower the forecast should be for given values of mean SPF and previous unemployment. Stated differently, if there is a small group giving extremely high forecasts for unemployment instead of a large group giving moderately high forecasts, there should be an adjustment in the final forecast. Also in the MRScons models, 95p has a negative coefficient in both regimes, where in one regime it is more significant than in the other. The regime with the more significant negative coefficient for 95p occurs more often and has a coefficient for mean or median SPF of above 1, while the other regime has a coefficient for this of around 1.

Note that these latter models actually would fit better under the 5p header. Where for most variables the 95th percentile are the most optimistic forecasters, for UNEMP these are the most pessimistic forecasters. Also the interpretation of these models is quite similar to the interpretation of the models found in the previous section, where the pessimists variable gets a negative parameter at least part of the time.

The conclusion on the 95p variable in a optimist setting, or at least for INPROD, is that it can probably signal recovery regimes, but we would like to find more significant and robust results.

5.4.4 Number of forecasts

Finally, for the number of survey forecasts we find various models that seem capable of outperforming the benchmark models in forecasting INPROD, see Table 5.1 and Table 5.4. The model with the best forecasting performance is a MRS model with as explanatory variables one lag and the mean of survey forecasts, with nr as the explanatory variable for the regime switches and with a varying variance of the error terms. We again estimated this model over the complete data set to see what it looks like. In figure 5.6 we see the probability of regime 2 in combination with the number of SPF forecasts. The first regime estimated by the model occurs most of the time, notably when the number of forecasts is above approximately 27.

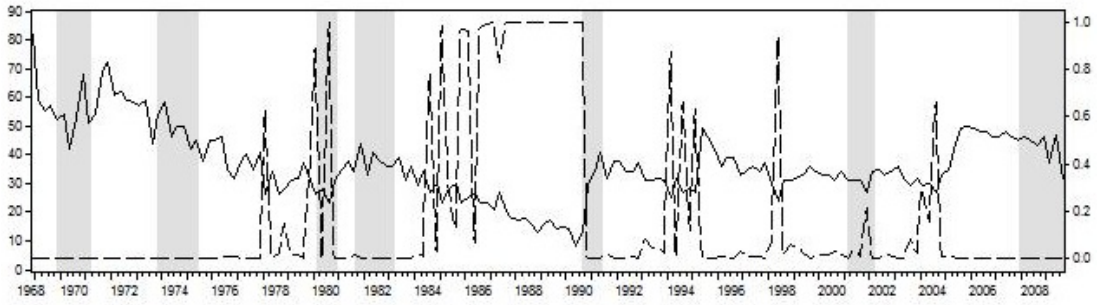

Figure 5.6: This figure shows the probability of regime two as estimated by the MRSvar-mean+nr-v model for growth of INPROD in combination with the number of SPF forecasts. For more information, see the caption of Figure 5.1.

Table 5.3 gives for the estimated coefficients and its significance. In regime one, when the number of forecasts is high, we find a negative intercept, a significantly positive coefficient for the lag, a significantly positive coefficient for the mean of SPF forecasts and a variance of the error terms of around 0.8. In regime two, when the number of forecasts is low, we find a positive intercept, a coefficient for the lag that is higher than in regime one and still significant, a coefficient for the mean of survey forecasts that is not significant anymore and a lower variance of the error terms.

In Figure 5.7 we observe when regime two occurs according to this model in combination with growth of INPROD. During regime two the dependent variable is much

more stable and not that volatile as during regime one.

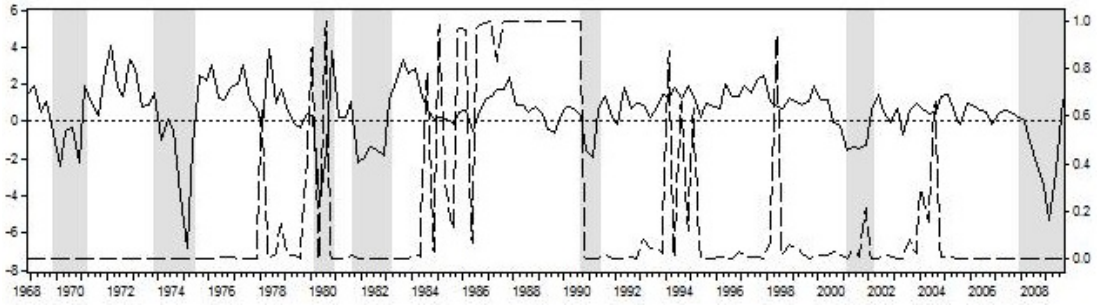

Figure 5.7: This figure shows the probability of regime two as estimated by the MRSvar-mean+nr-v model for growth of INPROD in combination with growth of INPROD. For more information, see the caption of Figure 5.2.

The forecasting results of this model are quite convincing. The RMSPE is lower than the RMSPE's of all the benchmark models and even lower than the SPF mean and SPF median, see Table 5.1. According to the standard test the difference is significant for all the benchmark models and according to the weighted tests the difference is also significant for SPF mean and SPF median, see Table 5.4.

There are two alternative conclusions which could be drawn from these results. The first is that when the dependent variable is more volatile, the expert forecasts should receive more weight and the number of forecasts can predict these different regimes. Thus, when the economic process is in a volatile regime, more experts produce forecasts than in more stable periods. This coincides with our premises in Section 5.2. It also fits the results of the first model presented in this section, where std predicts how volatile the variable of interest will be.

Another interpretation could be that when there are not enough forecasts the mean or median of the forecasts is less reliable nor informative and should not be used anymore or to a lesser extent. Whether the link between volatility of the dependent variable and the number of forecasts is a coincidence is not clear and should be investigated using other data sets.

Part of the models we find capable of producing accurate forecasts using a fixed es-

timation window are similar to the models we found using a rolling estimation window. The MRSvar-mean+nr-c, MRSvar-mean+nr-v, MRSvar-median+nr-c and the MRSvar-median+nr-v all produce forecasts with RMSPE's lower than all benchmark RMSPE's and in most cases the differences are significant in case of last release data, see Table 5.1, columns 3 and 6, and see Table 5.7. The estimated parameters and interpretations of these models are similar to that of the MRSvar-mean+nr-v estimation over the complete data set.

Two MRS models with constant transition probabilities and nr as one of the explanatory variables perform well, especially looking at first release data, see the third and sixth column of Table 5.1 again. In one regime the mean or median has a coefficient of around 0.8 and the coefficient of nr is insignificant. In the other regime, occurring less often, the mean or median has a coefficient of around 1.5 and nr a coefficient of around 0.02. This indicates that in specific situations more forecasters signify a higher growth of INPROD, ceteris paribus.

When we use a rolling estimation window of size 104 instead of 124, we see the opposite happening with the variable std. For std the second regime with the least observations, occurred frequently in the beginning of the data set and estimation with 104 observations was enough to identify this regime. The optimal models we find for nr are also MRS models, but now with a regime that has the least observations in it occurring more in the middle of the data set. If we use 104 observations instead of 124, two periods of this regime drop out of the estimation sample and the model cannot be estimated properly all 60 times. That is why we do find MRSvar models in which nr is used to model the regime switches which produce low RMSPE's too, but not significantly lower than those of the benchmarks anymore.

Finally, in each of the four other data sets this statistic seems to have predictive value. For UNEMP and HOUSING we find, as for INPROD, MRSvar models which outperform all the benchmarks in some way, see Table 5.6. The model that performs best for HOUSING is MRSvar-median+nr-c and it is estimated 40 times almost exactly the same. One regime occurs if there are more than approximately 25 forecasters,

has a constant of around 1.8, a lag coefficient of around 0.2 (significant) and a SPF median coefficient of around 1 (significant) and the other regime has a constant of around -2, a lag coefficient that is insignificant and a SPF median coefficient of around 1.7. The same kind of model that performs best for UNEMP, MRSvar-mean+nr-v, has two different sets of estimated parameters. One is similar to those for INPROD and HOUSING, with one minority regime occurring when nr is low. The other has a minority regime occurring when nr is high, above approximately 50.

For UNEMP we also find that MRScons-mean+nr-v performs very well. In this model there is one regime with an insignificant coefficient for nr and one regime with a small but significantly positive coefficient for nr. In this last regime the coefficient for the mean of SPF is close to zero. For NGDP and PGDP, MRScons models with nr forecast accurately, see Table 5.6.

We can conclude that the variable nr is very successful for forecasting. When enough data points are available to estimate the MRSvar models, this type of model can use nr to recognize different regimes in which the lag and location variable should get different coefficients. Some MRScons models are also successful, in which there is one regime where nr significantly predicts the dependent variable to be higher or lower.

5.5 Conclusions

Many studies have shown that the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia contains valuable information for forecasting macroeconomic variables. In our study we have shown that more forecast accuracy can be gained if not just the simple mean or median of SPF forecasts is used, but also if measures of disagreement amongst forecasters is used. All four new SPF statistics proposed, namely standard deviation, 5th percentile, 95th percentile and number of forecasts, showed to contain useful information for forecasting, especially when used in Markov regime-switching models.

One of the interesting findings is that the standard deviation of survey forecasts tends to go up around recessionary periods. Therefore, this statistic signals the beginning of a regime in which different weights should be used for the mean or median of SPF forecasts and the lag of the dependent variable. Also the 5th percentile and 95th percentile predict regime switches, but are also found to have predictive value in a linear way in specific regimes. Finally, the number of forecasts seems to be very useful to predict regimes in which the mean or median should receive a substantially different weight relative to the lag of the dependent variable than in the second regime.

The results found in our basic analysis, using industrial production data and a rolling estimation window and creating 40 forecasts, could be generalized to other situations, but we found a lack of robustness. This may be due to the fact that the MRS model with varying transition probabilities between the regimes, is not easy to estimate, while it turns out that this model is necessary to include the forecasting information contained in the SPF statistics.

Correct estimation of the MRSvar models is easier if the proper starting values are used in the estimation procedure. As no information was available what these starting values could be, we used a grid of starting values. Our results could be used in further research to choose more specific starting values or to use Bayesian techniques.

Furthermore, using a fixed estimation window does not seem such a bad idea in terms of forecasting accuracy according to the INPROD data. It might be a good idea to explore this method further. As only one model needs to be estimated in this case, more effort and time, for example by using a finer grid of starting values for the parameters, could be devoted to find an adequate MRS model.

Another idea could be to use other nonlinear time series models, such as smooth-transition models.

There are many issues unresolved in this paper and more research is needed on this. It is not clear for example if the number of forecasts predicts that a different economic regime is coming up or that it indicates that the mean or median of survey forecasters is less reliable. Furthermore, it might also be interesting to see if changes in

survey forecasts have predictive value. If the average of the forecasts does not change much, but every forecaster changes the forecast substantially compared to previous period, does this say anything about the economic situation or about the accuracy of the survey forecasts? We leave this issues for further research.

Chapter 6

Nederlandse samenvatting (Summary in Dutch)

Dit proefschrift omvat vier opzichzelfstaande hoofdstukken die afzonderlijk gelezen kunnen worden. Alle hoofdstukken gaan over voorspelsituaties waarin econometrische modellen en intuïtie van experts een rol spelen, maar het laatste hoofdstuk is enigszins anders dan de overige drie hoofdstukken. De eerste drie hoofdstukken gaan over wat het is dat de experts doen wanneer ze statistische modelvoorspellingen aanpassen en over wat dat aanpassingsgedrag mogelijk verbetert. Alhoewel de technieken die zijn beschreven en de resultaten die zijn verkregen in deze hoofdstukken toegepast kunnen worden op en gegeneraliseerd kunnen worden naar andere micro- en macro-economische datasets, hebben wij de technieken toegepast op en de resultaten gegenereerd voor verkoopdata. De empirische secties van deze hoofdstukken zijn allemaal gebaseerd op (een deel van) dezelfde unieke dataset, verkregen van een groot farmaceutisch bedrijf met het hoofdkantoor in Nederland en lokale kantoren in verschillende landen.

Het laatste hoofdstuk is gericht op onderzoek naar de manier waarop optimaal gebruik kan worden gemaakt van meerdere voorspellingen die gegenereerd zijn door meerdere experts voor één en dezelfde gebeurtenis. De situatie waarin meerdere van

zulke voorspellingen beschikbaar zijn doet zich meestal voor bij macro-economisch voorspellen en het empirische gedeelte van dit onderzoek is gebaseerd op voorspellingen uit de Survey of Professional Forecasters (SPF) (enquête onder professionele voorspellers), welke vrij verkrijgbaar zijn.

In meer detail omvatten de hoofdstukken het volgende.

Hoofdstuk 2 is gebaseerd op Legerstee et al. (2011). In de situatie zoals geanalyseerd in dit hoofdstuk, hebben experts beschikking over statistische model voorspellingen wanneer ze hun eigen voorspellingen creëren en het is niet gedocumenteerd wat het is dat de experts doen. We focussen in dit hoofdstuk op drie vragen, welke we proberen te beantwoorden aan de hand van beschikbare expertvoorspellingen en modelvoorspellingen. Ten eerste, is de expertvoorspelling gerelateerd aan de modelvoorspelling en hoe? Ten tweede, hoe wordt deze potentiële relatie beïnvloed door andere factoren? Ten derde, hoe beïnvloedt deze relatie voorspelnauwkeurigheid?

Wij introduceren een nieuw en innovatief hiërarchisch Bayes model met twee niveaus om deze vragen te beantwoorden. We passen de door ons geïntroduceerde methodologie toe op de grote dataset met voorspellingen en realisaties van verkoopdata van het farmaceutische bedrijf. Als resultaat vinden we dat expertvoorspellingen op verschillende manieren kunnen afhangen van modelvoorspellingen. Gemiddelde verkoophoeveelheden, volatiliteit in verkoophoeveelheden en de voorspelhorizon beïnvloeden deze afhankelijkheid. We laten ook zien dat theoretische implicaties van expertgedrag op voorspelnauwkeurigheid worden teruggevonden in de empirische data. In het algemeen geldt in onze dataset dat de experts die een paar simpele regels volgen, welke voorspelprestaties optimaliseren onder optimale omstandigheden, beter voorspellen dan het model.

Hoofdstuk 3 is gebaseerd op Legerstee and Franses (2011). Hier analyseren we wederom het gedrag van experts die voorspellingen geven voor maandelijkse verkoopdata, maar nu vergelijken we data van voor en van na het moment dat de experts verschillende soorten feedback hebben ontvangen over hun gedrag. We hebben data van 21 experts die zich in 21 verschillende landen bevinden en die voorspellingen maken

voor een verscheidenheid aan farmaceutische producten voor oktober 2006 tot en met september 2007. We bestuderen het gedrag van de experts door hun voorspellingen te vergelijken met die van een geautomatiseerd statistisch programma en we rapporteren de voorspelnauwkeurigheid over deze 12 maanden. In september 2007 ontvingen deze experts feedback over hun gedrag en kregen ze training op het hoofdkantoor, waarbij uitgebreid aandacht is besteed aan de technische details van het statische programma dat gebruikt wordt voor het creëren van de modelvoorspellingen. Vervolgens bestuderen we het gedrag van de experts in de drie maanden na de trainingssessie, dus in oktober 2007 tot en met december 2007. Onze belangrijkste conclusie is dat de expertvoorspellingen in de tweede periode minder afweken van de statistische modelvoorspellingen en dat hun nauwkeurigheid aanzienlijk verbeterde.

Hoofdstuk 4 is gebaseerd op Franses et al. (2011b). In dit hoofdstuk wordt een nieuwe en simpele methodologie geïntroduceerd om de verliesfuncties gerelateerd aan expertvoorspellingen te schatten. Onder de veronderstelling van conditionele normaliteit van de data en de voorspelverdeling, kan de asymmetrie parameter van de lin-lin en linex verliesfunctie makkelijk geschat worden, gebruik makende van een lineaire regressie. Deze regressie geeft ook een schatting van potentiële systematische scheefheid in de voorspellingen van de experts. De residuen van de regressie zijn de input voor een test op de validiteit van de normaliteitveronderstelling.

We passen onze aanpak wederom toe op de dataset met voorspellingen voor verkoop gemaakt door experts en we vergelijken de uitkomsten met die van statistische, modelgebaseerde voorspellingen voor dezelfde verkoopdata. We vinden substantieel bewijs voor asymmetrie in de verliesfunctie van de experts, met te lage voorspellingen strenger bestraft dan te hoge voorspellingen.

Hoofdstuk 5 is gebaseerd op Legerstee and Franses (2010). Voorspellingen van verschillende experts worden vaak gebruikt in voorspelmodellen en in voorspelcombinaties door het gemiddelde of de mediaan te nemen van de enquêtedata. In dit onderzoek nemen we een ander standpunt in en onderzoeken we de voorspellende waarde van potentiële meningsverschillen tussen voorspellers. De veronderstelling is dat het

niveau van verschil in mening een signaal kan zijn voor toekomstige structurele of tijdelijke veranderingen in een economisch proces of in de voorspellende waarde van de enquêtevoorspellingen.

In ons empirische werk onderzoeken we een verscheidenheid aan macro-economische variabelen en gebruiken we verschillende manieren om de mate van verschil in mening te meten, samen met verschillende maatstaven voor de locatie van de enquêtedata en autoregressieve termen. Voorspellingen van simpele lineaire modellen en voorspellingen van Markov regiemvariërende modellen met constante en met tijdsvariërende transitiekansen zijn geconstrueerd in reële tijd en vergeleken op voorspelnauwkeurigheid. Onze belangrijkste vondst is dat meningsverschil inderdaad voorspellende waarde kan hebben, vooral wanneer het gebruikt wordt in Markov regiemvariërende modellen.

Bibliography

Armstrong, J. (2001a). Combining forecasts. In Armstrong, J., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Acadamic Publishers, Boston, MA.

Armstrong, J. (2001b). *Principles of Forecasting*. Kluwer Academic Publishers, Boston, MA.

Armstrong, J. and Pagell, R. (2003). The ombudsman: Reaping benefits from management research: Lessons from the forecasting principles project. *Interfaces*, 33:91–97.

Athanasopoulos, G. and Hyndman, R. (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting*, 27:845–849.

Awirothananon, T. and Cheung, W. (2009). On joint determination of the number of states and the number of variables in markov-switching models: A monte carlo study. *Communications in Statistics Simulation and Computation*, 38:1757–1788.

Baillon, E. and Cabantous, A. (2009). Combining imprecise or conflicting probability judgments: A choice based study. ICBBR Working Paper Series No 2009_03.

Balzer, W., Sulsky, L., Hammer, L., and Sumner, K. (1992). Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes*, 53:35–54.

Björkman, M. (1972). Feedforward and feedback as determiners of knowledge and policy-notes on a neglected issue. *Scandinavian Journal of Psychology*, 13:152–158.

Blattberg, R. and Hoch, S. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36(8):887–899.

Bolger, F. and Önkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, 20:29–39.

Boulaksil, Y. and Franses, P. (2009). Experts' stated behavior. *Interfaces*, 39(2):168–171.

Bunn, D. and Salo, A. (1996). Adjustment of forecasts with model consistent expectations. *International Journal of Forecasting*, 12(1):163–170.

Capistrán, C. and Timmermann, A. (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, 41(2-3):365–396.

Carroll, C. (2003). The epidemiology of macroeconomic expectations. In Blume, L. and Durlauf, S., editors, *The Economy as an Evolving Complex System, III*. Oxford: Oxford University Press.

Christoffersen, P. and Diebold, F. (1996). Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11(5):561–571.

Christoffersen, P. and Diebold, F. (1997). Optimal prediction under asymmetric loss. *Econometric Theory*, 13(6):808–817.

Clark, T. and McCracken, M. (2001). Test of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105:85–110.

Clatworthy, M., Peel, D., and Pope, P. (forthcoming 2011). Are analysts' loss functions asymmetric? *Journal of Forecasting*.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–583.

Cooke, R. (1991). *Experts in Uncertainty; Opinion and Subjective Probability in Science*. New York: Oxford University Press.

D'Agostino, R. and Stephens, M. (1986). *Goodness-of-fit techniques*. Marcel Dekker, Inc., New York.

Diamantopoulos, A. and Mathews, B. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial and Decision Economics*, 10:51–59.

Diebold, F., Lee, J.-H., and Weinbach, G. (1994). Regime switching with time-varying transition probabilities. In Hargreaves, C., editor, *Nonstationary Time Series Analysis and Cointegration*, , Advanced Texts in Econometrics. Oxford: Oxford University Press.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Dovern, F., Fritsche, U., and Slacalek, J. (2009). Disagreement among forecasters in G7 countries. Europen Central Bank Working Paper Series No 1082.

Einhorn, H. and Hogharth, R. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13:171–192.

Elliott, G., Komunjer, I., and Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *The Review of Economic Studies*, 72(6):1107–1125.

Elliott, G., Komunjer, I., and Timmermann, A. (2008). Biases in macroeconomic forecasts: irrationality or asymmetric loss? *Journal of European Economic Association*, 6(1):122–157.

Elliott, G. and Timmermann, A. (2005). Optimal forecast combination under regime switching. *International Economic Review*, 46(4):1081–1102.

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Fildes, R. and Goodwin, P. (2007a). Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6):570–576.

Fildes, R. and Goodwin, P. (2007b). Good and bad judgement in forecasting: Lessons from four companies. *Foresight: The International Journal of Applied Forecasting*, 8:5–10.

Fildes, R., Goodwin, P., Lawrence, M., and Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25:3–23.

Franses, P., Kranendonk, H., and Lanser, D. (2011a). One model and various experts: Evaluating Dutch macroeconomic forecasts. *International Journal of Forecasting*, 27:482–495.

Franses, P. and Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, 25:35–47.

Franses, P. and Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, 29(3):331–340.

Franses, P. and Legerstee, R. (2011a). Combining SKU-level sales forecasts from models and experts. *Expert Systems with Applications*, 38:2365–2370.

Franses, P. and Legerstee, R. (2011b). Experts' adjustment to model-based SKU-level forecasts: Does the forecast horizon matter? *Journal of the Operational Research Society*, 62(3):537–543.

Franses, P., Legerstee, R., and Paap, R. (2011b). Estimating loss functions of experts. Tinbergen Institute Discussion Paper 2011-177/4.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Gönül, S., Önkal, D., and Goodwin, P. (2009). Expectations, use and judgmental adjustment of external financial and economic forecasts: An empirical investigation. *International Journal of Forecasting*, 28:19–37.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgement. *International Journal of Forecasting*, 16:85–99.

Goodwin, P. and Fildes, R. (1999a). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12(1):37–53.

Goodwin, P. and Fildes, R. (1999b). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12:37–53.

Granger, C. (2001). Overview of nonlinear macroeconometric empirical models. *Macroeconomic Dynamics*, 5:466–481.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. (1994). *Time Series Analysis*, chapter Modeling Time Series with Changes in Regime. Princeton, NJ: Princeton University Press.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediciton mean squared errors. *International Journal of Forecasting*, 13:281–291.

Heij, C., de Boer, P., Franses, P., Kloek, T., and van Dijk, H. (2004). *Econometric Methods with Applications in Business and Economics*, chapter 5.7, pages 396–418. Oxford University Press.

Lahiri, K. and Sheng, X. (2010). Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics*, 25:514–538.

Lamont, O. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization*, 48:265–280.

Laster, D., Bennett, P., and Geoum, I. (1999). Rational bias in macroeconomic forecasts. *The Quarterly Journal of Economics*, 114(1):293–318.

Lawrence, M., Goodwin, P., O'Connor, M., and Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22:493–518.

Legerstee, R. and Franses, P. (2010). Does disagreement amongst forecasters have predictive value? Tinbergen Institute Discussion Paper 2010-088/4.

Legerstee, R. and Franses, P. (2011). Do experts SKU forecasts improve after feedback? Tinbergen Institute Discussion Paper 2011-135/4.

Legerstee, R., Franses, P., and Paap, R. (2011). Do experts incorporate statistical model forecasts and should they? Tinbergen Institute Discussion Paper 2011-141/4.

Lim, K., O'Connor, M., and Remus, W. (2005). The impact of presentation media on decision making: Does multimedia improve the effectiveness of feedback? *Information and Management*, 42:305–316.

Mankiw, G., Reis, R., and Wolfers, J. (2003). Disagreement about inflation expectations. In Gertler, M. and Rogoff, K., editors, *NBER Macroeconomics Annual*. Cambridge: MIT Press.

Mathews, B. and Diamantopoulos, A. (1986). Managerial intervention in forecasting: An empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, 3:3–10.

Mathews, B. and Diamantopoulos, A. (1989). Judgemental revision of sales forecasts: A longitudinal extension. *Journal of Forecasting*, 8:129–140.

Mathews, B. and Diamantopoulos, A. (1990). Judgmental revision of sales forecasts: Effectiveness of forecast selection. *Journal of Forecasting*, 9:407–415.

Mathews, B. and Diamantopoulos, A. (1992). Judgmental revision of sales forecasts: The relative performance of judgementally revised versus non revised forecasts. *Journal of Forecasting*, 11:569–576.

Mathews, B. and Diamantopoulos, A. (1994). Towards a taxonomy of forecast error measures- a factor-comparative investigation of forecast error dimensions. *Journal of Forecasting*, 13:409–416.

McCracken, M. (2000). Robust out-of-sample inference. *Journal of Econometrics*, 99:195–223.

McNees, S. (1990). The role of judgment in macroeconomic forecasting accuracy. *International Journal of Forecasting*, 6:287–299.

Psaradakis, Z. and Spagnolo, N. (2006). Joint determination of the state dimension and autoregressive order for models with markov regime switchin. *Journal of Time Series Analysis*, 27(5):753–766.

Remus, W., O'Connor, M., and Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, 66:22–30.

Sanders, N. (1992). Accuracy of judgemental forecasts: A comparison. *Omega*, 20:353–364.

Sanders, N. (1997). The impact of task properties feedback on time series judgmental forecasting tasks. *Omega*, 25:135–144.

Smith, A., Naik, P., and Tsai, C.-L. (2006). Markov-switching model selection using Kullback-Leibler divergence. *Journal of Econometrics*, 134:553–577.

Stone, E. and Opel, R. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, 83:282–309.

Timmermann, A. (2000). Moments of markov switching models. *Journal of Econometrics*, 96(1):75–111.

Trapero, J., Fildes, R., and Davydenko, A. (2010). Nonlinear identification of judgmental forecasts effects at SKU level. *Journal of Forecasting*, 30(5):490–508.

Turner, D. (1990). The role of judgment in macroeconomic forecasting. *Journal of Forecasting*, 9:315–345.

van Dijk, D. and Franses, P. (2003). Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. *Oxford Bulletin of Economics and Statistics*, 65:727–744.

Varian, H. (1975). A Bayesian approach to real estate assessment. In Fienberg, S. and Zellner, A., editors, *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, pages 195–208. North-Holland, Amsterdam.

Wackerly, D., Mendenhall III, W., and Scheaffer, R. (2002a). *Mathematical Statistics with Applications*, chapter 10.8, pages 489–494. Duxbury Advanced Series, 6th edition.

Wackerly, D., Mendenhall III, W., and Scheaffer, R. (2002b). *Mathematical Statistics with Applications*, chapter 10.3, pages 467–473. Duxbury Advanced Series, 6 edition.

Wackerly, D., Mendenhall III, W., and Scheaffer, R. (2002c). *Mathematical Statistics with Applications*, chapter 10.9, pages 498–505. Duxbury Advanced Series, 6 edition.

Welch, E., Bretschneider, S., and Rohrbaugh, J. (1998). Accuracy of judgmental extrapolation of time series data. characteristics, causes, and remediation strategies for forecasting. *International Journal of Forecasting*, 14:95–110.

Zarnowitz, V. and Braun, P. (1993). Twenty-two years of the NBER-ASA quarterly economic oulook surveys: Aspects and comparisons of forecasting performance. In Stock, J. and Watson, M., editors, *Business Cycles, Indicators, and Forecasting*. University of Chicago Press.

Zarnowitz, V. and Lambros, L. (1987). Consensus and uncertainty in economic prediction. *The Journal of Political Economy*, 95(3):591–621.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

480 H. WU, *Essays on Top Management and corporate behavior*

481 X. LIU, *Three Essays on Real Estate Finance*

482 E.L.W. JONGEN, *Modelling the Impact of Labour Market Policies in the Netherlands*

483 M.J. SMIT, *Agglomeration and Innovations: Evidence from Dutch Microdata*

484 S. VAN BEKKUM, *What is Wrong With Pricing Errors? Essays on Value Price Divergence*

485 X. HU, *Essays on Auctions*

486 A.A. DUBOVIK, *Economic Dances for Two (and Three)*

487 A.M. LIZYAYEV, *Stochastic Dominance in Portfolio Analysis and Asset Pricing*

488 B. SCHWAAB, *Credit Risk and State Space Methods*

489 N. BASTÜRK, *Essays on parameter heterogeneity and model uncertainty*

490 E. GUTIÉRREZ PUIGARNAU, *Labour markets, commuting and company cars*

491 M.W. VORAGE, *The Politics of Entry*

492 A.N. HALSEMA, *Essays on Resource Management: Ownership, Market Structures and Exhaustibility*

493 R.E. VLAHU, *Three Essays on Banking*

494 N.E. VIKANDER, *Essays on Teams and the Social Side of Consumption*

495 E. DEMIREL, *Economic Models for Inland Navigation in the Context of Climate Change*

496 V.A.C. VAN DEN BERG, *Congestion pricing with Heterogeneous travellers*

497 E.R. DE WIT, *Liquidity and Price Discovery in Real Estate Assets*

498 C. LEE, *Psychological Aspects of the Disposition Effect: An Experimental Investigation*

499 M.M. RIDHWAN, *Regional Dimensions of Monetary Policy in Indonesia*

500 J. GARCÍA, *The moral herd: Groups and the Evolution of Altruism and Cooperation*

501 F.H. LAMP, *Essays in Corporate Finance and Accounting*

502 J. SOL, *Incentives and Social Relations in the Workplace*

503 A.I.W. HINDRAYANTO, *Periodic Seasonal Time Series Models with applications to U.S. macroeconomic data*

504 J.J. DE HOOP, *Keeping Kids in School: Cash Transfers and Selective Education in Malawi*

505 O. SOKOLINSKIY, *Essays on Financial Risk: Forecasts and Investor Perceptions*

506 T. KISELEVA, *Structural Analysis of Complex Ecological Economic Optimal Management Problems*

507 U. KILINC, *Essays on Firm Dynamics, Competition and Productivity*

508 M.J.L. DE HEIDE, *R&D, Innovation and the Policy Mix*

509 F. DE VOR, *The Impact and Performance of Industrial Sites: Evidence from the Netherlands*

510 J.A. NON, *Do ut Des: Incentives, Reciprocity, and Organizational Performance*

511 S.J.J. KONIJN, *Empirical Studies on Credit Risk*

512 H. VRIJBURG, *Enhanced Cooperation in Corporate Taxation*

513 P. ZEPPINI, *Behavioural Models of Technological Change*

514 P.H. STEFFENS, *It's Communication, Stupid! Essays on Communication, Reputation and (Committee) Decision-Making*

515 K.C. YU, *Essays on Executive Compensation - Managerial Incentives and Disincentives*

516 P. EXTERKATE, *Of Needles and Haystacks: Novel Techniques for Data-Rich Economic Forecasting*

517 M. TYSZLER, *Political Economics in the Laboratory*

518 Z. WOLF, *Aggregate Productivity Growth under the Microscope*

519 M.K. KIRCHNER, *Fiscal Policy and the Business Cycle - The Impact of Government Expenditures, Public Debt, and Sovereign Risk on Macroeconomic Fluctuations*

520 P.R. KOSTER, *The cost of travel time variability for air and car travelers*

521 Y. ZU, *Essays of nonparametric econometrics of stochastic volatility*

522 B. KAYNAR, *Rare Event Simulation Techniques for Stochastic Design Problems in Markovian Setting*

523 P. JANUS, *Developments in Measuring and Modeling Financial Volatility*

524 F.P.W. SCHILDER, *Essays on the Economics of Housing Subsidies*

525 S.M. MOGHAYER, *Bifurcations of Indifference Points in Discrete Time Optimal Control Problems*

526 C. ÇAKMAKLI, *Exploiting Common Features in Macroeconomic and Financial Data*

527 J. LINDE, *Experimenting with new combinations of old ideas*

528 D. MASSARO, *Bounded rationality and heterogeneous expectations in macroeconomics*

529 J. GILLET, *Groups in Economics*