TI 2011-141/4
Tinbergen Institute Discussion Paper

Do Experts incorporate Statistical Model Forecasts and should they?

Rianne Legerstee
Philip Hans Franses
Richard Paap

Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Do experts incorporate statistical model forecasts and should they?

Rianne Legerstee[1]

Philip Hans Franses

Richard Paap

Econometric Institute - Erasmus University Rotterdam

September 30, 2011

Abstract

Experts can rely on statistical model forecasts when creating their own forecasts. Usually it is not known what experts actually do. In this paper 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 a large data set of forecasts and realizations of SKU-level sales data from a 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.

Keywords: model forecasts; expert forecasts; forecast adjustment; Bayesian analysis; endogeneity

[1]We are very grateful to participants of the 31st Annual International Symposium on Forecasting in Prague on June 26-29, 2011 and to participants of the annual conference of the Netherlands Econometric Study Group in Rotterdam on June 10, 2011 for their helpful comments. Please address correspondence to: Rianne Legerstee, Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, Netherlands, e-mail: legerstee@ese.eur.nl

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 (2007), 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

$$EF = \alpha + \beta MF + I, \tag{1}$$

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 considerably. 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 (1) whereas the second level consists of equations that relate the parameters in (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 (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 4 we describe the models that we develop to test the hypotheses. Section 5 describes the data and the results of our case study. The final section concludes.

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.

2.1 What do experts do with model forecasts?

To refine notation, we define the relation between expert forecasts and model forecasts as

$$EF_{t+h|t} = \alpha + \beta MF_{t+h|t} + I_{t+h|t}, \tag{2}$$

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) 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 (they 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) 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[2], 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) appears when $\beta = 0$. Here, the expert does not 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) 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

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

suggests that β is smaller than 1, or, stated differently, closer to 0. In Franses and Legerstee (2009) the parameters in model (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), 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. Often $\beta \neq 1$ in (2).

b. When $\beta \neq 1$, often $\beta < 1$ in (2).

Hypothesis 2

a. Often $\alpha \neq 0$ in (2).

b. When $\alpha \neq 0$, often $\alpha > 0$ in (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) 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) 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 4 we propose an econometric model with which can put these hypotheses to a test.

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 (1999), 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), 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 MF is often endogenous in (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) 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 4. Before we turn to our econometric model, we first discuss various implications of expert behavior on forecast accuracy.

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

$$ESPE = \mathrm{E}[(R_{t+h} - EF_{t+h|t})^2], \tag{3}$$

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) is substituted in (3) we obtain

$$ESPE = \mathrm{E}[(R_{t+h} - \alpha - \beta MF_{t+h|t})^2] + \mathrm{E}[I_{t+h|t}^2] - 2\mathrm{E}[((R_{t+h} - \beta MF_{t+h|t})I_{t+h|t}], \quad (4)$$

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 $ESPE$.

3.1 Optimal settings

For ease of derivation, at first we assume that MF is exogenous in (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) given by

$$\alpha_{opt} = \mathrm{E}[R_{t+h}] - \beta \mathrm{E}[MF_{t+h|t}]. \quad (5)$$

$\frac{\partial ESPE}{\partial \beta} = 0$ and then substituting it with the optimal value for α in (5) gives the optimal value for β, that is,

$$\beta_{opt} = \frac{\mathrm{E}[MF_{t+h|t}R_{t+h}] - \mathrm{E}[MF_{t+h|t}]\mathrm{E}[R_{t+h}]}{\mathrm{E}[MF_{t+h|t}^2] - \mathrm{E}[MF_{t+h|t}]^2} = \frac{\mathrm{Cov}[MF_{t+h|t}, R_{t+h}]}{\mathrm{V}[MF_{t+h|t}]}, \quad (6)$$

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 $\mathrm{E}[MF_{t+h|t}^2]$, the more β_{opt} differs from 1. However, under the additional condition that $\mathrm{E}[MF_{t+h|t}R_{t+h}] = \mathrm{E}[MF_{t+h|t}^2]$, 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 $\mathrm{E}[MF_{t+h|t}^2]$, which are $\sum MF_{t+h|t}R_{t+h}$ and $\sum MF_{t+h|t}^2$,

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 (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 (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

$$2\text{Cov}[R_{t+h}I_{t+h|t}] > \text{V}[I_{t+h|t}], \tag{7}$$

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). 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 $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] = 0$, 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 $ESPE$:

$$\alpha_{opt} = \mathrm{E}[R_{t+h}] - \beta \mathrm{E}[MF_{t+h|t}], \tag{8}$$

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

$$\begin{aligned} \beta_{opt} &= \frac{\mathrm{E}[MF_{t+h|t}R_{t+h}] - \mathrm{E}[MF_{t+h|t}]\mathrm{E}[R_{t+h}] - \mathrm{E}[MF_{t+h|t}I_{t+h|t}]}{\mathrm{E}[MF_{t+h|t}^2] - \mathrm{E}[MF_{t+h|t}]^2} \\ &= \frac{\mathrm{Cov}[MF_{t+h|t}, R_{t+h}] - \mathrm{Cov}[MF_{t+h|t}I_{t+h|t}]}{\mathrm{V}[MF_{t+h|t}]}, \end{aligned} \tag{9}$$

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

$$\alpha_{opt} = \frac{\mathrm{Cov}[MF_{t+h|t}I_{t+h|t}]}{\mathrm{V}[MF_{t+h|t}]}\mathrm{E}[MF_{t+h|t}], \tag{10}$$

$$\beta_{opt} = 1 - \frac{\mathrm{Cov}[MF_{t+h|t}I_{t+h|t}]}{\mathrm{V}[MF_{t+h|t}]}. \tag{11}$$

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 (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 (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 (4). We observe that adding intuition is only beneficial if

$$2\mathrm{E}[(R_{t+h} - \beta MF_{t+h|t})I_{t+h|t}] > \mathrm{V}[I_{t+h|t}]. \tag{12}$$

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.

3.2 Implications and hypotheses

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

$$\mathrm{E}[R_{t+h}] = \mathrm{E}[MF_{t+h|t}], \tag{13}$$

$$\mathrm{E}[MF_{t+h|t}R_{t+h}] = \mathrm{E}[MF^2_{t+h|t}]. \tag{14}$$

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) that,

a. $\alpha = 0$, $\beta = 1$, and (7) is met for $I_{t+h|t}$, assuming that (13) and (14) are met and that $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] = 0$;

b. α is as in (10), β as in (11), and (12) is met for $I_{t+h|t}$, if (13) and (14) are met, but possibly $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] \neq 0$;

c. α is as in (8), β as in (9), and (12) is met for $I_{t+h|t}$, if (13) and (14) are *not* met and possibly $\mathrm{E}[MF_{t+h|t}I_{t+h|t}] \neq 0$.

Note that (7) and (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 (12) results in higher prediction errors for EF, where the amount of loss of precision depends on the interaction 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 (13) and (14) are met) and exogenous in (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 (13) and (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)

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 (13) and (14) are met, it might be interesting to test Hypothesis 6.

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.

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 A for a more detailed explanation of the data format. Using this notation we can then write (2) as

$$EF_{i,t} - MF_{i,t} = \alpha_i^* + \beta_i^* MF_{i,t} + \varepsilon_{i,t}, \tag{15}$$

with $\varepsilon_{i,t} \sim N(0, \sigma^2_{\varepsilon,i})$. Note that β_i^* in this model associates with $\beta - 1$ in (2) and α_i^* with α in (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

$$\alpha_i^* = \begin{cases} 0 & \text{if } P_i = 1 \\ \alpha_i^\dagger = z_i' \gamma_\alpha + \xi_i & \text{if } P_i = 0, \end{cases} \tag{16}$$

and

$$\beta_i^* = \begin{cases} 0 & \text{if } S_i = 1 \\ \beta_i^\dagger = z_i' \gamma_\beta + \eta_i & \text{if } S_i = 0, \end{cases} \tag{17}$$

with $\xi_i \sim N(0, \sigma^2_\xi)$ and $\eta_i \sim N(0, \sigma^2_\eta)$. 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 2.1. 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:

$$P_i = \begin{cases} 1 & \text{if } q_i = z_i'\psi_\alpha + \nu_i > 0 \\ 0 & \text{if } q_i = z_i'\psi_\alpha + \nu_i \leq 0, \end{cases} \tag{18}$$

and

$$S_i = \begin{cases} 1 & \text{if } w_i = z_i'\psi_\beta + \omega_i > 0 \\ 0 & \text{if } w_i = z_i'\psi_\beta + \omega_i \leq 0, \end{cases} \tag{19}$$

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

$$\kappa_i = \int_0^\infty \phi(q_i; z_i'\psi_\alpha, 1) dq_i, \tag{20}$$

and

$$\lambda_i = \int_0^\infty \phi(w_i; z_i'\psi_\beta, 1) dw_i, \tag{21}$$

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 (16), (17), (20) and (21) constitute the second level of our model.

Sofar we have assumed that the error terms in our basic equation (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 2.3. 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,

$$MF_{i,t} = \mu_i + \delta_i V_{i,t} + \zeta_{i,t}, \tag{22}$$

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 (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

$$EF_{i,t} - MF_{i,t} = \alpha_i^* + \beta_i^* MF_{i,t} + \varepsilon_{i,t}, \tag{23}$$

$$MF_{i,t} = \mu_i + \delta_i V_{i,t} + \zeta_{i,t}, \tag{24}$$

$$\alpha_i^* = \begin{cases} 0 & \text{if } P_i = 1 \\ \alpha_i^\dagger = z_i'\gamma_\alpha + \xi_i & \text{if } P_i = 0 \end{cases} \tag{25}$$

$$\beta_i^* = \begin{cases} 0 & \text{if } S_i = 1 \\ \beta_i^\dagger = z_i'\gamma_\beta + \eta_i & \text{if } S_i = 0, \end{cases} \tag{26}$$

$$P_i = \begin{cases} 1 & \text{if } q_i = z_i'\psi_\alpha + \nu_i > 0 \\ 0 & \text{if } q_i = z_i'\psi_\alpha + \nu_i \leq 0, \end{cases} \tag{27}$$

$$S_i = \begin{cases} 1 & \text{if } w_i = z_i'\psi_\beta + \omega_i > 0 \\ 0 & \text{if } w_i = z_i'\psi_\beta + \omega_i \leq 0. \end{cases} \tag{28}$$

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 B. 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.

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

$$DRMSPE_i = \sqrt{\frac{1}{T_i}\sum (R_{i,t} - MF_{i,t})^2} - \sqrt{\frac{1}{T_i}\sum (R_{i,t} - EF_{i,t})^2}. \qquad (29)$$

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

$$DRMSPE_i = r_i'\vartheta + \iota_i, \qquad (30)$$

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 (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) 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 (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 (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 (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 (30).

Concluding, we have for (30) the set of six explanatory variables

$$r_i' = [1,\ P_i,\ |\alpha_i^\dagger(1-P_i)|,\ S_i,\ |\beta_i^\dagger(1-S_i)|,\ \rho_{\Omega,i},\ \rho_{\varepsilon R,i}]. \tag{31}$$

See Table 1 for an overview of the variables in r_i, the names of the variables and for the hypothetical sign of the parameters in (30) following Hypothesis 6.

Table 1: A summary of the variables in r_i in model (30) and their hypothetical effect on $DRMSPE$ as denoted in (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	$\rho_{\varepsilon R,i}$	$+$

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 5.2 and 5.3 the results of our analysis are discussed.

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 maxi-

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

As an instrumental variable in (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.[3] So, as instrumental variable we use the most recent forecast 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 (16), (17), (20) and (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

[3] We use the same notation as in Appendix A. 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 $t - 1$.

a product.

The optimal values for α and β depend on conditions (13) and (14) as defined in Section 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 (13)) and if there is a significant difference between the mean of MF times R and the mean of MF^2 (condition (14)). For this, we used the common small-sample test for comparing two population means as described in Wackerly et al. (2002). We find that condition (13) is rejected in about 17% of the cases and condition (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 (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 (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 (13) holds in about 83% of the cases and condition (14) holds in about 93% of the cases.

5.2 Expert Behavior

To estimate the parameters of the model described in Section 4.1 we generate 80,000 iterations of the Gibbs sampler as described in Appendix B. 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.

The probability that $\beta^* = \beta - 1 = 0$ is varying, which can be seen from the histogram in Figure 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) is likely to be close to 1.

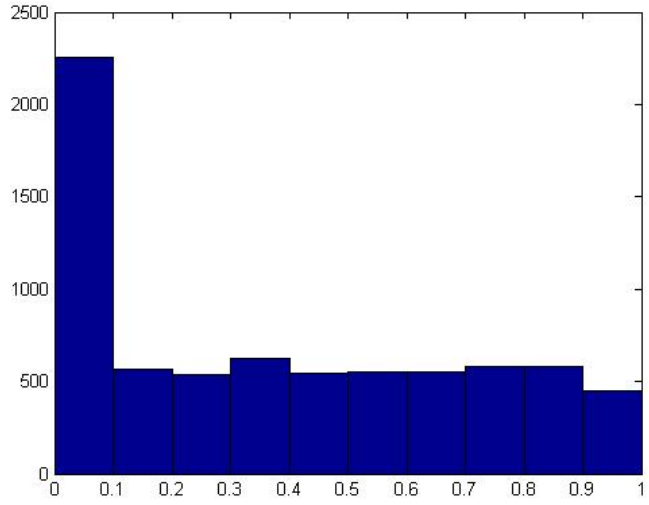

Figure 1: Histogram of posterior means for S_i in (19), for $i = 1, ..., N$.

Figure 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.

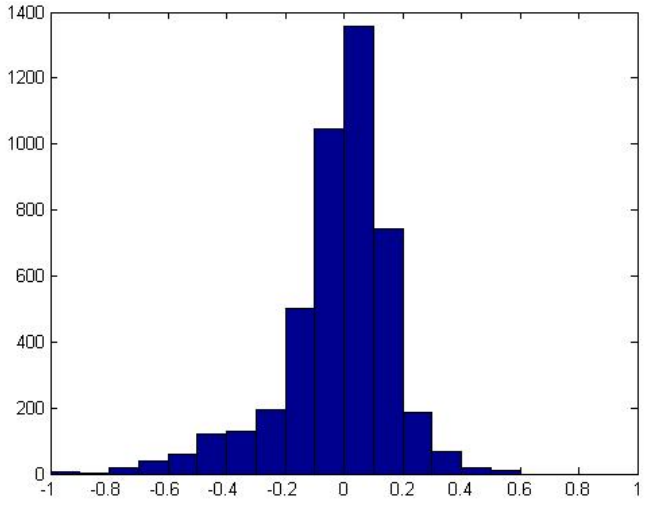

Figure 2: Histogram of posterior means for $\beta_i^\dagger$ in (17) for which the posterior mean for $S_i < 0.5$, the posterior mean for $\beta_i^\dagger > -1$ and the posterior mean for $\beta_i^\dagger < 1$, for $i = 1, ..., N$.

Figure 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.

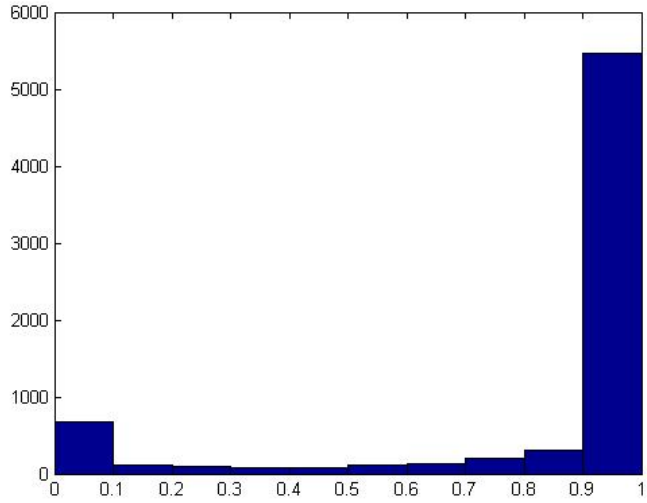

Figure 3: Histogram of posterior means for P_i in (18), for $i = 1, ..., N$.

Figure 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 research (see Section 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.

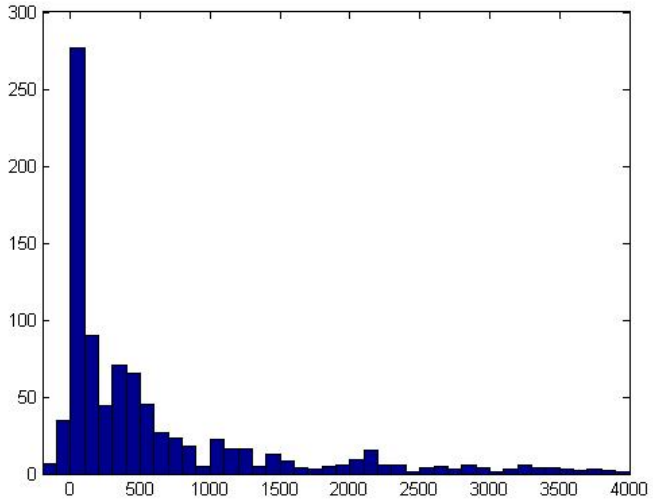

Figure 4: Histogram of posterior means for $\alpha_i^\dagger$ in (16) for which the posterior mean for $P_i < 0.5$ and the posterior mean for $-200 < \alpha_i^\dagger < 4000$, for $i = 1, ..., N$.

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 3 we derived that the optimal value of β_i is given by this fraction in (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 (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.

Table 2: Posterior means (and standard deviations) for the parameters in the second level of the model about expert behavior, described in Subsection 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)

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. 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). 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.

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[4]. 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). 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)). With Hypothesis 6 we stated that a lower or more negative correlation between MF and I in (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 (22) is significantly different from 0, so we can conclude that we used a fairly relevant instrument.

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

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

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.

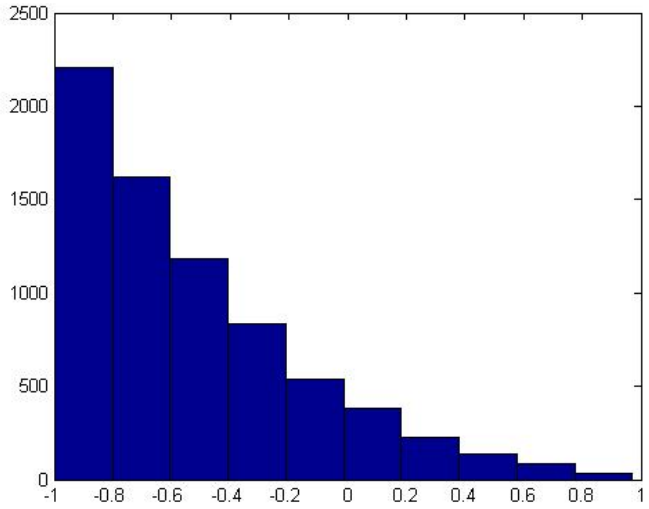

Figure 5: Histogram of posterior means for $\rho_{\Omega,i}$, correlation between $\varepsilon_{i,t}$ in (23) and $\zeta_{i,t}$ in (24), for $i = 1, ..., N$.

The endogeneity in (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 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.

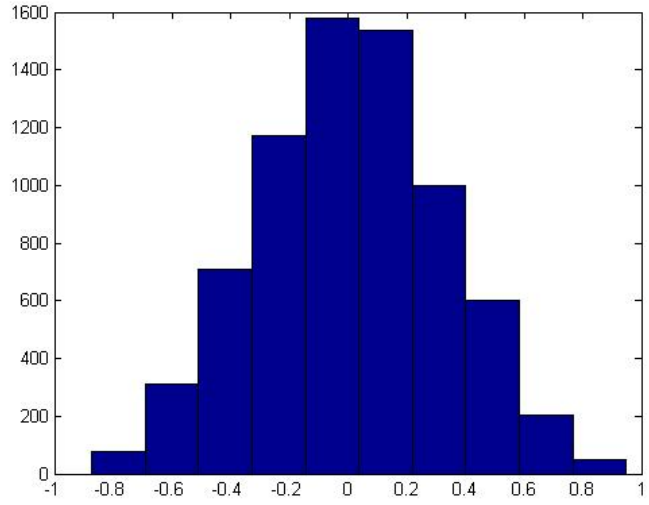

Figure 6: Histogram of the correlations between realized sales $R_{i,t}$ and the posterior mean for $\varepsilon_{i,t}$ from (15), for $i = 1, ..., N$.

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 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 (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.

5.3 Forecast Evaluation

In Table 3 we give the estimated coefficients of model (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) and the model forecasts and between the residuals in (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 3: Estimated coefficients of the forecast evaluation model (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 fore-

casts 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). 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 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 realized 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 (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.

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 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 relative 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 endo-

geneity 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.

References

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

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*, 12:163–170.

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.

Fildes, R. and Goodwin, P. (2007). 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. 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.

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

Goodwin, P. and Fildes, R. (1999). 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.

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.

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.

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

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

Trapero, J., Fildes, R., and Davydenko, A. (2010). Nonlinear identification of judgmental forecasts effects at sku level. *Journal of Forecasting*, online publication.

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

Wackerly, D., Mendenhall III, W., and Scheaffer, R. (2002). *Mathematical Statistics with Applications*, chapter 10.8, pages 467–473. Duxbury Advanced Series, 6th edition.

Appendices

A Typical data format

In this appendix we describe the data format as assumed in Section 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 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 A.1.

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 A.1. 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 A.1: 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	$t=1$	$t=2$	$\dots$	$t=T-1$	$t=T$
1	1	1	1	$X_{1,1}$	$X_{1,2}$	$\dots$	$X_{1,T-1}$	$X_{1,T}$
1	1	2	2	$X_{2,1}$	$X_{2,2}$	$\dots$	$X_{2,T-1}$	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	1	H_1	H_1	$X_{H_1,1}$	$X_{H_1,2}$	$\dots$	NA	NA
1	2	1	H_1+1	$X_{H_1+1,1}$	$X_{H_1+1,2}$	$\dots$	$X_{H_1+1,T-1}$	$X_{H_1+1,T}$
1	2	2	H_1+2	$X_{H_1+2,1}$	$X_{H_1+2,2}$	$\dots$	$X_{H_1+2,T-1}$	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	2	H_2	H_1+H_2	$X_{H_1+H_2,1}$	$X_{H_1+H_2,2}$	$\dots$	NA	NA
1	3	1	H_2+1	$X_{H_2+1,1}$	$X_{H_2+1,2}$	$\dots$	$X_{H_2+1,T-1}$	$X_{H_2+1,T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	J_1	H_{J_1}	$\sum_{j=1}^{J_1} H_j$	$X_{\sum_{j=1}^{J_1} H_j,1}$	$X_{\sum_{j=1}^{J_1} H_j,2}$	$\dots$	NA	NA
2	J_1+1	1	$\sum_{j=1}^{J_1} H_j+1$	$X_{\sum_{j=1}^{J_1} H_j+1,1}$	$X_{\sum_{j=1}^{J_1} H_j+1,2}$	$\dots$	$X_{\sum_{j=1}^{J_1} H_j+1,T-1}$	$X_{\sum_{j=1}^{J_1} H_j+1,T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
M	J	H_J	N	$X_{N,1}$	$X_{N,2}$	$\dots$	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 A.1 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 (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 5.1.

B Parameter estimation

In this appendix we describe the method used to estimate the parameters of the two-level Hierarchical Bayes model described in Section 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:

$$f_{kl,i} = f(y_i, x_i | P_i = k, S_i = l, \theta) = \prod_{t=1}^{T_i} \Phi(y_{i,t}, x_{i,t} | m_{kl,i,t}, \Omega_i) \tag{32}$$

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

$$\begin{aligned} m_{11,i,t} &= (x_{i,t}, \mu_i + \delta_i v_{i,t})' \\ m_{01,i,t} &= (\alpha_i^\dagger + x_{i,t}, \mu_i + \delta_i v_{i,t})' \\ m_{10,i,t} &= (x_{i,t} + \beta_i^\dagger x_{i,t}, \mu_i + \delta_i v_{i,t})' \\ m_{00,i,t} &= (\alpha_i^\dagger + x_{i,t} + \beta_i^\dagger x_{i,t}, \mu_i + \delta_i v_{i,t})'. \end{aligned} \tag{33}$$

The complete data likelihood is then

$$\begin{aligned} f(\{y_i\}_{i=1}^N, \{x_i\}_{i=1}^N, \{P_i\}_{i=1}^N, \{S_i\}_{i=1}^N | \theta) = \prod_{i=1}^N & (f_{11,i}\kappa_i\lambda_i)^{P_i S_i} \\ & (f_{01,i}(1-\kappa_i)\lambda_i)^{(1-P_i)S_i} (f_{10,i}\kappa_i(1-\lambda_i))^{P_i(1-S_i)} \\ & (f_{00,i}(1-\kappa_i)(1-\lambda_i))^{(1-P_i)(1-S_i)} \phi(\alpha_i | z_i'\gamma_\alpha, \sigma_\xi^2) \phi(\beta_i | z_i'\gamma_\beta, \sigma_\eta^2), \end{aligned} \tag{34}$$

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.

B.1 Sampling of P_i and S_i

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

$$Pr[P_i = 1|\theta, \text{data}] = \frac{\kappa_i(f_{11,i} + f_{10,i})}{\kappa_i(f_{11,i} + f_{10,i}) + (1 - \kappa_i)(f_{01,i} + f_{00,i})}, \tag{35}$$

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

$$Pr[S_i = 1|\theta, \text{data}] = \frac{\lambda_i(f_{11,i} + f_{01,i})}{\lambda_i(f_{11,i} + f_{01,i}) + (1 - \lambda_i)(f_{10,i} + f_{00,i})}. \tag{36}$$

B.2 Sampling of q_i and w_i

The full conditional posterior distribution of q_i is

$$q_i|\theta, \text{data} \sim \begin{cases} N(z_i'\psi_\alpha, 1)I[q_i > 0] & \text{if } P_i = 1 \\ N(z_i'\psi_\alpha, 1)I[q_i \leq 0] & \text{if } P_i = 0, \end{cases} \tag{37}$$

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.

B.3 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 $(\sum_{i=1}^N z_i z_i' + pr_{\psi_\beta}^{-1})^{-1}$.

B.4 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,

$$x_{i,t} = \mu_i + \delta_i v_{i,t} + \rho(y_{i,t} - \alpha_i^\dagger(1 - P_i) - x_{i,t} - \beta_i^\dagger x_{i,t}(1 - S_i)) + e_{i,t}, \tag{38}$$

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 $\tilde{y}_{i,t} = x_{i,t} - \rho(y_{i,t} - \alpha_i^\dagger(1 - P_i) - x_{i,t} - \beta_i^\dagger x_{i,t}(1 - S_i))$.

B.5 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 $\zeta_{i,t} = x_{i,t} - \mu_i - \delta_i v_{i,t}$.

B.6 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 $\sigma_\eta^2(\sum_{i=1}^{N} z_i z_i'(1-S_i))^{-1}$.

B.7 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

$$\frac{\sum_{t=1}^{N}\xi_i^2(1-P_i) + pr_{\sigma_\xi^2,sc}}{\sigma_\xi^2} \sim \chi^2\left(\sum_{i=1}^{N}(1-P_i) + pr_{\sigma_\xi^2,sh}\right). \tag{39}$$

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,

$$\frac{\sum_{i=1}^{N}\eta_i^2(1-S_i) + pr_{\sigma_\eta^2,sc}}{\sigma_\eta^2} \sim \chi^2\left(\sum_{i=1}^{N}(1-S_i) + pr_{\sigma_\eta^2,sh}\right), \tag{40}$$

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

B.8 Sampling of $\alpha_i^{\dagger}$

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

$$y_{i,t} = \alpha_i^{\dagger} + x_{i,t} + \beta_i^{\dagger}x_{i,t}(1-S_i) + \rho(x_{i,t} - \mu_i - \delta_i v_{i,t}) + e_{i,t}, \tag{41}$$

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 $\forall P_i = 0$,

$$\begin{aligned}\sigma_{e,i}^{-1}(y_{i,t} - x_{i,t} - \beta_i x_{i,t}(1-S_i) - \rho(x_{i,t} - \mu_i - \delta_i v_{i,t})) &= \sigma_{e,i}^{-1}\alpha_i^{\dagger} + \sigma_{e,i}^{-1}e_{i,t} \\ \sigma_\xi^{-1}z_i'\gamma_\alpha &= \sigma_\xi^{-1}\alpha_i^{\dagger} + \sigma_\xi^{-1}\xi_i \end{aligned} \tag{42}$$

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

$$B = A\alpha_i^\dagger + d, \tag{43}$$

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

$$\begin{gathered} B = (\sigma_{e,i}^{-1}(y_{i,1} - x_{i,1} - \beta_i x_{i,1}(1 - S_i) - \rho(x_{i,1} - \mu_i - \delta_i v_{i,1}), \\ \sigma_{e,i}^{-1}(y_{i,2} - x_{i,2} - \beta_i x_{i,2}(1 - S_i) - \rho(x_{i,2} - \mu_i - \delta_i v_{i,2}), ..., \\ \sigma_{e,i}^{-1}(y_{i,T_i} - x_{i,T_i} - \beta_i x_{i,T_i}(1 - S_i) - \rho(x_{i,T_i} - \mu_i - \delta_i v_{i,T_i}), \\ \sigma_\xi^{-1} z_i' \gamma_\alpha)' \\ A = (\sigma_{e,i}^{-1}, \sigma_{e,i}^{-1}, ..., \sigma_{e,i}^{-1}, \sigma_\xi^{-1})'. \end{gathered} \tag{44}$$

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.

B.9 Sampling of $\beta_i^\dagger$

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

$$y_{i,t} = \alpha_i^\dagger(1 - P_i) + x_{i,t} + \beta_i^\dagger x_{i,t} + \rho(x_{i,t} - \mu_i - \delta_i v_{i,t}) + e_{i,t}, \tag{45}$$

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 $\forall S_i = 0$,

$$\begin{gathered} \sigma_{e,i}^{-1}(y_{i,t} - \alpha_i^\dagger(1 - P_i) - x_{i,t} - \rho(x_{i,t} - \mu_i - \delta_i v_{i,t})) = \beta_i^\dagger(\sigma_{e,i}^{-1} x_{i,t}) + \sigma_{e,i}^{-1} e_{i,t} \\ \sigma_\eta^{-1} z_i' \gamma_\beta = \sigma_\eta^{-1}\beta_i^\dagger + \sigma_\eta^{-1}\eta_i, \end{gathered} \tag{46}$$

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.