Franses Philip Hans (Orcid ID: 0000-0002-2364-7777)

Evaluating heterogeneous forecasts for vintages of macroeconomic variables

Philip Hans Franses
Max Welz
Econometric Institute
Erasmus School of Economics

Abstract

There are various reasons why professional forecasters may disagree in their quotes for macroeconomic variables. One reason is that they target at different vintages of the data. We propose a novel method to test forecast bias in case of such unobserved heterogeneity. The method is based on so-called Symbolic Regression, where the variables of interest become interval variables. We associate the interval containing the vintages of data with the intervals of the forecasts. An illustration to 18 years of forecasts for annual USA real GDP growth, given by the Consensus Economics forecasters, shows the relevance of the method.

Key words: Forecast bias; Data revisions; Interval data; Symbolic regression
JEL Code: C53

We thank two anonymous reviewers for their detailed and helpful comments.

Correspondence: Econometric Institute, Erasmus School of Economics, POB 1738, NL-3000 DR Rotterdam, the Netherlands, phone: +31104081273, email: franses@ese.eur.nl

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/for.2835

Introduction and motivation

This paper is about the well-known Mincer Zarnowitz (1969) (MZ) auxiliary regression, which is often used to examine (the absence of) bias in forecasts[1]. This regression, in general terms, reads as

$$Realization = \beta_0 + \beta_1 Forecast + \varepsilon$$

Usually, the statistical test of interest concerns, $\beta_0 = 0$ and $\beta_1 = 1$, jointly.

The setting in this paper concerns macroeconomic variables. For many such variables it holds that these experience revisions. For variables like real Gross Domestic Growth (GDP), after the first release, there can be at least five revisions for various OECD countries[2].

The second feature of our setting is that forecasts are often created by a range of professional forecasters. In the present paper for example we will consider the forecasters collected in Consensus Economics[3]. To evaluate the quality of the forecasts from these forecasters, one often takes the average quote (the consensus) or the median quote, and sometimes also measures of dispersion like the standard deviation or the variance are considered. The latter measures give an indication to what extent the forecasters disagree. Recent relevant studies are Capistran and Timmermann (2009), Dovern, Fritsche, and Slacalek (2012), Lahiri and Sheng (2010), Laster, Bennett, and Geoum (1999), and Legerstee and Franses (2015). Reasons for disagreement could be heterogeneity across forecasters caused by their differing

[1] Bias in forecasts can come from including inappropriate information in the creating of the forecasts. Professional forecasters may rely on econometric models with a range of potentially relevant variables, but the forecasters may also decide not to incorporate econometric models at all and base their forecasts on intuition, or they may decide to manually adjust econometric model forecasts. The results summarized in Franses (2014) shows that such manual adjustment or fully ignoring an econometric model can lead to substantial bias in forecasts.

[2] http://www.oecd.org/sdd/na/revisions-of-quarterly-gdp-in-selected-oecd-countries.htm

[3] http://www.consensuseconomics.com/. Other professional forecasters' quotes can be found in the Survey of Professional Forecasters: https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/data-files/routput

reactions to news or noise, see Patton and Timmermann (2007), Engelberg, Manski and Williams (2009), and Clements (2010), and also information rigidities, see Coibion and Gorodnischenko (2012).
Recently, Clements (2019) suggested that there might be another reason why forecasters disagree, and that is, that they may target at different vintages of the macroeconomic data. Some may be concerned with the first (flash) quote, while others may have the final (say, after 5 years) value in mind. The problem however is that the analyst does not know who is doing what.

It is not easy to learn from the actual forecasts how they were created, nor is it easy to learn how forecast revisions are created. Clements (2019) proposes a few assumptions, and with these, he documents for a few variables that data revisions can be predictable. Aruoba (2008) also documents that sometimes data revisions can be viewed as noise, meaning that they can be predicted.

The question then becomes how one should deal with the MZ regression. Of course, one can run the regression for each vintage on the mean of the forecasts. But then still, without knowing who is targeting what, it shall be difficult to interpret the estimated parameters in the MZ regression. At the same time, why should one want to reduce or remove heterogeneity by only looking at the mean? It could be that the range from the vintages widens, but it could also be otherwise. We do not assume that the target of the forecasters interacts with the range from the vintages.

To alleviate these issues, in this paper we propose to keep intact the heterogeneity of the realized values of the macroeconomic variables as well as the unknown heterogeneity across the quotes of the professional forecasters. Our proposal relies on the notion to move away from scalar measurements to interval measurements. Such data are typically called symbolic data, see for example Bertrand and Goupil (1999) and Billard and Diday (2007). The MZ regression for such symbolic data thus becomes a so-called Symbolic Regression.

The outline of our paper is as follows. In the next section we provide more details about the setting of interest. For ease of reading, we will regularly refer to our illustration for annual USA real growth rates, but the material in this section can be translated to a much wider range of applications. The following section deals with the estimation methodology for the

Symbolic Regression. We will also run various simulation experiments to examine the reliability of the methods. Next, we will apply the novel MZ Symbolic Regression to the USA growth rates data and compare the outcomes with what one would have obtained if specific vintages were considered. It appears that the Symbolic MZ Regression is informative. The final section deals with a conclusion, limitations, and further research issues.

Replication files are made available at https://github.com/mwelz/symbreg.

Setting

Consider the I vintages of data for a macroeconomic variable y_t^i, where $i = 1,2,..,I$ and $t = 1,2,\dots,T$. In our illustration below we will have $I = 7$ and $t = 1996, 1997, \dots., 2013$, so $T = 18$. The sample ends in 2013 to be able to collect the seven vintages of data.

Professional forecasters, like the ones united in Consensus Economics forecasts, give quotes during the months m, where $m = 1,2,\dots,M$. For the Consensus Economics forecasters $M = 24$, and the months span January in year t-1, February in year t-1, …, December in year t-1, January in year t, until and including December in year t. An example of the data appears in Table 1, where the quotes are presented for May 13, 2013, for the years 2013 and 2014.

The forecasts can be denoted as

$$\hat{y}_{j,t|m} \text{ with } j = 1,2,\dots,J_{t,m}$$

The number of forecasters can change per month and per forecast target, hence we write $J_{t,m}$. In Table 1 this number is 29. For 2013, and in our notation, Table 1 considers $J_{2013,5}$ and for 2014 it is $J_{2014,17}$.

A key issue to bear in mind for later, and as indicated in the previous section, is that we do *not* observe

$$\hat{y}_{j,t|m}^{i} \text{ with } j = 1,2,\dots,J_{t,m} \ ,$$

that is, we do not know who of the forecasters is targeting which vintages of the data.
To run a Mincer Zarnowitz (MZ) regression, the forecasts per month are usually summarized by taking the median, by using a variance measure, or by the mean ("the consensus"), that is, by considering

$$\hat{y}_{t,m} = \frac{1}{J_{t,m}} \sum_{j=1}^{J_{t,m}} \hat{y}_{j,t|m}$$

The MZ regression then considered in practice is

$$y_t^i = \beta_0 + \beta_1 \hat{y}_{t,m} + \varepsilon_t$$

for $t = 1,2,\dots,T$, and this regression can be run for each $m = 1,2,\dots,M$. Under the usual assumptions, parameter estimation can be done by Ordinary Least Squares. Next, one computes the Wald test for the joint null hypothesis $\beta_0 = 0, \beta_1 = 1$.

Now, one can run this MZ test for each vintage of the data, but then still it is unknown what the estimated parameters in the MZ regression actually reflect. Therefore, we propose an alternative approach. We propose to consider, for $t = 1,2,\dots,T$, the interval

$$(\min_i y_t^i;\ \max_i y_t^i)$$

as the dependent variable, instead of y_t^i, and to consider

$$(\min_j \hat{y}_{j,t|m};\ \max_j \hat{y}_{j,t|m})$$

as the explanatory variable, instead of $\hat{y}_{t,m}$. These two new variables are intervals, and often they are called symbolic variables. The MZ regression thus also becomes a so-called Symbolic Regression, see Bertrand and Goupil (1999), Billard and Diday (2000, 2003, 2007).

Table 2 presents an exemplary dataset for May in year *t*, so *m* = 17. Figure 1 visualizes the same data in a scatter diagram. Clearly, instead of points in the simple regression case, the data can now be represented as rectangles.

How does Symbolic Regression work?

When we denote the dependent variable for short as y and the dependent variable as x, we can compute for the Symbolic MZ Regression

$$\hat{\beta}_1 = \frac{Covariance(y,x)}{Variance(x)}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

thereby drawing on the familiar OLS formulae.

Under the assumption that the data are uniformly distributed in the intervals[4], Billard and Diday (2000) derive the following results. At first, the averages are

$$\bar{y} = \frac{1}{2T} \sum_t (\max_i y_t^i + \min_i y_t^i)$$

and

[4] Even when there are clusters of forecasters who target at specific vintages, the data can be uniformly distributed. Or at least, it shall be hard to reject such a uniform distribution in practice. An interesting area for further research is the potentially plausible occurrence of outlying observations. That is, all forecasters behave similarly, and just one forecaster takes a position at far other end of the spectrum. For the data that we consider in this paper we do not observe such behavior, but for other variables this may occur.

$$\bar{x} = \frac{1}{2T}\sum_{t}(\max_{j}\hat{y}_{j,t|m} + \min_{j}\hat{y}_{j,t|m})$$

The covariance is computed as

$$Covariance(y,x) = \frac{1}{4T}\sum_{t}(\max_{i} y_t^i + \min_{i} y_t^i)\,(\max_{j}\hat{y}_{j,t|m} + \min_{j}\hat{y}_{j,t|m}) - \frac{1}{4T^2}\left[\sum_{t}(\max_{i} y_t^i + \min_{i} y_t^i)\right]\left[\sum_{t}(\max_{j}\hat{y}_{j,t|m} + \min_{j}\hat{y}_{j,t|m})\right]$$

Finally, the variance is computed as

$$Variance(x) = \frac{1}{4T}\sum_{t}(\max_{j}\hat{y}_{j,t|m} + \min_{j}\hat{y}_{j,t|m})^2 - \frac{1}{4T^2}\left[\sum_{t}(\max_{j}\hat{y}_{j,t|m} + \min_{j}\hat{y}_{j,t|m})\right]^2$$

This expression completes the relevant components to estimate the parameters.

Standard errors

To compute standard errors around the thus obtained parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we resort to the bootstrap. By collecting *T* random draws of pairs of intervals, with replacement, and by repeating this B times, we compute the bootstrapped standard errors. Together, they are used to compute the joint Wald test for the null hypothesis that $\beta_0 = 0, \beta_1 = 1$.

Simulations

To learn how Symbolic Regression and the bootstrapping of standard errors works, we run some simulation experiments. To save notation, we take as the Data Generating Process (DGP)

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

for $i = 1,2, \ldots, N$. We set $x_i \sim N(0,1)$ and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Next, we translate the thus generated y_i and x_i to intervals by creating

$$(y_i - |z_{1,i}|;\ y_i + |z_{2,i}|)$$
$$(x_i - |w_{1,i}|;\ x_i + |w_{2,i}|)$$

where

$$z_{j,i} \sim N(0, \sigma_z^2),\ \ j = 1,2$$
$$w_{j,i} \sim N(0, \sigma_w^2),\ \ j = 1,2$$

We set the number of simulation runs at 1000, and the number of bootstrap runs at B = 2000 (as suggested to be a reasonable number in Efron and Tibshirani, 1993). Experimentation with larger values of B did not show markedly different outcomes. The code is written in Matlab and R. We set *N* at 20 and 100, while $\alpha = 0$ or 5, and $\beta = -2, \text{or } 0, \text{or } 2$. The results are in Tables 3 to 6.

Table 3 shows that when we compare the cases where $\sigma_w^2 = 0.5$ versus $\sigma_w^2 = 2.0$ that a larger interval of the explanatory variable creates more bias than a larger interval for the dependent variable (compare $\sigma_z^2 = 0.5$ versus $\sigma_z^2 = 2.0$). Also, the bootstrapped standard errors get larger when the intervals of the data get wider, as expected.

Table 4 is the same as Table 3, but now $\sigma_\varepsilon^2 = 0.5$ is replaced by $\sigma_\varepsilon^2 = 2.0$. Overall this means that $\hat{\beta}$ deviates more from β when the variance σ_ε^2 increases. The differences across the deviations of $\hat{\alpha}$ versus α are relatively small.

Table 5 is the same as Table 3, but now $N = 20$ is replaced by $N = 100$. Clearly, a larger sample size entails less bias in the estimates, and also much smaller bootstrapped standard errors. But still, we see that $\hat{\alpha}$ is closer to α then is $\hat{\beta}$ to β.

Table 6 is similar to Table 4, but now for $N = 100$. A larger sample can offset the effects of increased variance σ_ε^2, as the standard errors are reasonably small.

In Table 7 we report on the simulations when we assume that there is autocorrelation in the forecast revisions. We now consider

$$\left(x_i - \rho x_{i-1} - |w_{1,i}|;\ x_i - \rho x_{i-1} + |w_{2,i}|\right),$$

with the convention $x_0 = 0$. We set the number of simulations runs again at 1000, and the number of bootstraps runs at B = 2000 (as suggested to be a reasonable number in Efron and Tibshirani, 1993). We set *N* at 100, while $\alpha = 0$, and $\beta = -2, \text{or } 0, \text{or } 2$, and we choose $\rho = 0.2\ or\ 0.5$. The results in Table 7 show that the method performs well, also when there is autocorrelation in the revisions.

Analysis of forecasts

We now turn to an illustration of the Symbolic MZ regression. We choose to consider the forecasts for annual growth rates of real GDP in the USA, for the years 1996 to and including 2013. This makes $T = 18$. Our data source[5] gives annualized growth rates per quarter[6]. As there are no vintages of true annual growth data available, we decide to further consider the averages of each time these four quarterly growth rates. The data intervals are presented in Table 2. The right-hand side columns of Table 2 concern the forecasts created in May of year *t*, which means the case where $m = 17$. This implies that we can consider 24 Symbolic MZ regressions, each for each of the 24 months.

Table 8 presents the estimation results, the bootstrapped standard errors and the p value of the Wald test for the null hypothesis that $\beta_0 = 0, \beta_1 = 1$. We see from the last column that a p value > 0.05 appears for the forecasts quoted in May in year *t*-1, and that after that the p value stays in excess of 0.05. However, if we look at the individual parameter estimates, we see that $\beta_1 = 0$ is with the 95% confidence interval until September, year *t*-1. So, Table 7 basically tells us that unbiased forecasts seem to appear from October, year *t*-1 onwards.

[5] http://www.oecd.org/sdd/na/revisions-of-quarterly-gdp-in-selected-oecd-countries.htm
[6] The relevant data in ALFRED go back to 2001, and this would make our sample even shorter, and hence we do not consider these data.

Let us now turn to the MZ regression in its standard format, that is, the explanatory variable is the mean of the forecasts and the variable to be explained in one of the vintages of the data. Table 9 presents the results for the first (flash) release real GDP annual growth rates, whereas Table 10 presents the results for the currently available vintage. We also have the results of all vintages in between, but these do not add much to the conclusions that can be drawn from Tables 9 and 10.

First, we see that the standard errors in Tables 9 and 10 are much smaller than the bootstrapped standard errors for the Symbolic MZ Regression. This of course does not come as a surprise as we have point data instead of intervals. For the first vintage of data in Table 9, we see from the p values for the Wald test in the last column that only since March, year *t*, the null hypothesis of no bias cannot be rejected (p value is 0.485). One month earlier, the p value is 0.071, but for that month we see that $\beta_1 = 1$ is not in 95% confidence interval (0.787 with a SE of 0.098). Note by the way that the forecasts created in the very last month of the current year (December, year *t*) are biased (p value of 0.012), at least for the first release data.

Table 10 delivers quite intriguing results for the forecasts concerning the most recent vintage of data. The p value of the Wald test becomes > 0.05 (that is, 0.083) for the quotes in May, year *t*, but note that $\beta_1 = 1$ is not in 95% confidence interval for 23 of the 24 months. Only for the forecasts in December, year *t*, the forecasts do not seem biased (p value of 0.115, and $\beta_1 = 1$ is in the 95% confidence interval (0.820 with SE of 0.088).

In sum, it seems that individual MZ regressions for vintages of data deliver confusing outcomes, which seem hard to interpret. Let alone that we effectively do not know who of the forecasters is targeting at which vintage. Moreover, it seems that outcomes of the Symbolic MZ Regression are much more coherent and straightforward to interpret. Of course, due to the very nature of the data, that is, intervals versus points, statistical precision in the Symbolic Regression is smaller, but the results seem to have much more face value and interpretability than the standard MZ regressions.

The power of our approach of course suffers from the notion that we look at annual data. We do not think that power loss is due to bootstrapping. In fact, for the first four months in Table 8, we do reject the null hypothesis. Also, as time proceeds the standard error get smaller quite

rapidly. The Symbolic Regression method incorporates the heterogeneity, that is fully ignored buy Ordinary Least Squares. So, we are tempted to argue that the bootstrapped standard errors reflect uncertainty more realistically than the OLS based standard errors. At the same time, the parameters in the MZ regression are approaching 0 and 1, respectively, as time proceeds, which is also something you would expect. This does not happen in Table 10.

Conclusion and discussion

Forecasts created by professional forecasters can show substantial dispersion. Such dispersion can change over time but can also concern the forecast horizon. The relevant literature has suggested various sources for dispersion. A recent contribution to this literature by Clements (2017) adds another potential source of heterogeneity, and this is that forecasters may target different vintages of the macroeconomic data. Naturally, the link between targets and forecasts is unknown to the analyst.

To alleviate this problem, we proposed an alternative version of the Mincer Zarnowitz (MZ) regression to examine forecast bias. This version adopts the notion that the vintages of the macroeconomic data can perhaps best be interpreted as interval data, where at the same time, the forecasts also have upper and lower bounds. Taking the data as intervals makes the standard MZ regression a so-called Symbolic MZ Regression. Simulations showed that reliable inference can be drawn from this auxiliary regression. An illustration for annual USA GDP growth rates showed its merits.

A limitation to the interval-based data analysis is the potential size of the intervals. In our case, the sample size is equal to 18 years. When more data become available, the method will become more reliable. A second limitation is that it is assumed that the data are uniformly distributed within the intervals. In our empirical exercise, we have a small number of observations in the intervals, so basically this assumption is an axiom. It shall not be reliable to formally test for the appropriateness of this assumption. Further research with alternative distributional assumptions shall be relevant. At present, our application considers only two variables, and it would be of interest to study the symbolic regression for more variables, as is also done in some of the relevant literature.

Further applications of the new regression should shine light on its practical usefulness. The method does have conceptual and face validity, but more experience with data and forecasts for more variables related to more countries should provide more credibility.

References

Aruoba, S. B (2008), Data revisions are not well behaved, *Journal of Money, Credit and Banking*, 40, 319-340.

Bertrand, P. and F. Goupil (1999), Descriptive statistics for symbolic data, in *Symbolic Data Analysis* (H.H. Bock and E. Diday, editors), Berlin: Springer Verlag, 103-124.

Billard, L. and E. Diday (2000), Regression analysis for interval-valued data, in *Data Analysis, Classification, and Related Methods* (H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen and M. Schader, editors), Berlin: Springer Verlag, 369-374.

Billard, L. and E. Diday (2003), From the statistics of data to the statistics of knowledge, *Journal of the American Statistical Association*, 98, 470-487.

Billard, L. and E. Diday (2007), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Chichester: Wiley

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

Clements, M.P. (2010), Explanations of the inconsistencies in survey respondents forecasts, *European Economic Review*, 54 (4), 536-549.

Clements, M.P. (2019), Do forecasters target first or later releases accounts data?, International Journal of Forecasting, 35, 1240-1249

Coibion, O. and Y. Gorodnichenko (2012), What can survey forecasts tell is about information rigidities?, *Journal of Political Economy*, 120, 116-159

Dovern, F., U. Fritsche and J. Slacalek (2012), Disagreement among forecasters in G7 countries, *The Review of Economics and Statistics* 94 (4), 1081-1096.

Efron, B. and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*. London: CRC Press.

Engelberg, J. C.F. Manski, and J. Williams (2009), Comparing the point predictions and subjective probability distributions of professional forecasters, *Journal of Business & Economic Statistics*, 27 (1), 30-41.

Franses, P.H. (2014), *Expert adjustments of model forecasts*, Cambridge: Cambridge University Press.

Genre, V., G. Kenny, A. Mayler and A. Timmermann (2013), Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* 29 (1), 108-121.

Lahiri, K., and X. Sheng (2010), Measuring forecast uncertainty by disagreement: The missing link, *Journal of Applied Econometrics* 25 (4), 514-538.

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

Legerstee, R. and P.H. Franses (2015), Does disagreement amongst forecasters have predictive value? *Journal of Forecasting* 34 (4), 290-302.

Patton, A.J. and A. Timmermann (2007), Testing forecast optimality under unknown loss, *Journal of the American Statistical Association*, 102, 1172-1184.

Figure 1: The intervals of Table 2.

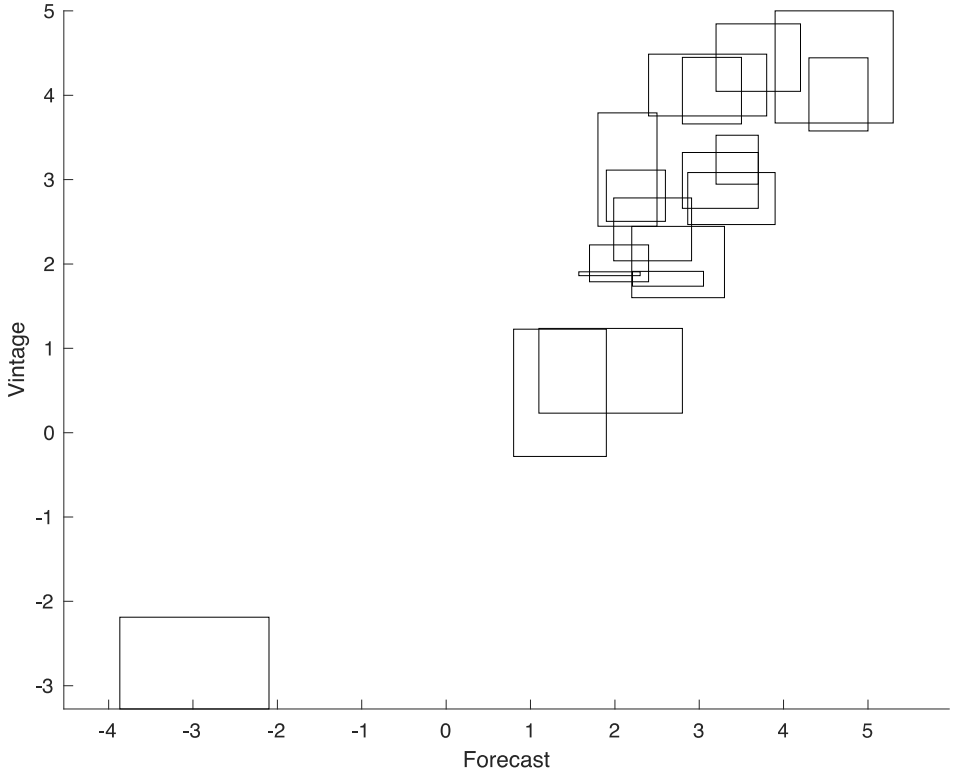

Table 1: An example of the data

Survey Date: May 13, 2013	Gross Domestic Product real, % change 2013	2014
Consensus (Mean)	1,932	2,702
High	2,300	3,380
Low	1,572	2,007
Standard Deviation	0,159	0,319
Number of Forecasts	29	29
UBS	2,300	3,000
American Int'l Group	2,200	2,600
First Trust Advisors	2,200	3,000
Ford Motor Company	2,172	2,996
Morgan Stanley	2,100	2,500
Eaton Corporation	2,053	2,887
Action Economics	2,000	2,800
RDQ Economics	2,000	2,600
General Motors	1,960	2,968
Goldman Sachs	1,959	2,914
Swiss Re	1,953	3,220
Macroeconomic Advisers	1,941	2,968
Moody's Analytics	1,940	3,380
Northern Trust	1,906	2,722
Citigroup	1,900	2,800
DuPont	1,900	3,000
Fannie Mae	1,900	2,500
Inforum - Univ of Maryland	1,900	2,600
Wells Capital Mgmt	1,900	2,600
Univ of Michigan - RSQE	1,880	2,735
Credit Suisse	1,868	2,300
PNC Financial Services	1,846	2,398
Nat Assn of Home Builders	1,843	2,622
IHS Global Insight	1,841	2,799
Barclays Capital	1,803	2,272
Wells Fargo	1,800	2,100
Bank of America - Merrill	1,756	2,684
The Conference Board	1,643	2,374
Georgia State University	1,572	2,007

Table 2: Forecasts and vintages as symbolic data. For the years 1996 to 2013 there are 7 vintages of quotes. For the month May in year *t* there are in between 20 to 30 forecasts. The data in this table are the lower and upper bounds of the intervals of these observations. The data are rounded (at two decimal places) for expository purposes.

	Vintages of real GDP growth		Forecasts	
	Lower bound	Upper bound	Lower bound	Upper bound
Year				
1996	2.45	3.79	1.80	2.50
1997	3.76	4.49	2.40	3.80
1998	3.66	4.45	2.80	3.50
1999	4.05	4.85	3.20	4.20
2000	3.67	5.00	3.90	5.30
2001	0.23	1.24	1.10	2.80
2002	1.60	2.45	2.20	3.30
2003	2.51	3.11	1.90	2.60
2004	3.58	4.44	4.30	5.00
2005	2.95	3.53	3.20	3.70
2006	2.66	3.32	2.80	3.70
2007	1.79	2.23	1.70	2.40
2008	-0.28	1.23	0.80	1.90
2009	-3.28	-2.19	-3.87	-2.10
2010	2.47	3.08	2.86	3.90
2011	1.74	1.91	2.21	3.05
2012	2.04	2.78	1.99	2.91
2013	1.86	1.91	1.57	2.30

Table 3: Simulation experiments for the case where $N = 20$ and $\sigma_\varepsilon^2 = 0.5$. The cells are average estimates of the parameters and associated standard errors (SE) across 1000 replications.

α	β	σ_z^2	σ_w^2	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)
0	-2	0.5	0.5	-0.008 (0.269)	-1.843 (0.168)
0	-2	0.5	2.0	-0.006 (0.317)	-1.581 (0.206)
0	-2	2.0	0.5	-0.011 (0.295)	-1.912 (0.207)
0	-2	2.0	2.0	-0.008 (0.350)	-1.631 (0.248)
0	0	0.5	0.5	-0.010 (0.210)	-0.014 (0.133)
0	0	0.5	2.0	-0.009 (0.216)	0.029 (0.126)
0	0	2.0	0.5	-0.012 (0.256)	-0.083 (0.179)
0	0	2.0	2.0	-0.012 (0.246)	-0.022 (0.171)
0	2	0.5	0.5	-0.011 (0.191)	1.814 (0.128)
0	2	0.5	2.0	-0.013 (0.199)	1.639 (0.137)
0	2	2.0	0.5	-0.014 (0.223)	1.745 (0.159)
0	2	2.0	2.0	-0.015 (0.227)	1.588 (0.164)
5	-2	0.5	0.5	4.992 (0.269)	-1.843 (0.178)
5	-2	0.5	2.0	4.994 (0.318)	-1.581 (0.198)
5	-2	2.0	0.5	4.989 (0.299)	-1.912 (0.210)
5	-2	2.0	2.0	4.991 (0.358)	-1.631 (0.250)
5	0	0.5	0.5	4.990 (0.261)	-0.014 (0.132)
5	0	0.5	2.0	4.991 (0.213)	0.029 (0.122)
5	0	2.0	0.5	4.988 (0.250)	-0.083 (0.171)
5	0	2.0	2.0	4.988 (0.253)	-0.022 (0.167)
5	2	0.5	0.5	4.989 (0.199)	1.814 (0.127)
5	2	0.5	2.0	4.987 (0.208)	1.639 (0.135)
5	2	2.0	0.5	4.986 (0.226)	1.745 (0.166)
5	2	2.0	2.0	4.985 (0.221)	1.588 (0.156)

Table 4: Simulation experiments for the case where $N = 20$ and $\sigma_\varepsilon^2 = 2.0$. The cells are average estimates of the parameters and associated standard errors (SE) across 1000 replications.

α	β	σ_z^2	σ_w^2	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)
0	-2	0.5	0.5	-0.015 (0.464)	-1.788 (0.260)
0	-2	0.5	2.0	-0.013 (0.494)	-1.502 (0.288)
0	-2	2.0	0.5	-0.017 (0.463)	-1.857 (0.291)
0	-2	2.0	2.0	-0.015 (0.516)	-1.552 (0.326)
0	0	0.5	0.5	-0.016 (0.407)	0.040 (0.244)
0	0	0.5	2.0	-0.016 (0.407)	0.108 (0.208)
0	0	2.0	0.5	-0.019 (0.426)	-0.029 (0.273)
0	0	2.0	2.0	-0.019 (0.427)	0.058 (0.254)
0	2	0.5	0.5	-0.018 (0.382)	1.868 (0.224)
0	2	0.5	2.0	-0.020 (0.375)	1.718 (0.208)
0	2	2.0	0.5	-0.021 (0.406)	1.800 (0.257)
0	2	2.0	2.0	-0.022 (0.395)	1.667 (0.232)
5	-2	0.5	0.5	4.985 (0.462)	-1.788 (0.265)
5	-2	0.5	2.0	4.988 (0.490)	-1.502 (0.287)
5	-2	2.0	0.5	4.983 (0.468)	-1.857 (0.287)
5	-2	2.0	2.0	4.985 (0.500)	-1.552 (0.319)
5	0	0.5	0.5	4.984 (0.411)	0.040 (0.234)
5	0	0.5	2.0	4.984 (0.408)	0.108 (0.226)
5	0	2.0	0.5	4.981 (0.448)	-0.029 (0.272)
5	0	2.0	2.0	4.982 (0.420)	0.058 (0.243)
5	2	0.5	0.5	4.982 (0.393)	1.868 (0.225)
5	2	0.5	2.0	4.980 (0.378)	1.718 (0.210)
5	2	2.0	0.5	4.979 (0.385)	1.800 (0.250)
5	2	2.0	2.0	4.978 (0.387)	1.667 (0.229)

Table 5: Simulation experiments for the case where $N = 100$ and $\sigma_\varepsilon^2 = 0.5$. The cells are average estimates of the parameters and associated standard errors (SE) across 1000 replications.

α	β	σ_z^2	σ_w^2	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)
0	-2	0.5	0.5	0.000 (0.093)	-1.878 (0.080)
0	-2	0.5	2.0	-0.000 (0.120)	-1.589 (0.095)
0	-2	2.0	0.5	-0.000 (0.114)	-1.866 (0.110)
0	-2	2.0	2.0	-0.001 (0.132)	-1.580 (0.115)
0	0	0.5	0.5	0.001 (0.088)	-0.036 (0.075)
0	0	0.5	2.0	0.001 (0.086)	-0.048 (0.068)
0	0	2.0	0.5	0.001 (0.109)	-0.024 (0.106)
0	0	2.0	2.0	0.001 (0.105)	-0.040 (0.097)
0	2	0.5	0.5	0.002 (0.110)	1.806 (0.094)
0	2	0.5	2.0	0.003 (0.138)	1.493 (0.117)
0	2	2.0	0.5	0.002 (0.128)	1.818 (0.118)
0	2	2.0	2.0	0.002 (0.149)	1.501 (0.140)
5	-2	0.5	0.5	5.000 (0.095)	-1.878 (0.082)
5	-2	0.5	2.0	5.000 (0.120)	-1.589 (0.097)
5	-2	2.0	0.5	5.000 (0.114)	-1.866 (0.111)
5	-2	2.0	2.0	5.000 (0.139)	-1.580 (0.112)
5	0	0.5	0.5	5.001 (0.086)	-0.036 (0.077)
5	0	0.5	2.0	5.001 (0.088)	-0.048 (0.070)
5	0	2.0	0.5	5.001 (0.107)	-0.024 (0.105)
5	0	2.0	2.0	5.001 (0.107)	-0.040 (0.094)
5	2	0.5	0.5	5.002 (0.108)	1.806 (0.093)
5	2	0.5	2.0	5.003 (0.138)	1.493 (0.120)
5	2	2.0	0.5	5.002 (0.127)	1.818 (0.116)
5	2	2.0	2.0	5.002 (0.149)	1.501 (0.143)

Table 6: Simulation experiments for the case where $N = 100$ and $\sigma_\varepsilon^2 = 2.0$. The cells are average estimates of the parameters and associated standard errors (SE) across 1000 replications.

α	β	σ_z^2	σ_w^2	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)
0	-2	0.5	0.5	0.002 (0.161)	-1.926 (0.129)
0	-2	0.5	2.0	0.001 (0.176)	-1.645 (0.131)
0	-2	2.0	0.5	0.001 (0.174)	-1.914 (0.146)
0	-2	2.0	2.0	0.001 (0.183)	-1.637 (0.139)
0	0	0.5	0.5	0.003 (0.158)	-0.084 (0.131)
0	0	0.5	2.0	0.003 (0.156)	-0.104 (0.119)
0	0	2.0	0.5	0.002 (0.177)	-0.072 (0.146)
0	0	2.0	2.0	0.002 (0.174)	-0.096 (0.137)
0	2	0.5	0.5	0.004 (0.180)	1.758 (0.146)
0	2	0.5	2.0	0.004 (0.200)	1.436 (0.158)
0	2	2.0	0.5	0.003 (0.190)	1.770 (0.164)
0	2	2.0	2.0	0.004 (0.210)	1.444 (0.174)
5	-2	0.5	0.5	5.002 (0.158)	-1.926 (0.126)
5	-2	0.5	2.0	5.001 (0.175)	-1.914 (0.128)
5	-2	2.0	0.5	5.001 (0.175)	-1.914 (0.148)
5	-2	2.0	2.0	5.001 (0.190)	-1.637 (0.141)
5	0	0.5	0.5	5.003 (0.161)	-0.084 (0.129)
5	0	0.5	2.0	5.003 (0.162)	-0.104 (0.122)
5	0	2.0	0.5	5.002 (0.175)	-0.072 (0.153)
5	0	2.0	2.0	5.002 (0.175)	-0.096 (0.141)
5	2	0.5	0.5	5.004 (0.176)	1.758 (0.148)
5	2	0.5	2.0	5.004 (0.197)	1.436 (0.161)
5	2	2.0	0.5	5.003 (0.191)	1.770 (0.165)
5	2	2.0	2.0	5.004 (0.211)	1.444 (0.185)

Table 7: Simulation experiments for the case where $N = 100$. The cells are average estimates of the parameters and associated standard errors (SE) across 1000 replications.

α	β	ρ	σ_ε^2	σ_z^2	σ_w^2	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)
0	-2	0.2	0.5	0.5	0.5	0.002 (0.102)	-1.765 (0.070)
0	-2	0.5	0.5	0.5	0.5	0.003 (0.123)	-1.491 (0.110)
0	0	0.2	0.5	0.5	0.5	0.000 (0.076)	0.001 (0.072)
0	0	0.5	0.5	0.5	0.5	0.000 (0.076)	0.001 (0.066)
0	2	0.2	0.5	0.5	0.5	-0.002 (0.102)	1.767 (0.097)
0	2	0.5	0.5	0.5	0.5	-0.003 (0.126)	1.493 (0.109)
0	-2	0.2	2.0	2.0	2.0	0.002 (0.187)	-1.420 (0.159)
0	-2	0.5	2.0	2.0	2.0	0.004 (0.196)	-1.236 (0.156)
0	0	0.2	2.0	2.0	2.0	0.000 (0.153)	0.003 (0.130)
0	0	0.5	2.0	2.0	2.0	0.000 (0.153)	0.003 (0.121)
0	2	0.2	2.0	2.0	2.0	-0.002 (0.186)	1.427 (0.160)
0	2	0.5	2.0	2.0	2.0	-0.004 (0.195)	1.242 (0.155)

Table 8: Symbolic regression results. Bootstrapped standard errors are in parentheses.

Forecast origin	β_0	β_1	p value Wald test
January, year *t*-1	2.879 (2.448)	-0.141 (0.762)	0.032
February, year *t*-1	3.041 (2.014)	-0.206 (0.648)	0.028
March, year *t*-1	2.689 (2.122)	-0.080 (0.658)	0.021
April, year *t*-1	2.683 (2.090)	-0.076 (0.659)	0.033
May, year *t*-1	2.147 (2.231)	0.118 (0.734)	0.055
June, year *t*-1	1.773 (2.485)	0.250 (0.786)	0.108
July, year *t*-1	0.649 (2.927)	0.655 (0.956)	0.394
August, year *t*-1	-0.104 (2.640)	0.941 (0.893)	0.703
September, year *t*-1	0.554 (2.682)	0.703 (0.959)	0.749
October, year *t*-1	-0.459 (1.417)	1.148 (0.502)	0.944
November, year *t*-1	-0.412 (1.395)	1.156 (0.501)	0.951
December, year *t*-1	-0.324 (0.889)	1.131 (0.318)	0.915
January, year *t*	-0.000 (0.812)	0.999 (0.269)	1.000
February, year *t*	-0.167 (0.559)	1.043 (0.188)	0.951
March, year *t*	0.052 (0.429)	0.987 (0.146)	0.992
April, year *t*	-0.087 (0.420)	1.016 (0.141)	0.966
May, year *t*	-0.009 (0.403)	0.976 (0.130)	0.880
June, year *t*	-0.075 (0.386)	0.990 (0.127)	0.789
July, year *t*	-0.142 (0.331)	1.025 (0.106)	0.856
August, year *t*	-0.068 (0.332)	1.011 (0.118)	0.956
September, year *t*	-0.077 (0.317)	1.000 (0.107)	0.855
October, year *t*	-0.057 (0.291)	1.011 (0.109)	0.965
November, year *t*	-0.095 (0.276)	1.024 (0.105)	0.923
December, year *t*	-0.087 (0.219)	1.006 (0.081)	0.760

Table 9: MZ results, based on the consensus forecasts, first release data. Standard errors are in parentheses.

Forecast origin	β_0		β_1		p value Wald test
January, year *t*-1	2.969	(0.258)	-0.028	(0.085)	0.000
February, year *t*-1	2.898	(0.276)	-0.024	(0.092)	0.000
March, year *t*-1	2.809	(0.293)	0.002	(0.097)	0.000
April, year *t*-1	2.708	(0.288)	0.028	(0.096)	0.000
May, year *t*-1	2.625	(0.283)	0.058	(0.094)	0.000
June, year *t*-1	2.533	(0.276)	0.090	(0.092)	0.000
July, year *t*-1	2.389	(0.266)	0.141	(0.088)	0.000
August, year *t*-1	2.271	(0.244)	0.163	(0.081)	0.000
September, year *t*-1	2.178	(0.258)	0.187	(0.086)	0.000
October, year *t*-1	1.558	(0.331)	0.363	(0.110)	0.000
November, year *t*-1	1.229	(0.361)	0.466	(0.120)	0.000
December, year *t*-1	0.900	(0.350)	0.592	(0.116)	0.014
January, year *t*	0.729	(0.361)	0.675	(0.120)	0.021
February, year *t*	0.441	(0.295)	0.787	(0.098)	0.071
March, year *t*	0.101	(0.284)	0.916	(0.094)	0.485
April, year *t*	0.101	(0.247)	0.937	(0.082)	0.661
May, year *t*	0.009	(0.212)	0.997	(0.070)	0.999
June, year *t*	0.009	(0.182)	1.002	(0.061)	0.986
July, year *t*	0.034	(0.162)	0.988	(0.054)	0.974
August, year *t*	-0.114	(0.141)	1.013	(0.047)	0.519
September, year *t*	-0.082	(0.124)	1.008	(0.041)	0.621
October, year *t*	-0.175	(0.085)	1.035	(0.028)	0.073
November, year *t*	-0.141	(0.067)	1.033	(0.022)	0.089
December, year *t*	-0.160	(0.055)	1.051	(0.018)	0.012

Table 10: MZ results, based on the consensus forecasts, most recent released data (computed: September 2018). Standard errors are in parentheses.

Forecast origin	β_0		β_1		p value Wald test
January, year *t*-1	3.046	(0.219)	-0.059	(0.071)	0.000
February, year *t*-1	2.983	(0.236)	-0.060	(0.076)	0.000
March, year *t*-1	2.910	(0.253)	-0.039	(0.082)	0.000
April, year *t*-1	2.811	(0.251)	-0.012	(0.081)	0.000
May, year *t*-1	2.737	(0.248)	0.015	(0.080)	0.000
June, year *t*-1	2.674	(0.245)	0.036	(0.079)	0.000
July, year *t*-1	2.566	(0.241)	0.075	(0.078)	0.000
August, year *t*-1	2.470	(0.227)	0.089	(0.074)	0.000
September, year *t*-1	2.390	(0.241)	0.109	(0.078)	0.000
October, year *t*-1	1.858	(0.316)	0.256	(0.102)	0.000
November, year *t*-1	1.587	(0.349)	0.341	(0.113)	0.000
December, year *t*-1	1.331	(0.355)	0.442	(0.115)	0.000
January, year *t*	1.208	(0.373)	0.509	(0.121)	0.000
February, year *t*	1.007	(0.351)	0.590	(0.114)	0.002
March, year *t*	0.760	(0.372)	0.687	(0.121)	0.035
April, year *t*	0.766	(0.353)	0.707	(0.114)	0.037
May, year *t*	0.684	(0.333)	0.765	(0.108)	0.083
June, year *t*	0.691	(0.323)	0.768	(0.105)	0.072
July, year *t*	0.700	(0.307)	0.759	(0.098)	0.046
August, year *t*	0.576	(0.310)	0.776	(0.101)	0.083
September, year *t*	0.602	(0.302)	0.773	(0.098)	0.066
October, year *t*	0.516	(0.291)	0.799	(0.094)	0.102
November, year *t*	0.535	(0.278)	0.802	(0.090)	0.087
December, year *t*	0.516	(0.272)	0.820	(0.088)	0.115