Statistical evaluation of the forecasts by The National Institute of Economic Research (KI)

David Sundén, Deputy Managing Director at Copenhagen Economics Sweden
Asger Lunde, Senior Associate, Professor at Aarhus University
Carl von Utfall Danielsson, Analyst at Copenhagen Economics

10 July 2014

Copenhagen Economics have been assigned by The Swedish Agency for Public Management to assess the forecasting precision of some of the key variables which The National Institute of Economic Research (KI) projects.

The content of the assessment and the analyses made have been decided by an Expert group consisting of Jørgen Elmeskov (Rigsstatistiker at Danmarks statistik) and Nigel Pain (OECD) in agreement with The Swedish Agency for Public Management.

Data for the assessment has kindly been provided by the staff at KI, DG Ecfin at the European Commission, and by Nigel Pain at the OECD.

This report includes a description of the data used, and an overview of the statistical analysis and tests made. It is accompanied by a statistical appendix consisting of performance measures, tests, and descriptive statistics.

The first part of the report discusses measures and methods used to evaluate the performance of KI's forecasting separately while the second part concerns KI's forecasting performance compared with other forecasting institutions.

I Evaluation of KI’s forecasts

The first part of the performance measures and statistical tests is only based on KI’s own forecasts and on four main variables. These are:

1. The GDP growth measured as percentage change in constant prices, not calendar-adjusted.

2. General government net lending measured as a share of GDP.

3. Unemployment measured as a share of the labour force, but with different definitions over time. For the period up to and including 2006 the targets for unemployment is defined as open unemployment, during the period 2007–2010 the targets follow the ILO definition and 2011 onwards they follow the EU definition.[1]

4. Inflation measured year by year in per cent but with a change in definition from KPIX to KPIF in August 2008. The KPIX is the underlying inflation and KPIF is the inflation with constant mortgage interest rate.

Besides these four main forecasting targets two deflator variables are also included:

5. GDP deflator measured as the GDP implicit price index.

6. Household consumption deflator measured as consumption expenditure implicit price index.

1.1 Evaluation criteria, data and measure definitions

All forecasts are evaluated based on the forecasting horizon of a single target, which spans from a lead of 8 quarters before the target to 1 quarter before the target.

In practice this means that a target, such as GDP growth in 2013, may have up to 8 different projections. The first in quarter 1 in 2012 (called Q8), the second in quarter 2 in 2012 (called Q7) and so on up to quarter 4 in 2013 (called Q1).

For GDP this can be represented as in the table below. One example is the 2008 GDP growth outcome of -0.2 per cent which was projected to be 3.4 per cent 8 quarters before (quarter 1 in 2007).

[1] There have been several changes in the definitions of the unemployment in the NIER forecasts. Until August 2007, full-time students searching for a job were not included in the definition. From August 2007 to March 2011, unemployment covered people between the ages of 16 and 64 who are out of work, want a job, have actively sought work in the previous four weeks and are available to start work within the next fortnight; or out of work and have accepted a job that they are waiting to start in the next fortnight From March 2011 and onward, the definitions is expanded to cover people between the ages of 15 and 74.

Table 1 Real GDP growth and KI forecasts Q1-Q8

Year	Outcome	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1997	1.8	1.9	2.1	2.1	2.0	2.7	2.2	2.2	2.4
1998	2.9	2.9	3.0	2.9	2.9	3.0	3.1	3.0	2.5
1999	3.8	3.6	3.8		2.2	2.3	3.3	3.3	2.7
2000	3.6	3.8	4.1	4.4	3.9	3.8	3.2		2.9
2001	1.2	1.2	1.6	2.0	2.8	3.4	3.6	3.5	3.3
2002	1.9	1.6	1.9	1.7	1.3	1.5	2.7	3.1	3.1
2003	1.6	1.5	1.3	1.3	1.4	1.8	2.7	2.7	2.7
2004	3.5	3.8	3.5	2.9	2.5	2.2	2.5	2.5	2.8
2005	2.7	2.7	2.4	2.1	3.0	3.2	3.0	2.7	2.6
2006	4.4	4.3	4.1	3.8	3.7	3.6	2.9	2.8	2.9
2007	2.6	2.7	3.5	3.6	3.9	3.6	3.3	3.2	3.2
2008	-0.2	0.8	1.7	2.4	2.5	3.0	3.8	3.7	3.4
2009	-4.9	-4.4	-5.0	-5.4	-3.9	-0.9	1.4	2.0	2.6
2010	5.5	5.6	4.3	3.7	2.4	2.7	1.5	0.8	0.9
2011	3.9	4.5	4.3	4.4	4.2	3.8	3.4	3.0	3.8
2012	0.8	0.9	1.3	0.7	0.4	0.6	1.9	2.9	3.1
2013	1.5	1.0	1.1	1.5	1.3	0.8	1.8	2.3	2.5
2014				2.6	2.4	2.5	2.5	2.3	2.2

Source: The National Institute of Economic Research

Here, it should be observed that all outcomes are measured in terms of the first published result for the year. The outcomes consequently do not include any revisions made after first publication and may not correspond to final published results.

Table 2 Net lending as a share of GDP and KI forecasts Q1-Q8

Year	Outcome	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1997									
1998									
1999									
2000	4.1	3.0	3.5	3.2	2.2				
2001	4.8	3.9	3.6	3.8	4.1	3.0	4.4	4.0	2.9
2002	1.1	1.3	1.7	1.6	1.4	1.4	1.6	2.6	3.5
2003	0.5	0.3	0.4	0.4	0.6	0.6	1.3	1.7	1.6
2004	1.1	0.4	0.7	0.4	0.2	0.4	1.2	1.1	1.1
2005	2.5	1.6	0.8	0.2	0.3	0.0	1.0	0.8	0.5
2006	2.1	2.3	2.1	2.1	1.8	0.5	-0.1	-0.8	-0.3
2007	3.0	3.2	2.6	2.4	2.3	2.1	2.2	1.8	2.2
2008	2.5	2.2	2.6	3.2	3.0	3.3	2.5	2.5	2.7
2009	-0.7	-1.5	-2.3	-2.3	-2.7	-1.3	0.9	1.9	1.9
2010	-0.3	-0.8	-0.6	-0.8	-1.2	-2.5	-3.5	-4.6	-4.6
2011	0.2	0.1	-0.1	0.2	0.1	0.0	-0.8	-1.1	-1.1
2012	-0.7	-0.5	-0.3	-0.1	-0.4	-0.4	0.3	0.5	0.4
2013	-1.3	-1.2	-1.3	-1.5	-1.4	-1.2	-0.4	-0.2	0.2
2014					-2.0	-1.6	-1.5	-1.2	-1.1

Source: The National Institute of Economic Research

The definition of net lending has undergone some revisions since the start of the evaluation period. The figures have been revised to reflect these.[2]

[2] Based on Annex B of Konjunkturinsitutet (2009), "Utvärdering av prognoser för offentliga finanser", some revisions of the data have been made. *Firstly*, due the premium pension system being moved from the government to the private sector, forecasts of revenues have been revised downwards by SEK 25bn for the forecasts of the 2006 outcome made before 2007 (i.e. all forecasts for 2006). Forecasts made before 2007 for the 2007 outcome have been revised down by SEK 26bn. Similarly, expenditure for the same periods has been revised down by SEK 5bn. *Secondly*, due to the expenditure of the Church of Sweden being moved from the government to the private sector, all outcomes and forecasts made before August 2000 have been revised down by SEK 10bn. *Finally*, due to changes in the National Accounts standards, all forecasts from before 2000 have been removed. Other changes are deemed to be of minor importance for the forecasts and outcomes.

The definitions for unemployment, employment and the labour force have changed over the years. In order not to include errors due to definition changes, target definitions and forecast definition have to match.

We have started from KI's target definition for each year. That is "open unemployment" up to and including 2006, the "ILO" definition for periods 2007–2010, and the "EU" definition 2011–2013. We have then matched the forecasts with the target definitions. We only include forecast that match the target definition, those that do not match are reported as missing (NA), see table 3.

Table 3 Unemployment and KI forecasts Q1-Q8

Year	Outcome	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1997	8.0	8.1	8.5	8.5	8.3	7.2	7.0	7.0	7.0
1998	6.5	6.6	6.7	6.7	6.7	6.9	7.7	7.7	7.8
1999	5.6	5.4	5.3		5.9	6.1	6.2	6.2	6.5
2000	4.7	4.7	4.7	4.6	4.6	4.6	4.5		5.7
2001	4.0	4.0	4.0	4.1	3.9	3.8	3.9	3.9	4.0
2002	4.0	4.0	4.0	4.2	4.3	4.3	3.9	3.7	3.6
2003	4.9	4.8	4.7	4.7	4.6	4.4	4.1	4.1	4.1
2004	5.5	5.5	5.5	5.6	5.6	5.3	4.6	4.6	4.3
2005	5.9	5.5	5.4	5.6	5.1	5.0	5.1	5.3	5.3
2006	5.4	5.4	5.4	5.1	4.9	4.6	5.0	5.4	4.6
2007	6.2	6.1	6.2						
2008	6.1	6.1	5.9	5.8	5.9	5.6	5.8		
2009	8.4	8.5	8.8	9.0	8.7	7.9	6.5	5.9	5.9
2010	8.4	8.5	8.5	8.9	9.1	10.1	11.4	11.5	10.7
2011	7.5	7.5	7.5	7.5					
2012	8.0	7.7	7.6	7.5	7.7	7.8	7.5	7.2	
2013	8.0	8.0	8.0	8.3	8.2	8.3	7.9	7.6	7.7
2014					7.9	7.7	7.8	8.3	8.2

Note: Missing values (NA) indicate that the projection definition do not match the target definition

Source: The National Institute of Economic Research

The measure of underlying inflation changed in 2008 from KPIX to KPIF. To avoid any errors from this definition change in the forecast evaluation we have set the target for 2008 and 2009 to be KPIF and removed all of KI's forecast of 2008 and 2009 with the old definition. They have been set to missing in the data, see Table 4.

Table 4 Inflation and KI forecasts Q1-Q8

Year	Outcome	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1997									
1998									
1999									
2000	1.4	1.4	1.2	1.3	1.4	1.4	1.4	1.5	
2001	2.8	2.7	2.6	2.4	1.5	1.4	1.1	1.3	1.4
2002	2.6	2.6	2.5	2.6	2.5	2.2	2.0	1.7	1.4
2003	2.3	2.3	2.3	2.1	2.4	2.2	1.8	1.8	1.9
2004	0.8	0.9	1.0	1.0	0.6	1.0	0.9	0.6	1.3
2005	0.8	0.8	0.7	0.5	0.6	1.1	1.4	1.3	1.3
2006	1.2	1.2	1.2	1.3	1.3	1.1	1.2	1.1	1.3
2007	1.2	1.2	1.2	1.1	0.9	1.2	1.2	1.4	1.5
2008	2.7	2.7	3.0	2.9					
2009	1.9	1.9	1.9	1.7	1.7	1.1	2.4	2.6	
2010	2.0	2.1	2.0	2.0	1.9	1.0	1.0	1.1	1.1
2011	1.4	1.4	1.5	1.6	1.7	1.6	1.3	1.3	1.3
2012	1.0	1.0	1.1	1.1	1.4	1.2	1.3	1.6	1.3
2013	0.9	0.8	1.0	1.0	1.0	1.0	1.3	1.5	1.6
2014					0.7	1.0	1.2	1.3	1.4

Source: The National Institute of Economic Research

The following performance measures are reported with their corresponding definitions:

Table 5 Definitions of measures

Measure	Definition
Mean Error (ME):	Target-Forec
Mean percentage error (MPE):	100*(Target - Forec)/Target
Mean absolute error (MAE):	abs(Target - Forec)
Mean absolute percentage error (MAPE):	100*abs((Target - Forec)/Target)
Root mean squared error (RMSE):	(Target - Forec)^2
Root mean squared percentage error (RMSPE):	100*((Target - Forec)/Target)^2

Note: Target is the outcome of the variable analysed and Forec is its forecast.

2 Performance measures in various sub periods

The first set of measures evaluates the forecasting performance for the full sample time period and for different sub periods.

The purpose of sub-dividing the sample time period is to evaluate if there have been any changes in KI forecasting performance over time and/or if any historical time periods have been harder to project, see section 2 in appendix 4 for the results.

3 Performance measures by decomposition

The second set of measures evaluates the forecasting performance of three main variables based on their subcomponents; GDP growth, general government net lending and unemployment.

The purpose of this is to evaluate if the errors of the main variables are more or less related to the forecasting errors of any of its components, see section 3 in appendix 4 for the results.

The main variables and their components are listed in the following sections.

3.1 GDP growth decomposition

The GDP variable is decomposed into:

1. Household consumption
2. Public consumption
3. Gross fixed capital formation
4. Stock building
5. Export
6. Import

The growth in overall GDP (period *t*) equals the sum of the growth of the components multiplied with their respective shares of GDP (period *t*-1). Using this relation it is possible to assess which subcomponent's error that "contributes" the most to the error in overall GDP growth.

In the tables presented in appendix 4 the mean errors of the GDP-growth components add up to the error of GDP-growth. Thus a large mean error for any of the components indicates a larger contribution to the error in total GDP growth.

It should be observed here that the stock-building variable has a significant variance and changes between years with up to 10 000 percentage points. This complicates the decomposition of the GDP growth variable making the theoretical equality of the growth rates not hold. To make the equality hold we have set stock building as a residual of all other variables in the calculations, thus excluding it from the analysis. The implication being that the error reported for stock-building is the residual error of all other variables. Any results reported for the stock-building variable are thus not valid and do not reflect KI's forecasting of the variable.

3.2 Net lending decomposition

The general government net lending variable is decomposed into general government revenues and general government expenditures. The mean error in the two sub variables add up to the mean error in the main variable. Thus a larger forecasting error in one of the sub-variables directly spills over to a larger error in the main variable.

3.3 Unemployment decomposition

Normally the unemployment rate is calculated as the number of unemployed as a share of the labour force, the employment rate is calculated as the number of employed as a share of the population, and the labour force participation rate as a share of the population.

By definition any forecast error in the *unemployment rate* thus consists of the error in the forecast of the *unemployed* and the error in the forecast of the *labour force*.

In order to indicate to what degree the error in the *unemployment rate* is dependent on the two separate variables we have redefined the unemployment rate to be calculated as a share of the population instead, but only in this section. Such a definition makes it conform to the definition of both the labour force participation rate and the employment rate.

Since unemployed and employed sum up to equal the labour force, the errors in the forecasts of unemployed and employed also sum up. From the point of view that only unemployment and employment are forecasted the labour force is just a resulting variable of the other two. In the tables the errors for all three variables are presented.

4 Testing for unbiasedness

The third set of measures evaluates unbiasedness in forecasting, see section 4 in appendix 4 for the results.

A simple test of whether the projections are systematically biased, or non-zero, is to regress the projection errors (e_t) on a constant:

$$e_t = \alpha + \varepsilon_t$$

where

$$e_t = Outcome_t - Projection_t$$

If the null hypothesis that $\alpha = 0$ cannot be rejected, then the projections are unbiased. We estimate α by OLS and construct 95% confidence intervals based on bootstrapped percentiles.

5 MZ regressions

The fourth set of measures evaluates forecasting efficiency, see section 5 in the statistical appendix for the results. It reports the result of estimating the following regression:

$$Outcome_t = \alpha + \beta \; Projection_t + \varepsilon_t$$

The joint hypothesis that $\alpha = 0$ and $\beta = 1$ is used to test for efficiency. Here efficiency refers to checking that the forecasts and their errors are uncorrelated. If there is a systematic relationship between these, then it could be used to help predict future errors, and in turn used to adjust the forecasting model. The concept of efficiency in this setting was suggested in Mincer and Zarnowitz (1969). See for example Pain and Britton (1992) for an extensive application.

The MZ regression in the above equation is also useful when comparing forecasts performance across horizons or providers. The R-squared form in this regression is closely related to the MSE, but easy to use in a comparison because it is scaled between 0 and 1.

6 Error correlation between main targets

The fifth set of measures shows the error correlations between the main variables, see section 6 in appendix 4 for the results.

II Comparisons of forecasts

The second part of the performance measures and statistical tests evaluates KI's forecasting performance in comparison with other forecasters.

There is only sufficient time series data for three of the four main targets to do the comparison; that is for the GDP growth, the unemployment rate and inflation.

Forecasts where the definition of the variable used in the forecast does not match that of the target variable as defined by KI have been removed from the analysis. For example, if the forecasting institution's definition of unemployment does not match the unemployment definition used by KI for a given target year, that forecast is excluded.

Furthermore, the data set containing the forecasts and targets are different from the one analysed in the previous section. The figures for KI and the targets are the same but are rounded to one decimal. The reason behind this is that the other institutes' forecasts are reported with one decimal. In order not to give KI a precision advantage compared to the other institutions, KI's figures also are rounded to one decimal. Consequently, the errors reported in this section do not exactly match the one reported in section I.

The institutions that KI are evaluated against are summarised in table below including the abbreviations used in the tables in appendix 4.

Table 6 Included institutions

Institution (Sv.)	Institution (Eng.)	Abbreviation
Europeiska Unionen	European Union	EU
Organisationen för ekonomiskt samarbete och utveckling	Organisation for Economic Co-operation and Development	OECD
Finansdepartementet	Ministry of Finance	FiD
Handelns Utredningsinstitut	HUI Research	HUI
Landsorganisationen i Sverige	The Swedish Trade Union Confederation	LO
Nordea	Nordea	Nordea
Riksbanken	The Riksbank	RB
Skandinaviska Enskilda Banken	Skandinaviska Enskilda Banken	SEB
Svenska Handelsbanken	Svenska Handelsbanken	SHB
Svenskt Näringsliv	The Confederation of Swedish Enterprise	SN
Swedbank	Swedbank	Swed

The evaluation of KI's forecasting performance as compared to other institutions is done in five sets of measures, tests and descriptive statistics described below.

7 Comparing the number of forecasts

In section 7 in appendix 4 we present a summary of the available number of forecasts for each of the included institutions. For each of the three main target variables GDP, unemployment and inflation three counts are given. The observation count of the full 1997-2013 sample followed in brackets by the number of observation in the 1997-2007 and the 2008-2013 periods. Note that the number of forecasts varies considerably across institutions. This is reflected in the following sections were we compute a particular measure or conduct a particular analysis only if a reasonable number of data points are at hand.

8 Performance measures

For the individual institutions the same performance measures as in section 2 and section 3 are displayed, see section 8 in appendix 4. Here the samples are split as indicated by the table of the number of forecasts (Table 7.1 in appendix 4). Missing values are reported in the tables if there were less than 12 target years in the full sample, less than 7 years for the period 1997-2007, and less than 5 years for 2008-2013.

9 Model confidence sets

Section 9 in appendix 4 present model confidence sets (MCS). Included in this comparison are the institutions that have at least 10 overlapping forecast for GDP, 9 for unemployment and 11 for inflation, respectively.

The MCS methodology is due to Hansen, Lunde, & Nason (2011). The outcome of this approach is a subset of forecasts that are not distinguishable from the best forecast across the complete set of forecasts. Defining the set of all competing forecasts as $M = \{A, B, C, \dots\}$, the MCS tests the null that no forecast is distinguishable against an alternative that at least one of the forecasts has a higher expected loss,

$$H_0 : E\, L(Y_t, F_t^i) = EL(Y_t, F_t^j)\ for\ all\ i, j \in M$$

vs.

$$H1 : E\, L(Y_t, F_t^i) = EL(Y_t, F_t^j)\ for\ some\ \ i \in M, for\ all\ j \in M \setminus i.$$

The MCS operates by iteratively deleting poorly performing forecasts to construct a set, $\widehat{M}^*$, that contains the forecast producing the lowest expected loss with probability weakly greater than the level of the test (e.g. 0.05), with the property that the probability that this set contains a sub-optimal forecast asymptotes to zero with the sample size. The MCS resembles in many respects a confidence interval for a parameter.

As an example of how this works consider Table 9.1 in appendix 4. Panel A of Table 9.1 presents MCS for GDP growth. The MCS is computed separately for each forecasting horizon. Consider the column denoted Q1 it has the MSE for each institution followed by the MCS p-value. Because of the scarcity of data we work with a 10% significance level. So all the institutions with a MCS p-value that is larger than 10 are in the 90% model confidence set (denoted $\widehat{M}^*{}_{90\%}$). Thus, at the one-step-ahead horizon one cannot reject that the institutions perform equally well. Now, there are cases where an institu-

tion seems to underperform. For Q3 FiD is out and SHB is out for Q6 and Q7. However, the overall conclusion that we think emerge from Table 9.1 is that there is not enough information in the data (too few observations) to tell the unconditional performance of the institutions apart.
Finally, let us explain why we apply the MCS approach. In general, comparing multiple alternatives is a non-standard problem. One of the complications that arise when comparing several alternatives is that spurious results may appear. A decent, but not superior, alternative can be “lucky” in a particular sample appearing to be better than all other alternatives. The more alternatives one compares, the higher the probability is that some alternative will appear superior by chance.

One often sees studies comparing many models using the so called Diebold-Mariano test applied pairwise (see Diebold and Mariano (1995)). Now when, say, M forecasting methods are being compared then there will be a total of $n = M(M-1)/2$ pairwise comparisons. If these n tests were independent and each test evaluated at an α significance level, then under the null for each test the relevant question is the following: What is the probability, α^*, of wrongly rejecting the null for one or more of these tests? For $n = 1$ this probability is obviously α per definition. For $n > 1$ we have

$$\alpha^* = P(\text{at least one Type I errors in } n \text{ tests}) = 1 - (1-\alpha)^n$$

Hence, if we choose an α significance level in each individual of the n tests, then the overall size of the tests is $1-(1-\alpha)^n$. Now, Bonferroni bounds would have us use $\alpha = \alpha^*/n$. So correctly sized pairwise DM tests will have no power in a setting like this! The worst case scenario would be that p-values from such standard two-sample pairwise comparisons of M objects at an α level should be multiplied by the number $M(M-1)/2$. Of course the pairwise comparisons are not independent tests, and hence the Bonferroni method would be much too conservative. The MCS methodology was designed to handle this problem, for the theory, simulations and examples see Hansen, Lunde and Nason (2011).

10 Encompassing regressions

The relative performance of two sets of projections can be assessed by forecast encompassing tests (see e.g. Newbold and Harvey, 2004), to test whether the KI projections contain all the information in the alternative forecast. For each of the 8 different forecast horizons, the main targets are regressed on KI and each of the alternative forecasts in turn:

$$Outcome_t = \beta_0 + \beta_1\, Projection_t^{KI} + \beta_2\, Projection_t^{alt} + \varepsilon_t$$

In section 10 in appendix 4 the estimates of β_1 and β_2 are presented with bootstrap p-values. The column entitled $\Delta R^2\%$ reports how much the R^2 increases when the alternative forecast is included in the regression.

11 Illustrations: Predictions and outcome

To visualise the forecasting performance of the different institutions for different horizons we have plotted the Q1-Q8 forecast for every target year and for every institution for three of the main variables; GDP-growth, unemployment rate and inflation, see section 11 in appendix 4 for the results.

12 References

Diebold, F.X. and Mariano, R.S. (1995), Comparing predictive accuracy. Journal of Business & Economic Statistics, 13, 253-263.

Hansen, P.R., A. Lunde, & J.M. Nason (2011), The model confidence set. Econometrica 79, 456–497.

Hendry D.F. and Clements, M. (1998), Forecasting Economic Time Series, Cambridge University Press.

Mincer, J., Zarnowitz, V. (1969), The evaluation of economic forecasts and expectations. In: Mincer, J. (Ed.), Economic Forecasts and Expectations. National Bureau of Economic Research, New York.

Newbold, P. and D.I. Harvey (2004), “Forecast combination and encompassing”, in M.P. Clements and D.F. Hendry (eds.), A Companion to Economic Forecasting, Blackwell Publishing, Oxford.