CESifo WORKING PAPERS

8266
2020

May 2020

Categorical Forecasts and Non-Categorical Loss Functions

Constantin Bürgi, Dorine Boumans

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8266*

Categorical Forecasts and Non-Categorical Loss Functions

Abstract

This paper introduces a new test of the predictive performance and market timing for categorical forecasts based on contingency tables when the user has non-categorical loss functions. For example, a user might be interested in the return of an underlying variable instead of just the direction. This new test statistic can also be used to determine whether directional forecasts are derived from non-directional forecasts and whether point forecast have predictive value when transformed into directional forecasts. The tests are applied to the categorical exchange rate forecasts in the ifo-Institute's World Economic Survey and to the point forecasts for quarterly GDP in the Philadelphia Fed's Survey of Professional Forecasters. We find that the loss function matters as exchange rate forecasters perform better under non-categorical loss functions, and the GDP forecasts have value up to two quarters ahead.

JEL-Codes: C120, C520, E370, F370.

Keywords: contingency tables, categorical forecast, profitability, World Economic Survey, directional accuracy, market timing, forecast value.

Constantin Bürgi
St. Mary's College of Maryland
47645 College Drive
USA – St. Mary's City, MD, 20686
crburgi@smcm.edu

Dorine Boumans
ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Germany – 81679 Munich
boumans@ifo.de

April 21, 2020
We would like to thank Leo Boumans, M. Hashem Pesaran, Stefan Sauer and Melissa Verin for their valuable comments and support.

1 Introduction

Starting at least with Merton (1981) and Henriksson and Merton (1981), there has been an extensive literature focusing on categorical and directional forecasts and their value for market timing. While some of the research since aimed at non-binary loss functions for categorical forecasts (e.g. Costantini et al. (2016), Blaskowitz and Herwartz (2011) or Anatolyev and Gerko (2005)), it mainly focused on the profitability of trading strategies, rather than more general loss functions. This paper aims to extend this literature on categorical forecasts by introducing new test statistics that allow for more general loss functions when testing the value and predictive performance as defined by Pesaran and Timmermann (1992) for contingency tables.[1] For example for directional forecasts, the new test statistics can weight the categorical forecasts by the specific point deviations of the non-categorical realizations from the cutoff. The tests can then be utilized to assess, whether the directional forecast is profitable rather than assessing whether the direction was predicted better than chance. As point forecasts should get larger deviations from a cutoff more often correct than chance, the weighted tests can also act as a test, whether categorical forecasts are derived from point forecasts and whether point forecast have predictive value. After presenting the new tests, the tests are applied to the directional foreign exchange forecasts in the World Economic Survey (WES) conducted by the ifo-Institute in Munich as well as the point forecasts for GDP in the Survey of Professional Forecasters (SPF) conducted by the Philadelphia Fed.

[1] A parametric test that just looks at correct and incorrect predictions like Diebold and Mariano (1995) might not be adequate here, as each category can have a different distribution. Also, extending the test to serially correlated data as in Pesaran and Timmermann (2009) or Blaskowitz and Herwartz (2014) is beyond the scope of this paper.

2 The Unweighted Case

Assume a user has T forecasts X_t with m categories for an underlying variable Y_t with the same categories. A common way to test the value of the forecast is to create a contingency table of proportions and then test their statistical independence according to the Pearson Chi-squared test.[2] Let $\hat{\pi}_{ij}$ denote the estimated proportions in cell (i,j) and $\hat{\pi}^*_{ij} = (\sum_i \hat{\pi}_{ij})(\sum_j \hat{\pi}_{ij})$ denote the corresponding expected proportions. The standard Pearson Chi-squared test statistic for this case is

$$S_T = T \sum_{i,j=1}^{m} \frac{(\hat{\pi}_{ij} - \hat{\pi}^*_{ij})^2}{\hat{\pi}^*_{ij}} \sim \chi^2_{(m-1)^2} \tag{1}$$

Under the null, $H_0 : \pi_{ij} = \pi^*_{ij} \forall i, j$, where $\pi^*_{ij} = (\sum_i \pi_{ij})(\sum_j \pi_{ij})$ there is statistical independence. If the null is rejected, the forecast is different from randomly guessing and it can be used in a market timing context. Instead of this independence test, Pesaran and Timmermann (1992) introduced a test of predictive accuracy for contingency tables. The null for that test is that the sum of the diagonal elements is different from their expected sum $H^*_0 : \sum_i (\pi_{ii} - \pi^*_{ii}) = 0$ with the test statistic

$$S^*_T = T \frac{(\sum_i (\hat{\pi}_{ii} - \hat{\pi}^*_{ii}))^2}{\hat{V}_T(\hat{\pi})} \sim \chi^2_1 \tag{2}$$

where

$$\hat{V}_T(\hat{\pi}) = \nabla g(\hat{\pi})'(diag(\hat{\pi}) - \hat{\pi}\hat{\pi}')\nabla g(\hat{\pi}) \tag{3}$$

and

$$g(\hat{\pi}) = \sum_i (\pi_{ii} - \pi^*_{ii}) \tag{4}$$

evaluated at $\hat{\pi}$. Note that in the two by two case, the H_0 is identical for both tests, but the Pesaran and Timmermann (1992) test has a higher variance in finite samples.[3] As a result, the Chi-squared test is more efficient in the two by two case and the Pesaran and Timmermann (1992) should only be used for if there are more than two categories.

[2] The Fisher exact test or the Yates continuity adjustments might be necessary for small samples.

[3] As mentioned in Pesaran and Timmermann (1992), the variance is equal to the Chi-squared variance plus $\frac{4}{n}\pi^*_{11}\pi^*_{22} > 0$, and the difference is only asymptotically negligible.

3 The Weighted Case

In many cases, the underlying variable Y_t is derived from a variable y_t that has more categories than Y_t (e.g. it might be continuous). A user might then be interested in getting $f(y_t)$ for some function $f(\cdot)$ correctly predicted, rather than Y_t. For example in the two-by two case, one might want to put a higher weight on observations that are far away from the cutoff c compared to observations that are closer to this threshold. Specifically, the weights might take the form of $|y_t - c|$. Similarly, in a three-by-three case with asset returns and outcomes up, down and unchanged, the researcher might put a smaller weight on getting the unchanged outcome right.

Denote $\hat{p}_{ij} = \sum_t w_t I(Y_t = i, X_t = j)$ the weighted proportions, where I is the indicator function, w is a vector of weights w_t and $\sum_t w_t = 1$. It is further assumed that the weighted proportions $\hat{p}$ are a consistent estimate of the true cell proportions p.[4] Then the π from the previous section can be replaced with with p and an asymptotically correct test statistic is obtained (e.g. see Bishop et al. (2007)). That is

$$S_T = \tilde{T} \sum_{i,j=1}^{m} \frac{(\hat{p}_{ij} - \hat{p}^*_{ij})^2}{\hat{p}^*_{ij}} \sim \chi^2_{(m-1)^2} \tag{5}$$

for the Pearson Chi-squared test and

$$S^*_T = \tilde{T} \frac{\left(\sum_i (\hat{p}_{ii} - \hat{p}^*_{ii})\right)^2}{\hat{V}_T(\hat{p})} \sim \chi^2_1 \tag{6}$$

for the predictive accuracy test. If the statistics still had a T instead of the $\tilde{T}$, the number of observations would likely be overestimated. Because some observations have a higher weight than others, the effective sample size might be different. Instead of T, one should use $\tilde{T} = T/\delta$ for the test statistic, where δ is an estimate for the generalized design effects. For the weighted Chi squared statistic, Rao and Scott (1984) pointed out that the statistic is the

[4] This assumption is relatively weak. If the weights are $|y_t - c|$ for example, this holds as long as y_t has finite second moments. Then $\hat{p}$ converges to π. If the weights only differ by category (e.g. equal weight on up and down but a smaller weight on unchanged), this also trivially holds.

weighted sum of squared normal variables, where the weights correspond to cell specific δs. They propose several different ways to obtain the appropriate δs that can change the degrees of freedom of the Chi squared statistic or change it to a F distribution and are summarized in Scott (2007). Many of these are already readily implemented in statistical packages as they are commonly used for survey data analysis. For the Pesaran and Timmermann (1992) test, $\tilde{T}$ can be estimated by dividing the weighted variance by the variance under random sampling which leads to an effective sample size of

$$\tilde{T} = T/\delta = T * \frac{\hat{V}_T(\hat{\pi})}{\hat{V}_T(\hat{p})} \tag{7}$$

Note that for a conservative estimate, one would set $\tilde{T} = T$ for $\delta < 1$.

Now that the two statistical tests have been adapted to accommodate weightings, it is possible to use these tests in a much wider setting. Specifically, the weighted Chi-squared test can determine if directional forecasts have value in a market timing setting or for any other loss functions. Similarly, the weighted Pesaran and Timmermann (1992) test can test the predictive accuracy with more general loss functions. For example, a model might not get the direction correct more often than a coin toss, but it might get all the large deviations from the cutoff right. This model is then deemed to not have good predictive accuracy based on the unweighted tests. However, if the observations are weighted according to the deviations from the cutoff, this changes and it now has good predictive accuracy. This result can also go the other way if the unweighted tests show a good predictive accuracy, but the weighted ones do not.

In addition to generalizing the application of these tests, the weighted statistics are also able to test, whether the forecasts are derived from point forecast. Point forecasts are often evaluated based on their mean squared error (MSE). Due to this, forecasters aim to predict values as close to the actual as possible and they should be better at predicting the direction of larger deviations from the cutoff than smaller ones. This is different for directional forecasters. A directional forecaster does not aim to get close to the actual as long as he gets the direction right and does not necessarily predict the direction of larger deviations from the cutoff better than smaller ones. If the weights are chosen to be the squared distance from the cutoff, the loss function is similar to a MSE and a comparison between the weighted and unweighted statistic might distinguish whether forecasts are based on point forecasts or not. One would expect to reject the null hypothesis of no predictive value for the weighted case if the categorical forecasts are based on point forecasts. As the direction might be easier to predict for points that are far from the cutoff, this is not a sufficient condition to determine whether directional forecasts are based on point forecasts. A sufficient (but not necessary) condition would be the failure to reject the null of no predictive value for the unweighted case.

Last but not least, this new measure can also be used to better test, whether point forecasts have predictive value. So far in order to determine whether point forecasts had value, they needed to be converted to directional forecasts and then the predictive value for the directional forecast were tested. However, forecasts might fail to predict the direction better than chance but still be valuable point forecasts if the observations with large deviations from the cutoff are correctly predicted. The weighting introduced in this paper allows for a test statistic that is robust against this issue. In line with the often used squared loss functions, one can use the squared deviations of the actual from the cutoff as weights. A point forecast then has predictive value, if the null of no predictive value can be rejected based on the weighted test. This benchmark comes directly from the market timing notion of directional forecast and might provide a theoretically well founded benchmark. Specifically, it might be a less arbitrary benchmark than the MSE being smaller than the one of a naive forecast like the ex post mean of the underlying variable.

4 Applications

The newly developed test is first applied to a subset of the directional exchange rate forecasts in the WES collected by the ifo-Institute in Munich. Every quarter, the WES collects six month ahead EUR/USD directional forecasts from around 20-30 experts in the major countries that use the Euro and our sample includes forecasts made in the period Q2 1999-Q2 2019. While individual forecasters provide three categories (up, down and unchanged), an average score by country is calculated across forecasters (attributing a 1 for up, 0 for unchanged and -1 for down at the individual level). The sign of the resulting average prediction in each period is used to create the unweighted two-by-two contingency table.[5] To obtain a weighted table, each period is weighted by the realized absolute log change in the currency over the period in question.[6] For example, the unweighted and weighted contingency tables for the sign of German forecasters are shown in Table 1. The unweighted Chi-squared, weighted F-stat p-values as well as the p-values for the weighted and unweighted Pesaran and Timmermann (1992) (PT) statistic are reported in Table 2.

Table 1: Contingency Tables for German Forecasters

Unweighted A/F	Down	Up
Down	27	17
Up	14	23
Weighted A/F	Down	Up
Down	29.86	15.78
Up	11.61	23.76

The table shows the number of periods in each of the categories. The weighted part does not add up to 81 due to rounding.

In line with the results by Meese and Rogoff (1983), there is little evidence that the

[5] We use a two-by-two table to get larger numbers of observations in each cell. Ties are broken by a coin toss.

[6] We use the log change instead of the percentage change, as percentage changes are not symmetric.

Table 2: P-values Across Countries

Country	Unweighted Chi	Weighted F	Unweighted PT	Weighted PT
Germany	0.059*	0.016**	0.031**	0.002***
France	0.915	0.817	0.742	0.789
Italy	0.308	0.166	0.218	0.088*
Spain	0.554	0.365	0.435	0.279

This table reports the p-value of the specified test. *, ** and *** imply significant at 10%, 5% and 1% level, respectively. Weighted are the directions weighted according to the change in the underlying variable.

EUR/USD exchange rate can be accurately predicted for forecasters in all countries except for Germany. Based on the tests, the null of no predictive value is rejected for German forecasters. Moreover, the predictive ability broadly increases with the weighted test relative to the unweighted version. This means that forecasters are more likely to get the large changes in the exchange rate right, rather than the small changes. In Germany, this is shown by the directional Chi statistic being only significant a the 10% level, but the weighted F statistic being significant at the 5% level.[7] This result is also in line with forecasters in Germany basing their directional forecasts on point forecasts for the exchange rate.

The second application is to test, how far out the average point forecast of the Survey of Professional Forecasters (SPF) for quarterly real GDP has predictive value. The sample used uses the forecasts made from Q4 1968 up to the ones made for Q4 2019. The survey collects forecasts for the current quarter (H0) and up to four quarters ahead (H4). The first release of GDP is used as the actual and in order to transform the point forecasts into directional forecasts, the mean of the actual is used as the cutoff. This means that if the average of

[7] The same is true for the PT statistic, where the unweighted PT statistic is significant at he 5% level but the weighted one at the 1% level.

individual forecasts for one period predicts faster GDP growth than the cutoff, it takes value one, otherwise 0. The weights used for the test statistic are the squared deviations of the actual from the cutoff. For the point forecast to have predictive value, it should reject the null of no predictive value for the weighted statistic.

Table 3: P-values for Real GDP Growth

Country	Weighted F	Weighted PT
H0	0.000***	0.000***
H1	0.000***	0.000***
H2	0.000***	0.000***
H3	0.325	0.462
H4	0.076*	0.605

This table reports the p-value of the specified test. *, ** and *** imply significant at 10%, 5% and 1% level, respectively.

The p-values of the weighted test statistics are reported in Table 3. There is clear evidence that forecasts have predictive value at the first three horizons. For these horizons, the null of no predictive performance can be rejected. For the last two horizons (three and four quarters ahead), the null of no predictive performance cannot be rejected. This implies that the average of the forecasters in the SPF cannot accurately predict the weighted direction beyond two quarters ahead and hence have a limited predictive ability. If the forecasts are split into three categories (up, down and unchanged), the results remain broadly unchanged as the null can be rejected for the first three horizons. For the last two horizons, the test does not have enough observations per cell, which is why a table has been omitted.

5 Conclusion

This paper extended the non-parametric predictive performance test by Pesaran and Timmermann (1992) as well as the Chi squared independence test for predictions to settings with non-categorical loss functions. While there have been other extensions of these tests aimed at assessing return loss functions, this paper provides a natural extension that allows for a much wider array of loss functions. This new test statistic allows further to determine, whether categorical forecasts are derived from point forecasts and can be used to asses the market timing value of point forecast beyond their direction alone.

The application to EUR/USD exchange rate forecasts has reconfirmed the difficulty in predicting exchange rates, even if it appears possible for some forecasters to accurately predict the direction six months ahead. Further research might be able to determine, whether this is a statistical artifact, or if the accurate prediction is due to superior models. Similarly with the application to the SPF real GDP forecasts, it was shown that forecasts have value up to two quarters ahead.

References

Anatolyev, S. and Gerko, A. (2005). A trading approach to testing for predictability. *Journal of Business & Economic Statistics*, 23(4):455–461.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Blaskowitz, O. and Herwartz, H. (2011). On economic evaluation of directional forecasts. *International journal of forecasting*, 27(4):1058–1065.

Blaskowitz, O. and Herwartz, H. (2014). Testing the value of directional forecasts in the presence of serial correlation. *International Journal of Forecasting*, 30(1):30–42.

Costantini, M., Cuaresma, J. C., and Hlouskova, J. (2016). Forecasting errors, directional accuracy and profitability of currency trading: The case of eur/usd exchange rate. *Journal of Forecasting*, 35(7):652–668.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Henriksson, R. D. and Merton, R. C. (1981). On market timing and investment performance. ii. statistical procedures for evaluating forecasting skills. *Journal of business*, pages 513–533.

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of international economics*, 14(1-2):3–24.

Merton, R. C. (1981). On market timing and investment performance. i. an equilibrium theory of value for market forecasts. *Journal of business*, pages 363–406.

Pesaran, M. H. and Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4):461–465.

Pesaran, M. H. and Timmermann, A. (2009). Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association*, 104(485):325–337.

Rao, J. N. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of statistics*, pages 46–60.

Scott, A. (2007). Rao-scott corrections and their impact. In *Proceedings of the 2007 joint statistical meetings, Salt Lake City, Utah.*