Economics Bulletin

Volume 36, Issue 2

Predicting events with an unidentified time horizon

Patrick De lamirande
CBU

Jason Stevens
University of Prince Edward Island

Abstract

Economists often employ binary choice models to determine if variables of interest such as asset prices or returns are able to predict the occurrence of significant events, most notably recessions. It is, however, unclear how the results of existing studies should be interpreted due to the common practice of testing the predictability of the event at multiple horizons. Presented with a set of test statistics, some may be tempted to conclude that the variable of interest is able to predict the event if the null hypothesis of non-predictability is rejected at any horizon. This paper demonstrates that this approach results in a significant probability of spuriously concluding that the event of interest is predictable. In light of this possibility, the ability of the term spread to predict US recessions is re-examined with corrected critical values, confirming that the results found in the existing literature are not the result of data-snooping.

Computational resources are provided by ACENET, the regional advanced research computing consortium for universities in Atlantic Canada. ACENET is funded by the Canada Foundation for Innovation (CFI), the Atlantic Canada Opportunities Agency (ACOA), and the provinces of Newfoundland & Labrador, Nova Scotia, and New Brunswick.
Citation: Patrick De lamirande and Jason Stevens, (2016) "Predicting events with an unidentified time horizon", *Economics Bulletin*, Volume 36, Issue 2, pages 729-735
Contact: Patrick De lamirande - patrick_delamirande@cbu.ca, Jason Stevens - jmstevens@upei.ca.
Submitted: September 19, 2014. **Published:** April 14, 2016.

1. Introduction

A significant literature employs binary choice models to evaluate the predictability of important events. Most notably, a large number of studies investigate whether financial variables are able to help predict the occurrence of recessions, with significant attention paid to the term spread. Examples of this include Dueker (1997), Funke (1997), Estrella and Mishkin (1998), Atta-Mensah and Tkacz (1998), Chauvet and Potter (2005), Moneta (2005), and Kauppi and Saikkonen (2008). It should be noted that recessions are not the only event of interest, Chen (2009) uses a similar approach to predict downturns in the stock market.

Unfortunately, the question of predictability is rarely posed for a specific time horizon. The standard approach is to test this hypothesis for several horizons [1 month, 1 quarter, 1 year, etc.], producing a collection of test statistics, each evaluating whether the variable of interest is able to predict the onset of a recession at a specific horizon.

Some may be tempted to conclude that the onset of a recession is predictable if any of these test statistics are found to be significant. This paper demonstrates that this approach yields a significant probability of reaching a spurious conclusion in a manner analogous to that discussed by White (2000). In other words, as the set of test statistics is expanded, the probability that a subset will exceed a fixed critical value grows, increasing the probability of reaching a spurious conclusion. While the dangers of data-snooping are well known, this particular form seems to have gone unrecognized within the literature.

More formally, Section 2 argues that reaching conclusions based on a subset of test statistics amounts to conducting inference when a nuisance parameter is not identified under the null hypothesis. It is well known that the distribution of such a statistic diverges from its standard counterpart, causing an elevated probability of reaching a spurious conclusion; see Hansen (1996) for a discussion of these issues.

Section 3 re-examines the ability of the term spread to predict U.S. recessions using corrected critical values obtained from a bootstrap procedure. The results of this exercise show that increasing the number of test statistics from 1 to 4 increases the 5% critical value by 30.62%. Even more dramatically, the use of 8 test statistics increases the 5% critical value by 36.7%. At the 10% level of significance, the results are even more dramatic with the critical values increasing by 37.4% and 56.5% for 4 and 8 test statistics respectively. This result demonstrates the importance of adjusting critical values to account for data snooping, and should prove useful in other applications. However, in the case of the term spread, it is found that its predictive value remains statistically significant, indicating that the results found in the existing literature are not the result of data snooping.

2. Spurious Inference

This section provides a brief overview of the standard approach to evaluating the ability of variables of interest to predict recessions, demonstrating that it produces a model with an unidentified nuisance parameter under the null hypothesis. As noted by Hansen (1996) and others, this increases the probability of spurious inference when conclusions are based on critical values obtained from standard distributions.

Given a dataset containing T observations (indexed as $t \epsilon [1, T]$), the model used to test if the observation of X helps predict the occurrence of Y at a horizon of k periods is specified as:

$$P(Y_t = 1) = F(\delta + \beta X_{t-k}) \quad (1)$$

In this application, δ and β are parameters to be estimated, with the null hypothesis of interest stated as $\beta = 0$. $F(\cdot)$ represents a cumulative distribution function; normal in the case of a probit and logistic in the case of a logit. While X is treated as a scalar in what follows to facilitate discussion, the underlying issue with identification also applies to multivariate analysis. This focus here is on in-sample analysis, but the conclusions to follow also apply to out-of-sample analysis.

It is important to distinguish between testing the general null hypothesis that X does not predict Y at any horizon and the specific null hypothesis that X does not predict Y at a fixed horizon of k periods (such as 1 year). In the specific case, the critical value of the test statistic (t, Wald, Likelihood ratio, Lagrange Multiplier, etc.) used to test the null hypothesis may be obtained from a standard distribution[1].

This is not the case for a test of the general hypothesis. In many applications, test statistics are obtained for $n > 1$ horizons. A reader may be tempted to conclude that X helps predict the occurrence of Y if the null hypothesis is rejected for any horizon. Mackinnon (2006) argues[2] that this approach is likely to be quite unreliable. In a specific example, Chow and Denning (1993) demonstrate that this approach causes spurious inference in the case of a variance ratio test for a unit root. In addition, Dufour and Taamouti (2010) argue this approach is inappropriate in the context of Granger causality tests covering multiple horizons. Intuitively, analogous to the argument of White (2000), as more test statistics are obtained, the probability that (at least) one will exceed a fixed critical value increases. Simply stated, when presented with a collection of test statistics, it is unclear whether a set of test statistics exceeding a fixed critical value represents evidence of a predictive relationship between the variables or is simply the result of accumulating multiple test statistics (data snooping).

If the collection of test statistics (ts) obtained from this approach is indexed as i: 1,…n, this approach is equivalent to maximizing the value of the test statistic through the selection of the time horizon (k).

$$ts^{\max} = \max[ts_1, ..., ts_n] \quad (2)$$

In technical terms, k represents an unidentified nuisance parameter under the null hypothesis ($\beta = 0$). Within the existing empirical literature, the issue of identification is most commonly encountered in testing for the presence of a structural break in a model's parameters (for example, Andrews (1993)). Hansen (1991) contains a substantial discussion of different models in which similar problems with identification exist. While a complete survey of this literature is beyond the scope of this paper, the universally accepted conclusion is that the use of critical values obtained from standard distributions (t, χ^2 , etc.) results in over-rejection of the null hypothesis.

[1] Although, often only asymptotically.
[2] Mackinnon was referring to use of a Maximum test statistic.

It is important to note that the exact critical values used in the analysis depend on the choice of which test statistics are included in (2). For example, Dueker (1997) uses horizons of 3,6, 9 and 12 months, while Funke (1997), Estrella and Mishkin (1998), and Moneta (2005) use each horizon between 1 and 8 quarters.

In most cases it is not possible to derive an analytical expression for the distribution of a Max test statistic, even under the null hypothesis. Fortunately, as noted by Mackinnon (2006), critical values for a maximum test statistic are readily obtained from bootstrap methods. The details of implementation are presented in the appendix. To minimize the computational burden of collecting bootstrap critical values, the analysis to follow is based on the LM statistic.

3. Application

This section re-examines the predictability of U.S. recessions using actual data. This analysis is presented in 3 steps. First, Section 3.1 describes the data used in the analysis. Section 3.2 presents some preliminary results as a benchmark. Finally, Section 3.3 provides the results of a bootstrap exercise employed to obtain corrected critical values.

3.1 Data

The existing literature has examined the ability of many different variables to predict the onset of recessions; far too many to be examined adequately here. The focus here is on the most widely studied variable; the term spread (TS). The recession indicator is that of the NBER. All data are obtained from the Federal Reserve Economic database, and span the period from the second quarter of 1953 through the second quarter of 2014. Descriptive statistics are presented in Table 1.

Table 1: Descriptive Statistics, 244 quarterly observations.

Variable	Mean	Median	Min	Max	SD	Skewness	Kurtosis
NBER	0.153	0.000	0.000	1.000	0.360	1.936	4.747
TS	1.471	1.465	-1.430	3.800	1.170	-0.073	-0.655

Source: Federal Reserve Economic Database.

3.2 Preliminary results

As a starting point for the analysis of the predictability of recessions in the U.S., the standard approach taken by the existing literature[3] is to test the predictability of a recession for each horizon several horizons; specifically each horizon between 1 and 12 quarters. Simply stated, testing $\beta = 0$ in Equation 1 for several values of k. The results of this exercise are presented in Table 2.

Table 2: Preliminary Results

Horizon	1	2	3	4	5	6	7	8	9	10	11	12
LM	1.95	5.50**	7.52***	8.55***	8.38***	7.30***	5.67**	3.92**	2.68	2.67	2.04	1.82

Notes: The values in the table represent the LM statistic testing the null hypothesis $\beta=0$ from Equation 1 for the corresponding horizon (number of quarters).

[3] The methodology used here is similar to that of Stevens (2014).

The results are consistent with the results of other studies in the literature, in that the null hypothesis is strongly rejected at several (7) horizons. While this appears to offer support for the predictive value of the term spread, the evidence is mixed with only a subset appearing to be significant.

Given the variation in these statistics, the question of interest is a simple one: is the significance of a subset of statistics the result of a real predictive relationship or a product of data snooping? The next section re-evaluates the significance of these statistics based on critical values obtained from the approach presented in the appendix.

3.3 Results

As the main goal of this paper is to demonstrate the need to use correct critical values when evaluating the predictability of recessions (or other events), critical values for 4 sets of horizons found in the existing literature are calculated.

The first is a single horizon of 4 quarters, analogous to that employed[4] by Chauvet and Potter (2005). In this case, data-snooping is not an issue and standard inference techniques remain valid. However, it is important to note that the LM (or other similar) statistic is only distributed as $\chi^2(1)$ asymptotically. Given the finite sample size of most macroeconomic data sets, critical values may (and often do) depart substantially[5] from their standard counterparts. In recognition of this, Table 4 includes the bootstrap critical values for the standard LM statistic to allow the consequences of data-snooping to be accurately quantified.

The second set of interest is each quarter between 1 and 4, which is analogous[6] to the study of Dueker (1997). The third set is to test predictability at each horizon between 1 and 8 quarters, as used by Funke (1997), Estrella and Mishkin (1998), and Moneta (2005). The final set is that of Atta-Mensah and Tkacz (1998) who test the predictability of Canadian recessions at all horizons between 1 and 8, as well as 12 quarters. The results are presented in Table 3.

Table 4: Estimated Critical and P-values for max-LM statistics derived from Table 2.

MLM	1	1-4	1-8	1-8,12
10%	2.925	3.923	4.578	4.811
		(34.1)	(56.5)	(64.6)
5%	3.811	4.770	5.350	5.536
		(25.2)	(40.4)	(45.3)
1%	5.434	6.278	6.706	6.790
		(15.5)	(23.4)	(25.9)
p-value	0.000	0.000	0.000	0.000

Notes: The critical values presented are obtained following the procedure described in the appendix, 10000 replications. The numbers in parenthesis represent the percentage increase

[4] It should be noted that they used monthly data, and tested the predictability of recessions at a single horizon of 12 months.
[5] See Dufour and Khalaf (2001) for a discussion of this point.
[6] He actually used monthly data, and tested predictability at 3, 6, 9 and 12 months.

relative to the benchmark critical values due to data snooping. The p-value represents the proportion of bootstrap test statistics whose value exceeds that of the MLM statistic for each horizon (8.55 in each case).

For each variable of interest, the 10%, 5%, and 1% critical values as well as the p-value are presented for each set of test statistics. The numbers in parenthesis represent the percentage increase in the max-LM statistic over its single value counterpart. For example, the 5% critical value is 25.2% higher than the benchmark when the Max-LM statistic is obtained from a set 4 statistics, and 40.4% higher when 8 statistics are used. The divergence is even more significant at the 10% level, with the critical values increasing by 34.1% and 56.5% when 4 and 8 statistics are used respectively.

In simple terms, the critical values increase as the number of horizons included in the set increase, which is consistent with intuition. The significant inflation in these critical values brought about by data-snooping introduces the possibility that the existing literature may overstate the predictability of recessions.

Having revealed the potential for spurious inference, it is important to note that the predictive value of the term spread is strongly supported, with the null hypothesis rejected at the 1% level in all cases. This finding strongly suggests that the strong evidence of a predictive relationship between the value of the term spread and the onset of recessions is not a product of data snooping.

4. Conclusion

The results presented in this paper demonstrate that treating the time horizon as variable (intentionally or unintentionally) can produce spurious conclusions when inference is based on critical values obtained from standard distributions. With respect to the existing literature, these results demonstrate that reaching a conclusion regarding the predictability of an event based on a collection of test statistics should be avoided.

While the results presented in Sections 2 and 3 demonstrate the potential for spurious conclusions, it is important to note that the use of corrected critical values generally support the results of existing studies, supporting the ability of the term spread to predict recessions.

A useful extension of this method is to evaluate the predictability of recessions across multiple horizons. The method used here only established the significance (or lack of) of the largest test statistic. However, it is obviously possible that the term spread can predict the onset of a recession at multiple horizons. Future research should be directed toward addressing this shortcoming.

References

Andrews, D (1993). Tests for parameter instability and structural change with unknown change point, Econometrica, 61, 821-856.

Atta-Mensah, J and Tkacz, G (1998). Predicting Canadian recessions using financial variables: a probit approach. Bank of Canada working paper, 98-5.

Chauvet, M and Potter, S (2005). Forecasting recessions using the yield curve, Journal of Forecasting, 24, 77-103.

Chen, S-S (2009). Predicting the bear stock Market: Macroeconomic variables as leading indicators. Journal of Banking and Finance, 33, 211-223.

Chow, K and Denning, K (1993). A simple multiple variance ratio test. Journal of Econometrics, 58, 385-401.

Davidson, R and MacKinnon, J (2006). Bootstrap methods in econometrics. T. C. Mills, K. Patterson, eds. Palgrave Handbooks of Econometrics: Vol. 1, Macmillan.

Dueker, M (1997). Strengthening the case for the yield curve as a predictor of U.S. recessions. Federal Reserve Bank of St.Louis Review, March/April, 41-51.

Dufour J-M, and Khalaf, L (2001). Monte Carlo methods in Econometrics, in Companion to Theoretical Econometrics. Oxford: Blackwell, 494-519.

Dufour, J-M. and Taamouti, A (2010). Short and long run causality measured: theory and evidence, *Journal of Econometrics*, 154, 42-56.

Estrella, A. and Mishkin, F. (1998) Predicting U.S. recessions: financial variables as leading indicators, The *Review of Economics and Statistics*, 80, 45-61.

Funke, N (1997). Predicting recessions: some evidence for Germany. Weltwirtschaftliches Archiv, 133, 90-102.

Hansen, B. (1991) Inference when a nuisance parameter is not identified under the null hypothesis, *Rochester Center for Economic Research*, Working Paper 291.

Hansen, B. (1996) Inference when a nuisance parameter is not identified under the null hypothesis, *Econometrica*, 64, 413-430.

Kauppi, H and Saikkonen, P (2008). Predicting recessions with dynamic binary response models. Review of Economics and Statistics, 90, 777-791.

MacKinnon, J (2006). Bootstrap methods in Econometrics, Economic Record, S2-S18.

Moneta, F. (2005). Does the yield spread predict recessions in the euro area? *International finance*, 8, 263-301.

Stevens, J (2014). Identification problems in Granger causality tests based on the net oil price increase. Applied Economics, 46, 102-110.

White, H. (2000) A reality check for data snooping, *Econometrica*, 68, 1097-1126.

Appendix: Corrected Critical Values

Due to the high degree of serial correlation exhibited by the dependent variable, all test statistics are constructed using the same covariance matrix employed by Estrella and Mishkin (1998). The method for testing the null hypothesis that a variable of interest does not predict the occurrence of a recession at any horizon is based on series of simple steps.

First, obtain a test statistic for the null hypothesis $\beta = 0$ for each period of interest. Note that a choice of n periods does not necessarily imply choosing periods 1-4. For example, Dueker (1997) tested horizons of 3, 6, 9 and 12 months. From this set, define the max-LM (MLM) statistic as:

$$MLM = Max\left[LM_1,...,LM_n\right] \tag{A1}$$

While a Probit (or any other binary choice model) is extremely non-linear, a bootstrap is simple to implement[7]; particularly in the case of an LM statistic[8]. Both critical and p-values for the Max-LM statistic are obtained from the following procedure:

For each of N replications:

1) Draw T + M observations (u) from a standard normal distribution, where T is the sample size and M is the number of autoregressive lags to be included in the next step.
2) Construct the following transformation of u:

$$u_t^* = u_t + \sum_{i=1}^{M} \rho_i u_{t-i} \tag{A2}$$

Where M and ρ_i are chosen to replicate the serial correlation observed in the actual dependent variable (under the null hypothesis). In the application presented in the next section, an AR(1) specification is chosen, with $\rho = 0.73$.

3) Construct the simulated dependent variable:

$$Y_t^* = \begin{cases} 1 & \text{if } \hat{\delta} + u_t^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{A3}$$

Where $\hat{\delta}$ is the estimated value of the constant when Equation 1 is estimated under the null hypothesis $\beta=0 \ \forall \ k \in [1,\ldots,n]$.

4) Using the simulated dependent variable, construct a simulated max-LM (SLM) test using the same approach which created the original test statistic.

This procedure produces a set of N simulated max-LM statistics. The α% critical value is simply the αN^{th} largest of the simulated statistics. The p-value of this test statistic is:

$$\frac{\sum_{i=1}^{N} I(SLM_i > MLM)}{N+1} \tag{A4}$$

Where I(·) is an indicator function which takes a value of 1 when the condition is satisfied, and zero otherwise.

[7] See Davidson and MacKinnon (2006) for details.

[8] As noted above, the distribution of this statistics substantially departs from its standard counterpart for n>1.