Asymptotic Variance of Brier (Skill) Score in the Presence of Serial Correlation

Kajal Lahiri
Liu Yang

CESIFO WORKING PAPER NO. 5290
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
ORIGINAL VERSION: MARCH 2015
THIS VERSION: SEPTEMBER 2015

ISSN 2364-1428

CESifo *Working Paper No. 5290*

Asymptotic Variance of Brier (Skill) Score in the Presence of Serial Correlation

Abstract

We derive autocorrelation-robust asymptotic variances of the Brier score and Brier skill score, which are generally applicable in circumstances with weak serial correlation. A simulation experiment and an empirical application from macroeconomics underscore the importance of taking care of serial correlation. We find that the conventional variances are too conservative to account for the sampling variability in estimating the Brier (skill) score.

JEL-Code: C010, C120, C530.

Keywords: probability forecasts, serial correlation, Brier score, Brier skill score, survey of professional forecasters.

Kajal Lahiri
Department of Economics
University at Albany, SUNY
USA – NY 12222
klahiri@albany.edu

*Liu Yang**
School of Economics
Nanjing University
P.R. China – Jiangsu 210093
lyang2@albany.edu

*corresponding author

We thank Daniel Wilks of Cornell’s Earth and Atmospheric Sciences department for some helpful suggestions. A shorter version of this paper will appear in Economics Letters.

1 Introduction

Probability forecasts for binary events are routinely used in diverse fields such as economics, psychology, and meteorology. For example, macroeconomists like to obtain probability forecasts for recessions. Commercial banks are required by regulators to evaluate the probability of defaults associated with loans they have made. Assessing the efficacy of the probability forecasts is of utmost importance to guide decision-making in these contexts. Typically, a number of scores, which are functions of the forecasts and actuals, are employed as evaluation metrics. Lahiri and Yang (2013) provided a survey on these scores from the perspective of economic forecasting. The Brier score (or the quadratic probability score) is probably the most commonly used and is the probabilistic analogue of the mean squared error. A large body of literature on economic forecasting takes the Brier score as the primary statistic to summarize the predictive performance of probability forecasts. Recent examples include Lahiri *et al.* (2013), Levanon *et al.* (2015) and Rudebusch and Williams (2009), only to name a few.

The Brier score is computed on the basis of a given sample on binary events and probability forecasts. Thus, the sampling uncertainty in estimating the score has to be properly accounted for in order to make a statistically meaningful inference. When observations are independently and identically distributed (i.i.d.) across the sample, Bradley *et al.* (2008) derived approximations to the variance of the Brier score and related Brier skill score. In macroeconomic forecasting, it is widely recognized that the sample of forecasts displays positive serial correlation as a consequence of the persistence in economic series, such as real GDP or inflation. One implication of the presence of serial correlation is that the usual sampling variance of the score obtained by assuming independence is no longer valid. This phenomenon has been recognized by Lahiri and Yang (2015) and Pesaran and Timmermann (2009) under various scenarios. Wilks (2010) has shown that the failure to accommodate positive serial correlation will significantly underestimate the standard error of the Brier (skill) score and the magnitude of underestimation depends on the event probability and the quality of the forecast. To correct the effects of serial correlation, Wilks proposed an adjustment factor, whose legitimacy is justified by a simulation experiment. Though informative under his specific data generating process, Wilks' adjusted variance might lose its validity when the underlying process deviates significantly from the assumed process. The main contribution of our paper is to reexamine the issue raised by Wilks (2010) and derive asymptotic variances of the Brier score and Brier skill score in general settings with weak serial correlation.

The remainder of the paper is organized as follows. Section 2 develops the asymptotic variance of the Brier (skill) score when the sample used to produce it is serially dependent. The finite-sample properties are investigated in Section 3. In Section 4, we apply the proposed methodology to examine the quality of the *Survey of Professional Forecasters* (SPF), with respect to its capacity of predicting real GDP declines in the United States. Some concluding remarks are given in Section 5.

2 Asymptotic variances of the Brier score and Brier skill score under weak serial correlation

Throughout the paper, Z is the binary variable to be predicted. When the target event occurs, $Z = 1$, and $Z = 0$ otherwise. P is the probabilistic forecast of the target event. To measure how well the forecast is related to the actual, Brier (1950) developed a score function based on a sequence $\{(Z_t, P_t) : t = 1, ..., T\}$ as $BS \equiv \frac{1}{T}\sum_{t=1}^{T}(Z_t - P_t)^2$. By construction, the Brier score BS lies strictly between 0 and 1, and it has a negative orientation in that lower BS indicates higher accuracy of P. When P exactly coincides with Z, that is, P is a perfect forecast, $BS = 0$. Another forecast, which is often taken as the benchmark, is $P = \pi \equiv P(Z = 1)$. In practice, the population probability $P(Z = 1)$ is rarely known, and thus it is usually replaced by its sample analogue $\bar{Z} \equiv \frac{1}{T}\sum_{t=1}^{T} Z_t$. The Brier score of this naive forecast is denoted as $BS_0 \equiv \frac{1}{T}\sum_{t=1}^{T}(Z_t - \bar{Z})^2$. A real-life forecast with skill is seldom perfect, yet often it can beat the naive benchmark. Consequently, the Brier score of a real-life forecast is often higher than 0 but lower than BS_0.

Sometimes, it is likely to yield a misleading conclusion regarding the performance of P

if we merely look at the Brier score. For example, suppose $P(Z=1)$ is very close to 0, that is, the target event $Z=1$ is rare. In this case, BS_0 could be very close to 0 as well. To see this, let $BS_0 \equiv \frac{1}{T}\sum_{t=1}^{T} Z_t^2 - 2\frac{1}{T}\sum_{t=1}^{T} Z_t\bar{Z} + \bar{Z}^2 = \bar{Z}(1-\bar{Z})$. With a rare event, $\bar{Z}$ is very small, which makes BS_0 quite close to 0. Although the naive forecast seemingly performs fairly well in terms of Brier score, it clearly has no skill because this benchmark cannot distinguish between the occasions when $Z=1$ occurs and those when $Z=1$ does not occur. To circumvent this pitfall of the Brier score in the case of rare events, the Brier skill score can be used. Given any forecast P, the Brier skill score is defined as $BSS \equiv \frac{BS_0 - BS}{BS_0} = 1 - \frac{BS}{BS_0}$ and it is the improvement of the forecast P relative to the naive baseline. The Brier skill score of the benchmark is 0. When P outperforms the benchmark, $BSS > 0$. Otherwise, $BSS < 0$. Other details of BSS can be found in Stephenson (2000).

Define Ω_T to be the covariance matrix of $\frac{1}{\sqrt{T}}(\sum_{t=1}^{T}(Z_t - P_t)^2, \sum_{t=1}^{T}(Z_t - \pi)^2)'$. The ijth component of a matrix A is denoted by A_{ij}. Let $BS^* \equiv E(Z_t - P_t)^2$, $BS_0^* \equiv E(Z_t - \pi)^2 = \pi(1-\pi)$ and $BSS^* \equiv 1 - \frac{BS^*}{BS_0^*}$. The goal of this section is to derive the asymptotic variances of BS and BSS when the sample used to generate them is serially correlated. The following assumptions are sufficient for this purpose.

Assumption 1 *For each t, $P_t \in [0,1]$ and $Z_t \sim$ Bernoulli(π), where $\pi \in (0,1)$.*

Assumption 2 *For some $r' > 1$, the process $\{(Z_t, P_t) : t = 1, \ldots\}$ is a mixing sequence with either uniform mixing coefficient ϕ_m or strong mixing coefficient α_m of size $2r'/(r'-1)$.*

Assumption 3 *(Z_t, P_t) is identically distributed across t.*

Assumption 4 *Ω_T is positive definite for each $T \in N$ and there exists $\varepsilon > 0$ and a natural number $N(\varepsilon)$ such that $|\Omega_T| > \varepsilon$ for all $T > N(\varepsilon)$.*

Assumption 1 rules out the case of non-stochastic Z when $\pi = 0$ or $\pi = 1$, which is of no interest. Assumption 2 allows for a certain degree of serial correlation in the sample, as long as its dependence shrinks towards zero at the stated rate. The population Brier (skill) score is not well defined unless Assumption 3 holds. Assumption 4 is required to ensure the existence of a positive definite long run covariance matrix Ω in Lemma 1.

Lemma 1 *Under Assumptions 1-4, there exists a symmetric positive definite matrix Ω such that $\Omega_T \to \Omega$ as $T \to \infty$.*

Theorem 1 *Under Assumptions 1-4, $\sqrt{T}\begin{pmatrix} BS - BS^* \\ BS_0 - BS_0^* \end{pmatrix} \xrightarrow{d} N(0, \Omega)$.*

Collorary 1 *Under Assumptions 1-4,* $\sqrt{T}(BS - BS^*) \xrightarrow{d} N(0, \Omega_{11})$.

Collorary 2 *Under Assumptions 1-4,* $\sqrt{T}(BSS - BSS^*) \xrightarrow{d} N(0, \frac{1}{BS_0^{*2}}(\Omega_{11} + \frac{BS^{*2}}{BS_0^{*2}}\Omega_{22} - 2\frac{BS^*}{BS_0^*}\Omega_{12}))$.

Colloraries 1 and 2 present the asymptotic distributions of BS and BSS respectively. If the data is independently identically distributed, $\Omega_T = \Omega$, which is the covariance matrix of $((Z_t - P_t)^2, (Z_t - \pi)^2)'$. However, in the presence of serial correlation, $\Omega_T \neq \Omega$ and Ω is the sum of the covariance matrix of $((Z_t - P_t)^2, (Z_t - \pi)^2)'$ and its autocovariance matrices of various orders. For example, $\Omega_{11} = Var((Z_t - P_t)^2) + 2\sum_{m=1}^{\infty} Cov((Z_t - P_t)^2, (Z_{t+m} - P_{t+m})^2)$ in Collorary 1. If the data exhibits positive autocorrelation, $Cov((Z_t - P_t)^2, (Z_{t+m} - P_{t+m})^2) > 0$ for any m, and thus $\Omega_{11} > Var((Z_t - P_t)^2)$. This implies that the conventional variance based on independence assumption could underestimate the true uncertainty in estimating the Brier score, and the degree of underestimation depends on the strength of autocorrelation.

The asymptotic variance of the Brier skill score in Collorary 2 is more complex. Motivated by Diebold and Mariano (1995) and Lopez (2001), an alternative expression of this variance is given by

$$\frac{\Omega_{11} + \Omega_{22}(1 - BSS^*)^2 - 2(1 - BSS^*)\Omega_{12}}{\pi^2(1-\pi)^2},$$

as shown in the appendix. As above, we can write Ω_{ij} as the sum of two parts, that is, $\Omega_{ij} = \Omega_{ij}^1 + \Omega_{ij}^2$, where Ω_{ij}^1 is the variance part and Ω_{ij}^2 is the autocovariance part.[1] Accordingly, the asymptotic variance of BSS is

$$\frac{\Omega_{11}^1 + \Omega_{22}^1(1 - BSS^*)^2 - 2(1 - BSS^*)\Omega_{12}^1}{\pi^2(1-\pi)^2} + \frac{\Omega_{11}^2 + \Omega_{22}^2(1 - BSS^*)^2 - 2(1 - BSS^*)\Omega_{12}^2}{\pi^2(1-\pi)^2}. \quad (1)$$

In view of (1), the inflation in variance due to serial correlation is characterized by $\frac{\Omega_{11}^2 + \Omega_{22}^2(1 - BSS^*)^2 - 2(1 - BSS^*)\Omega_{12}^2}{\pi^2(1-\pi)^2}$, whose magnitude is determined by the event probability π, the quality of the forecast BSS^* and the strength of serial correlation in Ω_{ij}^2. Consistent with the simulation evidence in Wilks (2010), a lower π, other things being equal, induces a larger inflation in variance.

To make use of Colloraries 1 and 2 to conduct statistical inference, the asymptotic vari-

[1] For instance, $\Omega_{11}^2 = 2\sum_{m=1}^{\infty} Cov((Z_t - P_t)^2, (Z_{t+m} - P_{t+m})^2)$.

ances must be estimated. For example, $BS_0 = \bar{Z}(1-\bar{Z})$ can be used to estimate BS_0^*. Similarly, BS^* is estimated by $BS = \frac{1}{T}\sum_{t=1}^{T}(Z_t - P_t)^2$. By Theorem 1, both BS_0 and BS are consistent. The long run covariance matrix Ω can be estimated in different ways, and are well known in the literature. The basic idea is to use a finite sum of sample autocovariance matrices to approximate the population infinite sum, allowing for the truncation lag to increase to infinity at an appropriate rate as the sample size grows. Under some weak regularity conditions, these estimators can be shown to be consistent. See Andrews (1991) and Newey and West (1987, 1994) for details.

3 A simulation experiment

To shed some light on the performance of the autocorrelation-robust variances in finite samples, a Monte Carlo simulation experiment is conducted in this section. Our setup will follow Lahiri and Yang (2015). The data generating process is

$$Z_t^* = \tau + \rho Z_{t-1}^* + \varepsilon_t^Z \text{ and } Y_t^* = \rho Y_{t-1}^* + \varepsilon_t^Y,$$

where ε_t^Z and ε_t^Y are normal white noise and mutually independent. The variance of ε_t^Z is 1, and the variance of ε_t^Y is determined in such a way that $Var(Y_t^*) = 1$. The binary target is obtained by $Z_t = I(Z_t^* > 0)$, and the probability forecast $P_t = \Phi(\mu_{Z_t}^* + Y_t^*)$, where $\mu_1^* = -\mu_{-1}^*$, $I(\cdot)$ is usual indicator function and $\Phi(\cdot)$ is the standard normal distribution function. The parameter τ is calibrated so that the corresponding probabilities of the target event are equal to various values. For each combination of parameter values (τ, ρ, μ_1^*), we simulate 1000 Monte Carlo replications of the processes. We consider samples of size T = 100, 200 and 500.

To construct confidence intervals, the asymptotic covariance matrix Ω must be estimated first. Suppose we treat the sample as i.i.d., as is often done in practice. The asymptotic covariance matrix is Ω_T, which can be estimated by the sample covariance matrix of $((Z_t - P_t)^2, (Z_t - \pi)^2)'$. When serial correlation is accommodated, we use Andrews (1991) quadratic spectral kernel estimator to approximate the long run variance Ω.[2] All computations are performed in the R system with the aid of functions in the package **sandwich**.

Table 1 summarizes the empirical coverage rates of two asymptotic 95% confidence intervals of BS and BSS when $\mu_1^* = 1$.[3] It is clear from the table that the conventional confidence intervals, by ignoring the serial correlation, suffer from severe bias in that its coverage rates are substantially lower than the nominal level when $\rho \neq 0$. For example, when $\rho = 0.7$, these intervals include the true values in only 75 out of 100 cases on average. In contrast, our robust intervals are uniformly superior in the presence of serial correlation. The improvement over the independent intervals rises with ρ. In the case of independence $(\rho = 0)$, two types of intervals show little difference. However, the coverage rates of the robust intervals are all above 85% with relatively strong correlation $(\rho = 0.7)$. As expected, a larger sample size leads to a smaller bias in estimating the asymptotic variance of the Brier (Skill) Score with our robust procedure. However, large T seems to have little help in reducing the bias associated with the independent intervals. Holding other parameters fixed, the performance of all intervals deteriorates as the target event becomes rarer. This makes sense as fewer observation with $Z_t = 1$ means less information content to be exploited. An interesting finding is that the independent intervals for BSS appear to be better than those for BS with serial correlation. This empirical evidence agrees with Wilks (2010), which found that the sampling variance of BSS is more robust to serial correlation than that of BS. Despite this smaller distortion, the independent intervals of BSS are still strictly dominated by our robust intervals as long as the serial correlation exists.

Table 2 parallels to Table 1 with $\mu_1^* = 2$.[4] Similar to what was found in Wilks (2010), in this scenario of high predictive accuracy, virtually all of the intervals perform more poorly. In other words, the impact of serial correlation on the sampling variabilities of BS and BSS gets stronger. Both intervals are not wide enough to completely reflect the larger variabilities. Only with a large sample $(T = 500)$ can the coverage rates of the robust intervals exceed 90% even when $\rho = 0.5$.

[2] Prior to calculating the robust intervals, the data is filtered by AR(1) prewhitening procedure as advocated by Andrew and Monahan (1992).

[3] The corresponding BSS is equal to 0.547.

[4] The corresponding BSS is equal to 0.907.

Table 1: Empirical coverage probabilities when $BSS = 0.547$

	$\pi^* = 0.05$		$\pi^* = 0.1$		$\pi^* = 0.3$		$\pi^* = 0.4$	
	Ind.	Corr.	Ind.	Corr.	Ind.	Corr.	Ind.	Corr.
BS	T=100							
$\rho = 0.0$	0.941	0.935	0.936	0.928	0.934	0.933	0.937	0.934
$\rho = 0.5$	0.787	0.893	0.798	0.896	0.849	0.908	0.872	0.918
$\rho = 0.7$	0.643	0.856	0.673	0.867	0.745	0.885	0.771	0.889
BSS	T=100							
$\rho = 0.0$	0.918	0.915	0.933	0.931	0.943	0.936	0.944	0.938
$\rho = 0.5$	0.861	0.900	0.850	0.893	0.864	0.915	0.886	0.931
$\rho = 0.7$	0.787	0.861	0.794	0.869	0.773	0.901	0.790	0.909
BS	T=200							
$\rho = 0.0$	0.942	0.941	0.945	0.942	0.947	0.945	0.944	0.940
$\rho = 0.5$	0.800	0.920	0.818	0.918	0.863	0.924	0.880	0.929
$\rho = 0.7$	0.658	0.906	0.874	0.889	0.756	0.904	0.776	0.913
BSS	T=200							
$\rho = 0.0$	0.937	0.937	0.947	0.946	0.949	0.947	0.947	0.945
$\rho = 0.5$	0.882	0.926	0.871	0.920	0.863	0.929	0.883	0.937
$\rho = 0.7$	0.826	0.907	0.791	0.890	0.774	0.912	0.779	0.924
BS	T=500							
$\rho = 0.0$	0.949	0.947	0.945	0.942	0.951	0.949	0.948	0.947
$\rho = 0.5$	0.782	0.919	0.826	0.932	0.872	0.940	0.883	0.941
$\rho = 0.7$	0.663	0.909	0.682	0.908	0.753	0.919	0.779	0.924
BSS	T=500							
$\rho = 0.0$	0.944	0.943	0.948	0.948	0.950	0.949	0.948	0.947
$\rho = 0.5$	0.884	0.930	0.853	0.929	0.864	0.939	0.879	0.943
$\rho = 0.7$	0.809	0.910	0.736	0.906	0.749	0.920	0.777	0.926

Notes: The columns “Ind.” and “Corr.” contain empirical coverage probabilities for the independent and Andrews’ HAC-based autocorrelation-robust 95% confidence intervals respectively.

Table 2: Empirical coverage probabilities when $BSS = 0.907$

	$\pi^* = 0.05$		$\pi^* = 0.1$		$\pi^* = 0.3$		$\pi^* = 0.4$	
	Ind.	Corr.	Ind.	Corr.	Ind.	Corr.	Ind.	Corr.
BS	T=100							
$\rho = 0.0$	0.879	0.878	0.887	0.886	0.879	0.878	0.881	0.878
$\rho = 0.5$	0.772	0.825	0.779	0.828	0.815	0.847	0.820	0.846
$\rho = 0.7$	0.645	0.775	0.665	0.778	0.706	0.795	0.716	0.801
BSS	T=100							
$\rho = 0.0$	0.900	0.898	0.904	0.901	0.890	0.886	0.889	0.884
$\rho = 0.5$	0.830	0.858	0.826	0.856	0.832	0.864	0.830	0.858
$\rho = 0.7$	0.766	0.825	0.763	0.823	0.738	0.823	0.740	0.822
BS	T=200							
$\rho = 0.0$	0.916	0.914	0.914	0.912	0.912	0.910	0.912	0.912
$\rho = 0.5$	0.818	0.893	0.821	0.877	0.848	0.886	0.854	0.890
$\rho = 0.7$	0.690	0.862	0.688	0.829	0.746	0.851	0.754	0.847
BSS	T=200							
$\rho = 0.0$	0.923	0.920	0.924	0.923	0.920	0.917	0.916	0.914
$\rho = 0.5$	0.871	0.911	0.857	0.897	0.856	0.896	0.858	0.895
$\rho = 0.7$	0.808	0.887	0.764	0.857	0.757	0.862	0.761	0.862
BS	T=500							
$\rho = 0.0$	0.935	0.934	0.934	0.934	0.934	0.932	0.936	0.934
$\rho = 0.5$	0.835	0.910	0.842	0.915	0.868	0.920	0.879	0.923
$\rho = 0.7$	0.697	0.877	0.716	0.880	0.763	0.892	0.774	0.895
BSS	T=500							
$\rho = 0.0$	0.940	0.938	0.942	0.941	0.937	0.936	0.938	0.936
$\rho = 0.5$	0.878	0.924	0.856	0.920	0.867	0.921	0.883	0.926
$\rho = 0.7$	0.788	0.896	0.741	0.888	0.760	0.894	0.775	0.899

Notes: The columns “Ind.” and “Corr.” contain empirical coverage probabilities for the independent and Andrews’ HAC-based autocorrelation-robust 95% confidence intervals respectively.

In summary, our robust confidence interval corrects for the presence of serial correlation and thus they are strictly preferred to the conventional counterpart. In terms of its finite-sample coverage rates, the robust interval is still subject to a moderate magnitude of bias, which increases with ρ and *BSS*, and decreases with T and π^*. To further reduce, if not eliminate, this bias, we conjecture that an alternative method to estimate the long run covariance matrix Ω might be useful. See Lahiri and Yang (2015) and Sun (2013, 2014) for details.

4 An empirical illustration

In this section, we will use the Brier score and Brier skill score to assess the performance of the probability forecasts of real GDP declines recorded in the *Survey of Professional Forecasters*. The main purpose here is to compare the asymptotic confidence intervals of BS and BSS using our autocorrelation-robust variances with those constructed by assuming independence. In each quarter, the respondents of this survey are asked to indicate the probability they would attach to a decline in the level of real GDP in the current and the next four quarters. The target variable is the same for all of these five horizons. Our sample spans from 1968:Q4 to 2015:Q1. During this period, the fraction of real GDP declines is about 12.9%, meaning that it is a relatively uncommon event. Lahiri and Wang (2013) carried out a comprehensive evaluation on the accuracy of these subjective forecasts.

Our robust analysis is motivated by Figure 1, which presents the autocorrelation functions of SPF forecasts and actuals. Clearly, all series display positive autocorrelation. All of the autocorrelation coefficients up to three-quarter lags are significantly different from zero. Table 3 shows BS and BSS for SPF forecasts. As horizon rises, the performance of professional forecasters deteriorates, as reflected by rising BS and declining BSS. In this example, the Brier score of the naive benchmark is about 0.116. As a result, the four-quarter-ahead forecasts are found to be even worse than this benchmark, as is obvious from its negative BSS. Two types of 95% confidence intervals of BS and BSS are also given in Table 3. By ignoring the positive serial correlation in Figure 1, the independent intervals are uniformly narrower than the autocorrelation-robust intervals. Notably, for Q2, the former approach suggests significantly positive skill, which is disputed by the appropriate autocorrelation-robust 95% confidence interval.

Table 3: Brier (skill) scores and their 95% confidence intervals

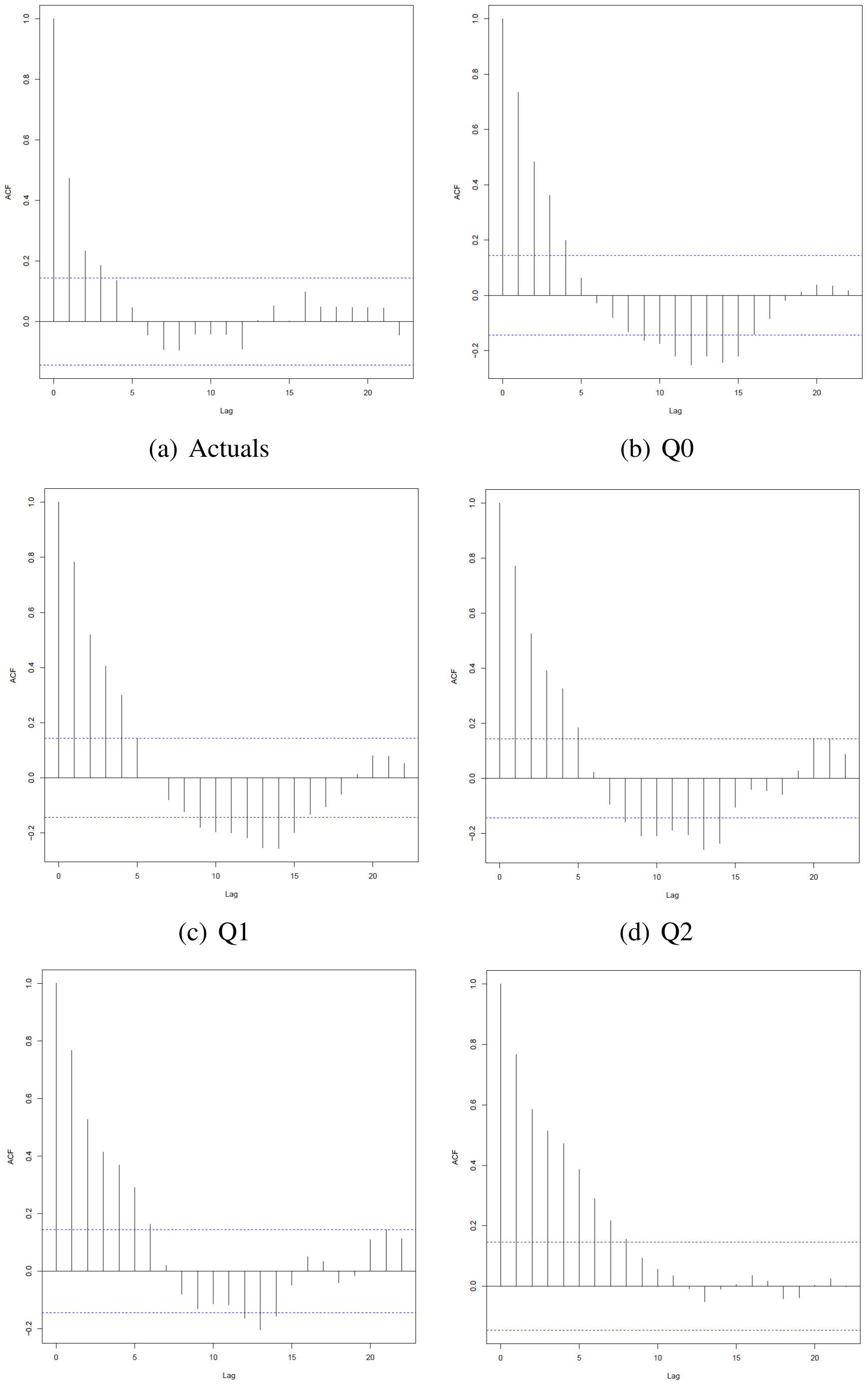
Statistic	Q0	Q1	Q2	Q3	Q4
BS	0.062	0.083	0.098	0.115	0.119
Ind.	(0.042,0.082)	(0.062,0.105)	(0.073,0.124)	(0.084,0.146)	(0.086,0.152)
Corr.	(0.035,0.090)	(0.049,0.117)	(0.053,0.143)	(0.064,0.166)	(0.070,0.168)
BSS	0.446	0.259	0.126	0.013	-0.040
Ind.	(0.258,0.635)	(0.094,0.425)	(0.006,0.245)	(-0.056,0.082)	(-0.109,0.030)
Corr.	(0.196,0.696)	(0.042,0.477)	(-0.006,0.258)	(-0.078,0.104)	(-0.139,0.059)

Notes: "Ind." is the 95% confidence interval based on independence assumption. "Corr." is the 95% confidence interval with Ω being estimated by Andrew's approach.

5 Conclusions

This paper addresses the problem of correcting the effects of serial correlation on the sampling properties of the Brier (skill) score, initially investigated by Wilks (2010). The proposed asymptotic variance is more general and thus applicable in circumstances with weak serial correlation. Using a simulation experiment and an empirical example with SPF probability forecasts, we confirm that by ignoring the positive serial correlation, the conventional variance is too conservative to account for the sampling uncertainty in estimating the Brier (skill) score.

Figure 1: Autocorrelation functions of SPF and Actuals



(a) Actuals

(b) Q0

(c) Q1

(d) Q2

(e) Q3

(f) Q4

Notes: The dotted lines are 95% confidence band about the zero line. “Qi” is the ith quarter SPF forecast (i=current, 1, 2, 3, 4).

Mathematical Appendix

Proof (Lemma 1): This lemma can be shown following the same reasoning in Lemma 1 of Lahiri and Yang (2015). □

Proof (Theorem 1): According to the central limit theorem for mixing sequences (cf. White (2000)), $\frac{1}{\sqrt{T}}\begin{pmatrix} \sum_{t=1}^{T}((Z_t-P_t)^2-E(Z_t-P_t)^2) \\ \sum_{t=1}^{T}((Z_t-\pi)^2-E(Z_t-\pi)^2) \end{pmatrix}$ converges in distribution to $N(0,\Omega)$. Since $\sqrt{T}(BS_0-BS_0^*)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}((Z_t-\bar{Z})^2-E(Z_t-\pi)^2)$, we have

$$\begin{aligned} & \sqrt{T}(BS_0-BS_0^*)-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}((Z_t-\pi)^2-E(Z_t-\pi)^2) \\ = & \frac{1}{\sqrt{T}}\sum_{t=1}^{T}((Z_t-\bar{Z})^2-E(Z_t-\pi)^2)-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}((Z_t-\pi)^2-E(Z_t-\pi)^2) \\ = & \frac{1}{\sqrt{T}}\sum_{t=1}^{T}((Z_t-\bar{Z})^2-(Z_t-\pi)^2) \\ = & \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(2Z_t(\pi-\bar{Z})+\bar{Z}^2-\pi^2) \\ = & (\bar{Z}-\pi)(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(-2Z_t+\bar{Z}+\pi)) \\ = & \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(Z_t-\pi)(\pi-\bar{Z}). \end{aligned}$$

Since $\pi-\bar{Z}=o_p(1)$ by law of large number and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(Z_t-\pi)=O_p(1)$ by central limit theorem, $\sqrt{T}(BS_0-BS_0^*)-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}((Z_t-\pi)^2-E(Z_t-\pi)^2)=o_p(1)$. Consequently, $\sqrt{T}\begin{pmatrix} BS-BS^* \\ BS_0-BS_0^* \end{pmatrix}\xrightarrow{d}N(0,\Omega)$. □

Proof (Collorary 1): It is a direct consequence of Theorem 1. □

Proof (Collorary 2): Note that $BSS=1-\frac{BS}{BS_0}$ and $BSS^*=1-\frac{BS^*}{BS_0^*}$. The result follows from Theorem 1 and the Delta method. □

An alternative proof (Collorary 2): Note that the two competing forecasts to be compared are P_t and $\bar{Z}$. By Diebold and Mariano (1995), $\sqrt{T}(\frac{1}{T}\sum_{t=1}^{T}((Z_t-\bar{Z})^2-(Z_t-P_t)^2)-E(Z_t-\pi)^2+E(Z_t-P_t)^2)\xrightarrow{d}N(0,\Omega_{11}+\Omega_{22}-2\Omega_{12})$. Following the same line of Theorem 1, we have

$$\sqrt{T}\begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T}((Z_t-\bar{Z})^2-(Z_t-P_t)^2)-E(Z_t-\pi)^2+E(Z_t-P_t)^2 \\ \frac{1}{T}\sum_{t=1}^{T}(Z_t-\bar{Z})^2-E(Z_t-\pi)^2 \end{pmatrix} \xrightarrow{d} N(0,\tilde{\Omega}),$$

where $\tilde{\Omega}=\begin{pmatrix} \Omega_{11}+\Omega_{22}-2\Omega_{12} & \Omega_{22}-\Omega_{12} \\ \Omega_{22}-\Omega_{12} & \Omega_{22} \end{pmatrix}$. By the Delta method,

$$\sqrt{T}(BSS-BSS^*) \xrightarrow{d} N(0,\frac{1}{BS_0^{*2}}(1,-BSS^*)\tilde{\Omega}(1,-BSS^*)').$$

The result follows by noting that

$$\begin{aligned}
&\frac{1}{BS_0^{*2}}(1,-BSS^*)\tilde{\Omega}(1,-BSS^*)' = \frac{\Omega_{11}+\Omega_{22}(1-BSS^*)^2-2(1-BSS^*)\Omega_{12}}{\pi^2(1-\pi)^2} \\
= \quad & \frac{\Omega_{11}+\Omega_{22}-2\Omega_{12}+BSS^{*2}\Omega_{22}-2BSS^*\Omega_{22}+2BSS^*\Omega_{12}}{\pi^2(1-\pi)^2} \\
= \quad & \frac{1}{BS_0^{*2}}(\Omega_{11}+\Omega_{22}-2\Omega_{12}+\Omega_{22}-2\frac{BS^*}{BS_0^*}\Omega_{22}+\frac{BS^{*2}}{BS_0^{*2}}\Omega_{22}-2\Omega_{22}+2\Omega_{12}+2\frac{BS^*}{BS_0^*}\Omega_{22}-2\frac{BS^*}{BS_0^*}\Omega_{12}) \\
= \quad & \frac{1}{BS_0^{*2}}(\Omega_{11}+\frac{BS^{*2}}{BS_0^{*2}}\Omega_{22}-2\frac{BS^*}{BS_0^*}\Omega_{12}).
\end{aligned}$$

□

R Code to Compute Asymptotic Variance of the Brier (Skill) Score

$install.packages(”RODBC”)$

$install.packages(”sandwich”)$

$channel < -odbcConnectExcel(”Datapath”)$

$spf < -sqlFetch(channel,”sheetname”)$

$P < -spf[,1]$

$Z < -spf[,2]$

$obs < -nrow(spf)$

$alpha < -0.05$

$BS < -mean((Z-P)^2)$

$BS0 < -mean(Z)*(1-mean(Z))$

$BSS < -1-BS/BS0$

$Gradient < -matrix(nrow = obs, ncol = 2)$

$for(iin1 : obs)Gradient[i,] = c((Z[i]-P[i])^2,(Z[i]-mean(Z))^2)$

$AsyCovI = cov(Gradient)/obs$

$AsyCov = lrvar(Gradient, type = ”Andrews”, prewhite = TRUE, adjust = TRUE)$

$c(BS - qnorm(1 - alpha/2) * sqrt(AsyCovI[1,1]), BS + qnorm(1 - alpha/2) * sqrt(AsyCovI[1,1]))$

$c(BS - qnorm(1 - alpha/2) * sqrt(AsyCov[1,1]), BS + qnorm(1 - alpha/2) * sqrt(AsyCov[1,1]))$

$c(BSS - qnorm(1-alpha/2)*sqrt(1/BS0^2*(AsyCovI[1,1]+(BS/BS0)^2*AsyCovI[2,2]-2*BS/BS0*AsyCovI[1,2])), BSS+qnorm(1-alpha/2)*sqrt(1/BS0^2*(AsyCovI[1,1]+(BS/BS0)^2*AsyCovI[2,2]-2*BS/BS0*AsyCovI[1,2])))$

$c(BSS - qnorm(1-alpha/2)*sqrt(1/BS0^2*(AsyCov[1,1]+(BS/BS0)^2*AsyCov[2,2]-2*BS/BS0*AsyCov[1,2])), BSS+qnorm(1-alpha/2)*sqrt(1/BS0^2*(AsyCov[1,1]+(BS/BS0)^2*AsyCov[2,2]-2*BS/BS0*AsyCov[1,2])))$

References

Andrews, D. W. K. (1991), 'Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation', *Econometrica* **59**, 817–858.

Andrews, D. W. K. and Monahan, J. C. (1992), 'An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator', *Econometrica* **60**, 953–966.

Bradley, A. A., Schwartz, S. S. and Hashino, T. (2008), 'Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score', *Weather and Forecasting* **23**, 992–1006.

Brier, G. W. (1950), 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review* **78**, 1–3.

Diebold, F. X. and Mariano, R. S. (1995), 'Comparing Predictive Accuracy', *Journal of Business & Economic Statistics* **13**, 253–263.

Lahiri, K., Monokroussos, G. and Zhao, Y. (2013), 'The Yield Spread Puzzle and the Information Content of SPF Forecasts', *Economics Letters* **118**, 219–221.

Lahiri, K. and Wang, J. G. (2013), 'Evaluating Probability Forecasts for GDP Declines using Alternative Methodologies', *International Journal of Forecasting* **29**, 175–190.

Lahiri, K. and Yang, L. (2013), Forecasting Binary Outcomes, *in* A. Timmermann and G. Elliott, eds, 'Handbook of Economic Forecasting Volume 2B', North-Holland Amsterdam, pp. 1025–1106.

Lahiri, K. and Yang, L. (2015), 'Confidence Bands for ROC Curves with Serially Dependent Data', *Journal of Business & Economic Statistics* forthcoming.

Levanon, G., Manini, J., Ozyildirim, A., Schaitkin, B. and Tanchua, J. (2015), 'Using Financial Indicators to Predict Turning Points in the Business Cycle: The Case of the Leading Economic Index for the United States', *International Journal of Forecasting* **31**, 426–445.

Lopez, J. A. (2001), ‘Evaluating the Predictive Accuracy of Volatility Models’, *Journal of Forecasting* **20**, 87–109.

Newey, W. K. and West, K. D. (1987), ‘A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix’, *Econometrica* **55**, 703–708.

Newey, W. K. and West, K. D. (1994), ‘Automatic Lag Selection in Covariance Matrix Estimation’, *Review of Economic Studies* **61**, 631–653.

Pesaran, M. H. and Timmermann, A. (2009), ‘Testing Dependence among Serially Correlated Multi-Category Variables’, *Journal of the American Statistical Association* **104**, 325–337.

Rudebusch, G. D. and Williams, J. C. (2009), ‘Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve’, *Journal of Business & Economic Statistics* **27**, 492–503.

Stephenson, D. B. (2000), ‘Use of the ‘Odds Ratio’ for Diagnosing Forecast Skill’, *Weather Forecasting* **15**, 221–232.

Sun, Y. (2013), ‘A Heteroskedasticity and Autocorrelation Robust F Test using an Orthonormal Series Variance Estimator’, *Econometrics Journal* **16**, 1–26.

Sun, Y. (2014), ‘Let’s Fix it: Fixed-b Asymptotics versus Small-b Asymptotics in Heteroscedasticity and Autocorrelation Robust Inference’, *Journal of Econometrics* **178**, 659–677.

White, H. (2000), *Asymptotic Theory for Econometricians*, Academic Press.

Wilks, D. S. (2010), ‘Sampling Distributions of the Brier Score and Brier Skill Score under Serial Dependence’, *Quarterly Journal of the Royal Meteorological Society* **136**, 2109–2118.