A Single-Sample Estimate of Shrinkage in Meteorological Forecasting

PAUL W. MIELKE JR.

Department of Statistics, Colorado State University, Fort Collins, Colorado

KENNETH J. BERRY

Department of Sociology, Colorado State University, Fort Collins, Colorado

CHRISTOPHER W. LANDSEA*

NOAA Climate and Global Change Fellowship, NOAA/AOML/Hurricane Research Division, Miami, Florida

WILLIAM M. GRAY

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

(Manuscript received 20 June 1996, in final form 20 May 1997)

ABSTRACT

An estimator of shrinkage based on information contained in a single sample is presented and the results of a simulation study are reported. The effects of sample size, amount, and severity of nonrepresentative data in the population, inclusion of noninformative predictors, and least (sum of) absolute deviations and least (sum of) squared deviations regression models are examined on the estimator. A single-sample estimator of shrinkage based on drop-one cross-validation is shown to be highly accurate under a wide variety of research conditions.

1. Introduction

Meteorologists have long recognized the importance of accurately quantifying statistical forecast skill. One of the primary tools of meteorological forecasting is multiple regression analysis (Murphy and Winkler 1984) where, given data on a response variable y_i and associated predictor variables x_{ij}, where $j = 1, \ldots, p$; $i = 1, \ldots, n$; p denotes the number of predictors; and n represents the number of events; the goal is to find some function of the x_{ij} values that is an accurate and precise predictor of y_i. It is generally recognized that any estimate of forecast skill grounded in a multiple regression model that is based on a sample of observations is characteristically higher than the forecast skill that would be obtained from a multiple regression model that is based on the entire population of observations (Mosteller and Tukey 1977; Picard and Cook 1984; Michaelsen 1987; Barnston and Van den Dool 1993). It is also widely accepted that the fit of the multiple regression model to new sample data is nearly always less precise than the fit of the same multiple regression model to the original sample data on which the model was based. This is reflected in lower forecast skill levels obtained when sample-based multiple regression models are used to predict future events.

It is useful to have elementary terms to distinguish between the fit of a multiple regression model to the sample data on which the model has been determined and the fit of the same multiple regression model to an independent sample of data. The former is termed ''retrospective'' fit and the latter is termed ''validation'' fit (Copas 1983). The term ''shrinkage'' denotes the drop in skill from retrospective fit to validation fit (Copas 1983) and indicates how useful the sample-based regression coefficients will be for prediction on other datasets. For purposes of clarification, shrinkage involves the following four-step procedure. First, a multiple regression model is fit to a sample dataset by optimizing the regression coefficients relative to a fitting criterion, for example, least squares. Second, the goodness of fit of the multiple regression model is measured by an index, such as a squared multiple correlation coefficient. Third, the obtained multiple regression model is applied to an independent sample dataset and a second goodness-of-fit index is obtained for the independent dataset.

* Current affiliation: NOAA/AOML/Hurricane Research Division, Miami, Florida.

Corresponding author address: Dr. Paul W. Mielke Jr., Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877. E-mail: mielke@lamar.colostate.edu

Fourth, a ratio of the two indices is constructed where the goodness-of-fit index from the original dataset is the denominator. This ratio is termed shrinkage since it is usually less than unity.

Mielke et al. (1996) investigated the effects of sample size, type of regression model, and noise-to-signal ratio on the degree of shrinkage in five populations that differed in the amount and degree of contaminated data. Shrinkage was defined as the ratio of the validation fit of a sample regression equation to the retrospective fit of the same sample regression equation where the validation fit was assessed on five independent samples, averaged over 10 000 simulations. While this index of shrinkage is both rigorous and comprehensive, the use of six independent samples precludes its use in routine research situations. In this paper, an estimate of shrinkage is developed that is based on a single sample and can easily be employed by research meteorologists. Comparisons with the index of shrinkage given by Mielke et al. (1996) indicate that the single-sample estimate of shrinkage is very accurate under a wide variety of conditions. The single-sample estimate of shrinkage is related to cross-validation methods that have become standard for assessing the predictive validity of forecast skill.

2. Cross-validation

Historically, users of multiple regression procedures have developed methods to assess how accurately sample regression coefficients estimate the corresponding population regression coefficients. The usual procedure is to test the sample regression coefficients on an independent set of sample data. This practice has come to be known as ''cross-validation.'' A comprehensive historical background on cross-validation is provided by Stone (1974, 1978), Geisser (1975), Mosteller and Tukey (1977), and Snee (1977). Camstra and Boomsma (1992) present an extensive overview of the use of cross-validation in regression, where the emphasis is on the prediction of individual observations, and in covariance structure analysis, where the emphasis is on future values of variances and covariances. Michaelsen (1987) and Elsner and Schmertmann (1994) describe and discuss cross-validation methods as they pertain to meteorological forecasting.

It is widely recognized that to be useful, any sample regression equation must hold for data other than those on which the regression equation was developed. When sample data are used to determine the regression coefficients that best predict the response variable from the set of predictor variables, assuming that the variables to be used in the regression equation have already been selected, prediction performance is usually overestimated (Picard and Cook 1984). Because the sample regression coefficients are determined by an optimizing process that is conditioned on the sample data, the regression equation generally provides better predictions for the sample data on which it is based than for any other dataset. This is sometimes referred to as ''testing on the training data'' (Glick 1978). It should be noted that the use of cross-validation precludes any manipulation of the dataset prior to the development of the regression model and subsequent cross-validation.

In general, cross-validation consists of determining the regression coefficients in one sample and applying the obtained coefficients to the predictor scores of another sample. The initial sample is termed the ''calibration'' or ''training'' sample and the second sample is called the ''validation'' or ''test'' sample (Browne 1975a,b; Huberty et al. 1987; Camstra and Boomsma 1992; MacCallum et al. 1994). The calibration sample is used to calculate the regression coefficients, and the predictive validity of the fitted equation is verified on the validation sample.

As defined, cross-validation requires two samples. Because a second sample is often not readily available, an alternative approach is often used in which a large sample is randomly split into two subsamples. One subsample is specified as the calibration sample and the second sample is designated the validation sample. The many problems associated with this approach to cross-validation are summarized in Lachenbruch and Mickey (1968), Picard and Cook (1984), and Picard and Berk (1990). Setting aside the obvious loss of information in splitting samples (Browne and Cudeck 1992), a significant problem is the difficulty in procuring large samples, which are not available in many research situations. In addition, when calibration sample sizes are small, the regression coefficients are less precise than those that would be obtained if the entire sample had been used (Horst 1966). Mosier (1951) suggested a double cross-validation procedure where the regression coefficients are calculated for both the calibration and validation samples and the two regression equations are cross-validated on the sample that was not used to establish the regression coefficients. Questions have been raised as to exactly what should be done when the results of the two cross-validations differ (Snee 1977). It has been suggested that if the two sets of regression coefficients are not too different, then a new set of coefficients may be obtained from the combined calibration and validation samples (Mosier 1951). While no estimate of predictive validity is available for the combined sample, Mosier (1951) posited that it may be approximated by the average of the predictive validities obtained for the original calibration and validation samples.

Cross-validation is certainly not limited to just two samples. The data can be divided into more than two samples and multiple cross-validations can be obtained. Multiple cross-validation involves partitioning an available sample of size n into a calibration sample of size $n - k$ and a validation sample of size k. The cross-validation procedure is realized by withholding each validation sample of size k, calculating a regression

TABLE 1. Population 1: Initial population consisting of 3958 noncontaminated events. Columns are (C1) true population ρ values (C2) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, (C3) average of five sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C2, (C4) average of 10 000 drop-one $\hat{\rho}$ values estimated for each of the 10 000 samples of C2, and corresponding ratios (C3/C2), (C4/C2), (C4/C3), and (C3/C1).

Sample size	Case	Model	C1	C2	C3	C4	C3/C2	C4/C2	C4/C3	C3/C1
15	10	LAD	0.51495	0.83216	0.21959	0.19520	0.264	0.235	0.889	0.426
		LSD	0.51154	0.76579	0.24721	0.20980	0.323	0.277	0.849	0.483
	6	LAD	0.51130	0.69947	0.32214	0.29055	0.461	0.415	0.902	0.630
		LSD	0.50917	0.64059	0.34883	0.31267	0.545	0.488	0.896	0.685
25	10	LAD	0.51495	0.69659	0.34693	0.33158	0.498	0.476	0.956	0.674
		LSD	0.51154	0.63931	0.37427	0.35416	0.585	0.554	0.946	0.732
	6	LAD	0.51130	0.61963	0.39741	0.37757	0.641	0.609	0.950	0.777
		LSD	0.50917	0.57839	0.41795	0.39942	0.723	0.691	0.956	0.821
40	10	LAD	0.51495	0.62613	0.41279	0.40342	0.659	0.644	0.977	0.802
		LSD	0.51154	0.58533	0.43132	0.42250	0.737	0.722	0.980	0.843
	6	LAD	0.51130	0.57687	0.43965	0.42829	0.762	0.742	0.974	0.860
		LSD	0.50917	0.54955	0.45455	0.44493	0.827	0.810	0.988	0.893
65	10	LAD	0.51495	0.58265	0.45361	0.44876	0.779	0.770	0.989	0.881
		LSD	0.51154	0.55438	0.46425	0.45954	0.837	0.829	0.990	0.908
	6	LAD	0.51130	0.55102	0.46701	0.46115	0.848	0.837	0.987	0.913
		LSD	0.50917	0.53274	0.47611	0.47099	0.894	0.884	0.989	0.935
100	10	LAD	0.51495	0.55843	0.47627	0.47168	0.853	0.845	0.990	0.925
		LSD	0.51154	0.53790	0.48182	0.47803	0.896	0.889	0.992	0.942
	6	LAD	0.51130	0.53651	0.48269	0.47818	0.900	0.891	0.991	0.944
		LSD	0.50917	0.52353	0.48814	0.48429	0.932	0.925	0.992	0.959
160	10	LAD	0.51495	0.54290	0.49184	0.48998	0.906	0.903	0.996	0.955
		LSD	0.51154	0.52759	0.49302	0.49183	0.934	0.932	0.998	0.965
	6	LAD	0.51130	0.52727	0.49364	0.49187	0.936	0.933	0.996	0.965
		LSD	0.50917	0.51780	0.49598	0.49463	0.958	0.955	0.997	0.974
250	10	LAD	0.51495	0.53325	0.50076	0.50011	0.939	0.938	0.999	0.972
		LSD	0.51154	0.52141	0.49982	0.49965	0.959	0.958	1.000	0.977
	6	LAD	0.51130	0.52160	0.50012	0.50004	0.959	0.959	1.000	0.978
		LSD	0.50917	0.51454	0.50081	0.50080	0.973	0.973	1.000	0.984
500	10	LAD	0.51495	0.52527	0.50865	0.50660	0.968	0.964	0.996	0.988
		LSD	0.51154	0.51661	0.50562	0.50355	0.985	0.975	0.996	0.988
	6	LAD	0.51130	0.51685	0.50578	0.50422	0.979	0.976	0.997	0.989
		LSD	0.50917	0.51206	0.50500	0.50294	0.986	0.982	0.991	0.992

model from the remaining calibration sample of size $n - k$, and validating each of the $\binom{n}{k}$ possible regression models on the remaining sample of size k held in reserve. Since $k = 1$ requires validating only n regression models on the remaining sample of size 1 held in reserve, this special case is both easily implemented and commonly used. In various literature, the case where $k = 1$ is termed drop-one cross-validation, leave-one-out cross-validation, hold-one-out cross-validation, or the U method. Stone (1978) provides a thorough review of drop-one cross-validation. Drop-one cross-validation is an exhaustive method involving substantial redundancy in the participation of each data point (far more redundancy when $k > 1$). However, the exhaustive features of drop-one cross-validation may provide a comprehensive evaluation of predictive accuracy and a solid estimate of predictive skill (Barnston and Van den Dool 1993).

Drop-one cross-validation is usually credited to Lachenbruch (1967) or Lachenbruch and Mickey (1968). However, Toussaint (1974) has traced the drop-one method to earlier sources under different names (Glick 1978). Currently, the drop-one method is the cross-validation procedure of choice and it is not unusual to see the term cross-validation virtually equated with the drop-one method (e.g., Nicholls 1985; Livezey et al. 1990).

For many researchers, the method of choice for cross-validation is to create a model on one sample and test the model on a second sample drawn from the same population; alternatively, a model is created on a substantial portion of a sample and tested on the remaining portion of the sample. In either case, the selection of predictors can be based on information in the population or some other out-of-sample source, or the selection of predictors can involve subset selection based on in-sample information. In addition, the regression coefficients are nearly always based on information in the calibration sample. Much of the early work in cross-validation specifically limited analyses to fixed models where the number and variety of predictors is determined a priori and not based on subset selection (e.g., Browne 1975a, 1975b; Camstra and Boomsma 1992; MacCallum et al. 1994). Thus, cross-validation in this context implies validation of the sample regression coefficients only. In those cases where subset selection is based on the sample information, cross-validation implies validation of

TABLE 2. Population 2: Contaminated population consisting of 3998 events consisting of the initial population of 3958 events and 40 moderately extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, (C3) average of five sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C2, (C4) average of 10 000 drop-one $\hat{\rho}$ values estimated for each of the 10 000 samples of C2, and corresponding ratios (C3/C2), (C4/C2), (C4/C3), and (C3/C1).

Sample size	Case	Model	C1	C2	C3	C4	C3/C2	C4/C2	C4/C3	C3/C1
15	10	LAD	0.48886	0.83077	0.20662	0.18590	0.249	0.224	0.900	0.423
		LSD	0.45120	0.76387	0.23315	0.20030	0.305	0.262	0.859	0.517
	6	LAD	0.48220	0.69081	0.30215	0.28085	0.437	0.407	0.930	0.627
		LSD	0.44984	0.63008	0.32887	0.29919	0.522	0.475	0.910	0.731
25	10	LAD	0.48886	0.69099	0.32943	0.31682	0.477	0.459	0.962	0.674
		LSD	0.45120	0.63311	0.35703	0.33872	0.564	0.535	0.949	0.791
	6	LAD	0.48220	0.60467	0.37074	0.35322	0.613	0.584	0.953	0.769
		LSD	0.44984	0.56220	0.39249	0.37486	0.698	0.667	0.955	0.873
40	10	LAD	0.48886	0.61657	0.38947	0.38038	0.632	0.617	0.977	0.797
		LSD	0.45120	0.57548	0.40904	0.39914	0.711	0.694	0.976	0.907
	6	LAD	0.48220	0.55632	0.40805	0.39739	0.733	0.714	0.974	0.846
		LSD	0.44984	0.52658	0.42208	0.41204	0.802	0.782	0.976	0.938
65	10	LAD	0.48886	0.56622	0.42434	0.41918	0.749	0.740	0.988	0.868
		LSD	0.45120	0.53715	0.43587	0.43264	0.811	0.805	0.993	0.966
	6	LAD	0.48220	0.52586	0.43418	0.42843	0.826	0.815	0.987	0.900
		LSD	0.44984	0.50036	0.43615	0.43242	0.872	0.864	0.991	0.970
100	10	LAD	0.48886	0.53914	0.44413	0.43980	0.824	0.816	0.990	0.909
		LSD	0.45120	0.51555	0.44819	0.44515	0.869	0.863	0.993	0.993
	6	LAD	0.48220	0.51103	0.45081	0.44633	0.882	0.873	0.990	0.935
		LSD	0.44984	0.48556	0.44265	0.43940	0.912	0.905	0.993	0.984
160	10	LAD	0.48886	0.51867	0.45922	0.45747	0.885	0.882	0.996	0.939
		LSD	0.45120	0.49590	0.45266	0.45185	0.913	0.911	0.998	1.003
	6	LAD	0.48220	0.49982	0.46294	0.46082	0.926	0.922	0.995	0.960
		LSD	0.44984	0.47246	0.44554	0.44477	0.943	0.941	0.998	0.990
250	10	LAD	0.48886	0.50767	0.46914	0.46727	0.924	0.920	0.996	0.960
		LSD	0.45120	0.48257	0.45410	0.45271	0.941	0.938	0.997	1.006
	6	LAD	0.48220	0.49408	0.47041	0.46900	0.952	0.949	0.997	0.964
		LSD	0.44984	0.46486	0.44727	0.44579	0.962	0.959	0.997	0.994
500	10	LAD	0.48886	0.49896	0.47922	0.47885	0.960	0.960	0.999	0.980
		LSD	0.45120	0.46904	0.45367	0.45338	0.967	0.967	0.999	1.005
	6	LAD	0.48220	0.48999	0.47780	0.47740	0.975	0.974	0.999	0.991
		LSD	0.44984	0.45838	0.44897	0.44878	0.979	0.979	1.000	0.998

the subset selection process *and* the sample regression coefficients.

The advent of double cross-validation brought additional complications. Given fixed predictors, the regression coefficients from each sample are tested on the other sample and any differences can be consolidated by some form of weighted averaging of the regression coefficients (Subrahmanyam 1972). However, given sample-based subset selection, there is the added complication that each sample will select a different number and/or a different set of predictors. It is much more difficult to resolve discrepancies between the two sample validation results. Browne (1970) provides results of random sampling experiments demonstrating the effects of not fixing the predictors beforehand. With drop-one cross-validation it is possible to conceive of up to *n* different but overlapping sets of predictors and up to *n* different values for the regression coefficients for each predictor. The satisfactory and optimal combining of these differences appears very difficult; see, for example, Browne and Cudeck (1989) and MacCallum et al. (1994).

Cross-validation is not without its critics and there is evidence that suggests some possible drawbacks to drop-one cross-validation. Glick (1978) and Hora and Wilcox (1982) provide simulation studies of drop-one cross-validation in discriminant analysis. Both studies indicate that the estimates have relatively high variability over repeated sampling, possibly due to the repeated use of the original data. The results of both Glick (1978) and Hora and Wilcox (1982) were based on discriminant analysis, which has a binary error function. Efron (1983) notes that cross-validation performs somewhat better given a smooth residual sum of squares error function. Finally, some investigators note that a model that fits the validation sample as well as the calibration sample is not necessarily a validated model. Maltz (1994), for example, argues that cross-validation may only show that the procedure used to split the sample did, in fact, divide the sample into two similar subgroups.

If a specific sample dataset exhibits a high first-order autoregressive pattern, drop-one cross-validation may overestimate the validation fit. For example, if a single sample consists of cases selected from a time series, then the cases in a given cycle (e.g., a month, a year,

TABLE 3. Population 3: Contaminated population of 3998 events of the initial population of 3958 events and 40 very extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, (C3) average of five sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C2, (C4) average of 10 000 drop-one $\hat{\rho}$ values estimated for each of the 10 000 samples of C2, and corresponding ratios (C3/C2), (C4/C2), (C4/C3), and (C3/C1).

Sample size	Case	Model	C1	C2	C3	C4	C3/C2	C4/C2	C4/C3	C3/C1
15	10	LAD	0.44873	0.83121	0.20082	0.18153	0.242	0.218	0.904	0.448
		LSD	0.29776	0.76468	0.22665	0.19527	0.296	0.255	0.862	0.761
	6	LAD	0.43722	0.69002	0.29357	0.27530	0.425	0.399	0.938	0.671
		LSD	0.27225	0.62930	0.31866	0.29310	0.506	0.466	0.920	1.170
25	10	LAD	0.44873	0.69172	0.31845	0.30758	0.460	0.445	0.966	0.710
		LSD	0.29776	0.63445	0.34486	0.32807	0.544	0.517	0.951	1.158
	6	LAD	0.43722	0.60065	0.35366	0.34098	0.589	0.568	0.964	0.809
		LSD	0.27225	0.55827	0.37435	0.36124	0.671	0.647	0.965	1.375
40	10	LAD	0.44873	0.61802	0.37532	0.36669	0.607	0.593	0.977	0.836
		LSD	0.29776	0.57769	0.39366	0.38507	0.681	0.667	0.978	1.322
	6	LAD	0.43722	0.54451	0.37879	0.37480	0.696	0.688	0.989	0.866
		LSD	0.27225	0.51698	0.39520	0.39098	0.764	0.756	0.989	1.452
65	10	LAD	0.44873	0.56667	0.40517	0.40388	0.715	0.713	0.997	0.903
		LSD	0.29776	0.53918	0.41701	0.41612	0.773	0.772	0.998	1.400
	6	LAD	0.43722	0.49965	0.38841	0.38981	0.777	0.780	1.004	0.888
		LSD	0.27225	0.47894	0.39580	0.39865	0.826	0.832	1.007	1.454
100	10	LAD	0.44873	0.53523	0.41757	0.41608	0.780	0.777	0.996	0.931
		LSD	0.29776	0.51541	0.42518	0.42465	0.825	0.824	0.999	1.428
	6	LAD	0.43722	0.47492	0.39691	0.39539	0.836	0.833	0.996	0.908
		LSD	0.27225	0.44794	0.38686	0.38850	0.864	0.867	1.004	1.421
160	10	LAD	0.44873	0.50458	0.42088	0.42080	0.834	0.834	1.000	0.938
		LSD	0.29776	0.48765	0.42113	0.42239	0.864	0.866	1.003	1.414
	6	LAD	0.43722	0.45708	0.40649	0.40508	0.889	0.886	0.997	0.930
		LSD	0.27225	0.41056	0.36825	0.37137	0.897	0.905	1.008	1.353
250	10	LAD	0.44873	0.48421	0.42473	0.42132	0.877	0.870	0.992	0.947
		LSD	0.29776	0.45964	0.40951	0.40740	0.891	0.886	0.995	1.375
	6	LAD	0.43722	0.44837	0.41433	0.41141	0.924	0.918	0.993	0.948
		LSD	0.27225	0.37856	0.34883	0.34694	0.921	0.916	0.995	1.281
500	10	LAD	0.44873	0.46656	0.43340	0.43378	0.929	0.930	1.001	0.966
		LSD	0.29776	0.40724	0.37716	0.37488	0.926	0.921	0.994	1.267
	6	LAD	0.43722	0.44242	0.42437	0.42319	0.959	0.957	0.997	0.971
		LSD	0.27225	0.33445	0.31859	0.31641	0.953	0.946	0.993	1.170

or a decade) may be highly correlated. In such cases a drop-k cross-validation may be required to mitigate the cyclic pattern, where k exceeds the length of the cycle. Michaelsen (1987) has researched the effects of autoregressive effects on cross-validation in statistical climate forecast models.

3. Statistical measures

Let the population and sample sizes be denoted by N and n, respectively, let y_i denote the response variable, and let $x_{i1}, \ldots, x_{ip}$ denote the p predictor variables associated with the ith of n events. Consider the linear regression model given by

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i,$$

where $\beta_0, \ldots, \beta_p$ are $p + 1$ unknown parameters and e_i is the error term associated with the ith of n events. Two types of regression models are of interest: least (sum of) absolute deviations (LAD) regression models and least (sum of) squared deviations (LSD) regression models. The LAD and LSD prediction equations are given by

$$\tilde{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j x_{ij},$$

where $\tilde{y}_i$ is the predicted value of y_i and $\tilde{\beta}_0, \ldots, \tilde{\beta}_p$ minimize the expression

$$\sum_{i=1}^{n} |e_i|^v$$

with $v = 1$ and $v = 2$ associated with the LAD and LSD regression models, respectively.

A measure of agreement is employed to determine the correspondence between the y_i and $\tilde{y}_i$ values, for $i = 1, \ldots, n$. Many researchers have utilized measures of agreement in assessing prediction accuracy, for example, Willmott (1982), Willmott et al. (1985), Kelly et al. (1989), Tucker et al. (1989), Gray et al. (1992), McCabe and Legates (1992), Badescu (1993), Elsner and Schmertmann (1993), Hess and Elsner (1994), Cotton et al. (1994), and Lee et al. (1995). Watterson (1996)

TABLE 4. Population 4: Contaminated population of 4158 events of the initial population of 3958 events and 200 moderately extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, (C3) average of five sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C2, (C4) average of 10 000 drop-one $\hat{\rho}$ values estimated for each of the 10 000 samples of C2, and corresponding ratios (C3/C2), (C4/C2), (C4/C3), and (C3/C1).

Sample size	Case	Model	C1	C2	C3	C4	C3/C2	C4/C2	C4/C3	C3/C1
15	10	LAD	0.36924	0.82319	0.17630	0.16583	0.214	0.201	0.941	0.477
		LSD	0.31192	0.75362	0.19886	0.17653	0.242	0.234	0.888	0.638
	6	LAD	0.36698	0.66307	0.24413	0.23173	0.368	0.349	0.949	0.665
		LSD	0.30599	0.59807	0.26934	0.24821	0.450	0.415	0.922	0.880
25	10	LAD	0.36924	0.66658	0.26978	0.26159	0.405	0.392	0.970	0.731
		LSD	0.31192	0.60568	0.29496	0.28121	0.487	0.464	0.953	0.946
	6	LAD	0.36698	0.54840	0.28455	0.27086	0.519	0.494	0.952	0.775
		LSD	0.30599	0.50205	0.30704	0.29335	0.612	0.584	0.955	1.003
40	10	LAD	0.36924	0.57417	0.30939	0.30583	0.539	0.533	0.988	0.838
		LSD	0.31192	0.52983	0.33079	0.32601	0.624	0.615	0.986	1.060
	6	LAD	0.36698	0.48241	0.30925	0.29728	0.641	0.616	0.961	0.843
		LSD	0.30599	0.44501	0.32037	0.31236	0.720	0.702	0.975	1.047
65	10	LAD	0.36924	0.50310	0.33020	0.32513	0.656	0.646	0.985	0.894
		LSD	0.31192	0.46771	0.34046	0.33620	0.728	0.719	0.987	1.091
	6	LAD	0.36698	0.43752	0.32591	0.31565	0.745	0.721	0.967	0.888
		LSD	0.30599	0.39896	0.32084	0.31322	0.804	0.785	0.976	1.049
100	10	LAD	0.36924	0.45960	0.34149	0.33438	0.743	0.728	0.979	0.925
		LSD	0.31192	0.42314	0.33801	0.33372	0.799	0.772	0.987	1.084
	6	LAD	0.36698	0.41285	0.33750	0.32911	0.817	0.797	0.975	0.920
		LSD	0.30599	0.36963	0.31813	0.31226	0.861	0.845	0.982	1.040
160	10	LAD	0.36924	0.42656	0.34929	0.34577	0.819	0.811	0.990	0.946
		LSD	0.31192	0.38503	0.33172	0.32978	0.862	0.857	0.994	1.063
	6	LAD	0.36698	0.39421	0.34556	0.34131	0.877	0.866	0.988	0.942
		LSD	0.30599	0.34668	0.31471	0.31177	0.908	0.899	0.991	1.028
250	10	LAD	0.36924	0.40505	0.35441	0.35257	0.875	0.870	0.995	0.960
		LSD	0.31192	0.35909	0.32608	0.32541	0.908	0.906	0.998	1.045
	6	LAD	0.36698	0.38328	0.35161	0.34934	0.917	0.911	0.994	0.958
		LSD	0.30599	0.33194	0.31188	0.31066	0.940	0.936	0.996	1.019
500	10	LAD	0.36924	0.38708	0.36071	0.35695	0.932	0.922	0.990	0.977
		LSD	0.31192	0.33582	0.32013	0.31692	0.953	0.944	0.990	1.026
	6	LAD	0.36698	0.37622	0.36011	0.35656	0.957	0.948	0.990	0.981
		LSD	0.30599	0.31932	0.30961	0.30651	0.970	0.960	0.990	1.012

provides a comprehensive comparison of various measures of agreement.

In this simulation study, the measure of agreement for both the LAD and LSD prediction equations is given by

$$\rho = 1 - \frac{\delta}{\mu_\delta},$$

where

$$\delta = \frac{1}{n}\sum_{i=1}^{n} |y_i - \tilde{y}_i|$$

and μ_δ is the average value of δ over all $n!$ equally likely permutations of $y_1, \ldots, y_n$ relative to $\tilde{y}_1, \ldots, \tilde{y}_n$ under the null hypothesis that the n pairs (y_i and $\tilde{y}_i$ for $i = 1, \ldots, n$) are merely the result of random assignment. This reduces to the simple computational form given by

$$\mu_\delta = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |y_i - \tilde{y}_j|.$$

Since ρ is a chance-corrected measure of agreement, $\rho = 1.0$ implies that all paired values of y_i and $\tilde{y}_i$ for $i = 1, \ldots, n$ fall on a line with unit slope that passes through the origin (i.e., a perfect forecast). The choice of ρ rather than the Pearson product–moment correlation coefficient (r) or r^2 (the coefficient of determination) is that the latter are measures of linearity and not measures of agreement. Also, the choice of the mean absolute error for δ rather than the mean squared error is that extreme values influence the squared Euclidean differences of the MSE far more than the Euclidean differences of the MAE. These choices are elaborated by Mielke et al. (1996).

4. Data and simulation procedures

The present study investigates the accuracy and utility of a single-sample estimator of shrinkage. Also considered are the effects of sample size, type of regression model (LAD and LSD), and noise-to-signal ratio in five populations that differ in amount and degree of contaminated data. Sample sizes (n) of 15, 25, 40, 65, 100, 160, 250, and 500 events are obtained from a fixed population of $N = 3958$ events, which, for the purpose of this study, is not contaminated with extreme cases;

TABLE 5. Population 5: Contaminated population of 4158 events of the initial population of 3958 events and 200 very extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, (C3) average of five sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C2, (C4) average of 10 000 drop-one $\hat{\rho}$ values estimated for each of the 10 000 samples of C2, and corresponding ratios (C3/C2), (C4/C2), (C4/C3), and (C3/C1).

Sample size	Case	Model	C1	C2	C3	C4	C3/C2	C4/C2	C4/C3	C3/C1
15	10	LAD	0.16541	0.82410	0.15671	0.14784	0.190	0.179	0.943	0.947
		LSD	0.13645	0.75637	0.17684	0.15684	0.234	0.207	0.887	1.296
	6	LAD	0.10284	0.65648	0.21212	0.21141	0.323	0.322	0.997	2.063
		LSD	0.08999	0.59195	0.23220	0.22483	0.392	0.380	0.968	2.580
25	10	LAD	0.16541	0.67046	0.23397	0.22978	0.349	0.343	0.982	1.414
		LSD	0.13645	0.61289	0.25467	0.24579	0.416	0.401	0.965	1.866
	6	LAD	0.10284	0.52745	0.22739	0.22995	0.431	0.436	1.011	2.211
		LSD	0.08999	0.48267	0.24628	0.24774	0.467	0.513	1.006	2.737
40	10	LAD	0.16541	0.57754	0.26244	0.26566	0.454	0.460	1.012	1.587
		LSD	0.13645	0.53669	0.27945	0.28149	0.521	0.524	1.007	2.048
	6	LAD	0.10284	0.42604	0.21735	0.22235	0.510	0.522	1.023	2.113
		LSD	0.08999	0.39821	0.23444	0.24104	0.589	0.605	1.028	2.605
65	10	LAD	0.16541	0.48744	0.26077	0.26526	0.535	0.544	1.017	1.577
		LSD	0.13645	0.46297	0.27581	0.27978	0.596	0.604	1.014	2.021
	6	LAD	0.10284	0.33718	0.19921	0.19730	0.591	0.585	0.990	1.937
		LSD	0.08999	0.30874	0.20295	0.20390	0.657	0.660	1.005	2.255
100	10	LAD	0.16541	0.40913	0.24710	0.24839	0.604	0.607	1.005	1.494
		LSD	0.13645	0.38873	0.25578	0.25838	0.658	0.665	1.010	1.875
	6	LAD	0.10284	0.27822	0.18593	0.18110	0.668	0.651	0.974	1.808
		LSD	0.08999	0.23511	0.17010	0.16837	0.723	0.716	0.990	1.890
160	10	LAD	0.16541	0.33696	0.23219	0.23301	0.689	0.692	1.004	1.404
		LSD	0.13645	0.30317	0.22288	0.22643	0.735	0.747	1.016	1.633
	6	LAD	0.10284	0.23075	0.17241	0.16991	0.747	0.736	0.985	1.676
		LSD	0.08999	0.17580	0.14063	0.13848	0.800	0.788	0.985	1.563
250	10	LAD	0.16541	0.28582	0.21974	0.22088	0.769	0.773	1.005	1.328
		LSD	0.13645	0.23539	0.19305	0.19305	0.820	0.820	1.000	1.415
	6	LAD	0.10284	0.19584	0.15837	0.15498	0.809	0.791	0.979	1.540
		LSD	0.08999	0.13849	0.11975	0.11718	0.865	0.846	0.979	1.331
500	10	LAD	0.16541	0.23315	0.20148	0.19745	0.864	0.847	0.980	1.218
		LSD	0.13645	0.17699	0.16198	0.15951	0.915	0.901	0.985	1.187
	6	LAD	0.10284	0.15889	0.14052	0.13570	0.884	0.854	0.966	1.366
		LSD	0.08999	0.11112	0.10326	0.09952	0.929	0.896	0.964	1.147

a fixed population of $N = 3998$ events consisting of the initial population and 40 moderately extreme events (1% moderate contamination); a fixed population of $N = 3998$ events consisting of the initial population and 40 very extreme events (1% severe contamination); a fixed population of $N = 4158$ events consisting of the initial population and 200 moderately extreme events (5% moderate contamination); and a fixed population of $N = 4158$ events consisting of the initial population and 200 very extreme events (5% severe contamination). The 3958 available primary events used to construct each of the five populations used in this study consist of a response variable and $p = 10$ predictor variables. Specifics of the meteorological data used to construct these five populations are given in Mielke et al. (1996).

TABLE 6. Probability of no contaminated values in each sample of size n.

	Contamination	
Sample size (n)	1%	5%
15	0.8600	0.4774
25	0.7777	0.2916
40	0.6688	0.1392
65	0.5202	0.0406
100	0.3658	0.0072
160	0.2001	0.0004
250	0.0810	4.4×10^{-6}
500	0.0066	2.0×10^{-11}

Two prediction models are considered for each of the five populations. The first prediction model (case 10) consists of $p = 10$ independent variables, and the second prediction model (case 6) consists of $p = 6$ independent variables. In case 10, 4 of the 10 independent variables in the initial population of $N = 3958$ events were found to contribute no information to the predictions. Case 6 is merely the prediction model with the four noninformative independent variables of case 10 deleted. Both the case 10 and case 6 prediction models were constructed from the initial fixed population of $N = 3958$ events. The reason for the two prediction models is to examine the effect of including noninformative independent variables (i.e., noise) in a prediction model.

TABLE 7. Population 1: Initial population consisting of 3958 noncontaminated events. The SD($\hat{\rho}$|C2) column contains the standard deviations of the 10 000 $\hat{\rho}$ values composing each C2 value in Table 1. The SD($\hat{\rho}$|C4) column contains the standard deviations of the 10 000 drop-one $\hat{\rho}$ values composing each C4 value in Table 1.

Sample size	Case	Model	SD($\hat{\rho}$|C2)	SD($\hat{\rho}$|C4)
15	10	LAD	0.07205	0.21811
		LSD	0.10145	0.20870
	6	LAD	0.10536	0.23722
		LSD	0.12495	0.21354
25	10	LAD	0.07976	0.19051
		LSD	0.09381	0.15925
	6	LAD	0.09650	0.18076
		LSD	0.10405	0.15057
40	10	LAD	0.07542	0.14298
		LSD	0.08140	0.11277
	6	LAD	0.08507	0.13386
		LSD	0.08608	0.10771
65	10	LAD	0.06600	0.10259
		LSD	0.06736	0.08098
	6	LAD	0.07015	0.09610
		LSD	0.06946	0.07882
100	10	LAD	0.05482	0.07727
		LSD	0.05479	0.06276
	6	LAD	0.05717	0.07272
		LSD	0.05582	0.06118
160	10	LAD	0.04579	0.05734
		LSD	0.04508	0.04804
	6	LAD	0.04702	0.05450
		LSD	0.04545	0.04737
250	10	LAD	0.03696	0.04148
		LSD	0.03620	0.03692
	6	LAD	0.03762	0.04071
		LSD	0.03636	0.03651
500	10	LAD	0.02695	0.02921
		LSD	0.02622	0.02684
	6	LAD	0.02717	0.02884
		LSD	0.02615	0.02679

TABLE 8. Population 2: Contaminated population of 3998 events consisting of the initial population of 3958 events and 40 moderately extreme events. The SD($\hat{\rho}$|C2) column contains the standard deviations of the 10 000 $\hat{\rho}$ values composing each C2 value in Table 2. The SD($\hat{\rho}$|C4) column contains the standard deviations of the 10 000 drop-one $\hat{\rho}$ values composing each C4 value in Table 2.

Sample size	Case	Model	SD($\hat{\rho}$|C2)	SD($\hat{\rho}$|C4)
15	10	LAD	0.07354	0.21597
		LSD	0.10359	0.20627
	6	LAD	0.10890	0.23385
		LSD	0.12936	0.21345
25	10	LAD	0.08266	0.18894
		LSD	0.09707	0.15777
	6	LAD	0.10427	0.18602
		LSD	0.11167	0.15612
40	10	LAD	0.07913	0.14657
		LSD	0.08435	0.11843
	6	LAD	0.09503	0.14472
		LSD	0.09713	0.12155
65	10	LAD	0.07164	0.11311
		LSD	0.07299	0.09119
	6	LAD	0.08132	0.11164
		LSD	0.08418	0.09778
100	10	LAD	0.06286	0.08795
		LSD	0.06408	0.07473
	6	LAD	0.06886	0.08581
		LSD	0.07336	0.08112
160	10	LAD	0.05246	0.06648
		LSD	0.05481	0.06168
	6	LAD	0.05509	0.06478
		LSD	0.06151	0.06593
250	10	LAD	0.04379	0.05167
		LSD	0.04796	0.05224
	6	LAD	0.04497	0.04962
		LSD	0.05194	0.05471
500	10	LAD	0.03150	0.03451
		LSD	0.03697	0.03729
	6	LAD	0.03189	0.03289
		LSD	0.03864	0.03872

5. Findings and discussion

The results of the study are summarized in Tables 1–5. In Tables 1–5, each row is specified by 1) a sample size (n), 2) $p = 10$ (case 10) and $p = 6$ (case 6) independent samples, and 3) LAD and LSD regression analyses. In each of the five tables the first column (C1) contains the true ρ values for the designated population and the second column (C2) contains the average of 10 000 randomly obtained sample estimates of ρ, $\hat{\rho}$, where the $\tilde{y}$ values are based on the sample regression coefficients for each of the 10 000 independent samples, that is, a measure of *retrospective fit.* The third column (C3) measures the effectiveness of validating sample regression coefficients. In this column the sample regression coefficients from 10 000 random samples were first obtained from column C2, then for each of these 10 000 sets of sample regression coefficients an additional five independent random samples of the same size ($n = 15, \ldots, 500$) were drawn from the population. The sample regression coefficients from C2 were then applied to each of the five new samples, and $\hat{\rho}$ values were computed for each of these five samples for a total of 50 000 $\hat{\rho}$ values. The average of the 50 000 $\hat{\rho}$ values is reported in column C3, yielding a measure of *validation fit.* The fourth column (C4) contains the average of 10 000 randomly obtained drop-one sample $\hat{\rho}$ values where each of the $\hat{\rho}$ values is based on the same sample data that yields one of the 10 000 sample $\hat{\rho}$ values composing the averages in column C2. Thus, each value in column C4 represents the average of n times 10 000 $\hat{\rho}$ values. The fifth column (C3/C2) contains the ratio of the average $\hat{\rho}$ value of C3 to the corresponding $\hat{\rho}$ value of C2, that is, the index of *shrinkage.* The sixth column (C4/C2) contains the ratio of the average $\hat{\rho}$ value of C4 to the average $\hat{\rho}$ value of C2, that is, the drop-one single-sample estimator of shrinkage, as measured by C3/C2. The seventh column (C4/C3) contains the ratio of the average $\hat{\rho}$ value of C4/C2 to the average $\hat{\rho}$ value of C3/C2, that is, the ratio of the drop-one single-sample estimator of shrinkage to the index of shrinkage. The eighth column (C3/C1) contains the ratio of the validation fit of C3 to the corresponding true fit, measured by the population ρ value given in C1. The values of

TABLE 9. Population 3: Contaminated population consisting of 3998 events consisting of the initial population of 3958 events and 40 very extreme events. The SD($\hat{\rho}$|C2) column contains the standard deviations of the 10 000 $\hat{\rho}$ values composing each C2 value in Table 3. The SD($\hat{\rho}$|C4) column contains the standard deviations of the 10 000 drop-one $\hat{\rho}$ values composing each C4 value in Table 3.

Sample size	Case	Model	SD($\hat{\rho}$|C2)	SD($\hat{\rho}$|C4)
15	10	LAD	0.07312	0.21340
		LSD	0.10296	0.20394
	6	LAD	0.10986	0.23235
		LSD	0.13057	0.21374
25	10	LAD	0.08271	0.18859
		LSD	0.09680	0.16023
	6	LAD	0.10943	0.19101
		LSD	0.11702	0.16443
40	10	LAD	0.07881	0.15062
		LSD	0.08353	0.12591
	6	LAD	0.10850	0.16098
		LSD	0.10854	0.14063
65	10	LAD	0.07311	0.12111
		LSD	0.07326	0.10234
	6	LAD	0.10760	0.14048
		LSD	0.10912	0.13000
100	10	LAD	0.07011	0.10276
		LSD	0.06856	0.09096
	6	LAD	0.09916	0.12123
		LSD	0.10987	0.12467
160	10	LAD	0.06532	0.09020
		LSD	0.06718	0.08629
	6	LAD	0.08366	0.09932
		LSD	0.10760	0.11722
250	10	LAD	0.06024	0.07645
		LSD	0.07112	0.08589
	6	LAD	0.07153	0.08012
		LSD	0.10170	0.10799
500	10	LAD	0.04760	0.05133
		LSD	0.07220	0.07339
	6	LAD	0.05308	0.05542
		LSD	0.08288	0.08283

TABLE 10. Population 4: Contaminated population consisting of 4158 events consisting of the initial population of 3958 events and 200 moderately extreme events. The SD($\hat{\rho}$|C2) column contains the standard deviations of the 10 000 $\hat{\rho}$ values composing each C2 value in Table 4. The SD($\hat{\rho}$|C4) column contains the standard deviations of the 10 000 drop-one $\hat{\rho}$ values composing each C4 value in Table 4.

Sample size	Case	Model	SD($\hat{\rho}$|C2)	SD($\hat{\rho}$|C4)
15	10	LAD	0.07562	0.20328
		LSD	0.10652	0.19460
	6	LAD	0.12349	0.23447
		LSD	0.14319	0.21392
25	10	LAD	0.09434	0.18918
		LSD	0.10862	0.16428
	6	LAD	0.13018	0.19988
		LSD	0.13574	0.17395
40	10	LAD	0.09796	0.16053
		LSD	0.10279	0.13597
	6	LAD	0.12286	0.16902
		LSD	0.12368	0.14667
65	10	LAD	0.09569	0.13179
		LSD	0.09648	0.11449
	6	LAD	0.11052	0.13891
		LSD	0.10641	0.11780
100	10	LAD	0.08759	0.10968
		LSD	0.08668	0.09591
	6	LAD	0.09751	0.11473
		LSD	0.09047	0.09552
160	10	LAD	0.07687	0.08986
		LSD	0.07262	0.07640
	6	LAD	0.08342	0.09390
		LSD	0.07305	0.07508
250	10	LAD	0.06777	0.07482
		LSD	0.05996	0.06017
	6	LAD	0.07159	0.07611
		LSD	0.05947	0.05962
500	10	LAD	0.05184	0.05465
		LSD	0.04164	0.04147
	6	LAD	0.05314	0.05635
		LSD	0.04148	0.04148

columns C1, C2, C3, C3/C2, and C3/C1 are contained in Mielke et al. (1996).

It should be noted in this context that both C3 and C4 are free from any selection bias. Selection bias occurs when a subset of predictor variables is selected from the full set of predictor variables in the population based on information contained in the sample. In this study, selection bias has been controlled by selecting the two sets of predictor variables (i.e., cases 10 and 6) from information contained in the population and not from information contained in any sample. Specifically, in the case of C3, the predictor variables were selected from information in the population, the regression coefficients were based on information contained in the sample for these (10 or 6) predetermined predictor variables, then the regression coefficients were applied to five new independent samples of the same size and drawn from the same population. This process was repeated for 10 000 samples, producing 50 000 $\hat{\rho}$ values. Each C3 value is an average of these 50 000 $\hat{\rho}$ values. Thus, while there is an optimizing bias due to retrospective fit, there is no selection bias. In the case of C4, the predictor variables were again selected from information contained in the population and the regression coefficients were based on information contained in the sample, after dropping one observation. A $\hat{\rho}$ value was calculated on the set of $n - 1$ y and $\tilde{y}$ values, and the procedure was repeated n times, dropping a different observation each time. The entire process was repeated for 10 000 samples, producing n times 10 000 $\hat{\rho}$ values. Each C4 value is an average of these n times 10 000 $\hat{\rho}$ values. Thus, there is no selection bias. The advantage to this approach is that the optimizing bias can be isolated and examined while the selection bias is controlled. In addition, this approach is more conservative as validation fit is almost always better when subset selection is included (MacCallum et al. 1994). The drawback to this approach is that the results cannot be generalized to studies that selected both prediction variables and regression coefficients based on sample information and, in addition, shrinkage may be increased.

The ratio values in column C3/C2 in Tables 1–5 provide a comprehensive index of shrinkage that serves as a benchmark against which the accuracy of the drop-

TABLE 11. Population 5: Contaminated population consisting of 4158 events consisting of the initial population of 3958 events and 200 very extreme events. The SD($\hat{\rho}$|C2) column contains the standard deviations of the 10 000 $\hat{\rho}$ values composing each C2 value in Table 5. The SD($\hat{\rho}$|C4) column contains the standard deviations of the 10 000 drop-one $\hat{\rho}$ values composing each C4 value in Table 5.

Sample size	Case	Model	SD($\hat{\rho}$|C2)	SD($\hat{\rho}$|C4)
15	10	LAD	0.07494	0.19074
		LSD	0.10476	0.18378
	6	LAD	0.12873	0.22530
		LSD	0.14899	0.20967
25	10	LAD	0.09339	0.17836
		LSD	0.10667	0.16084
	6	LAD	0.14751	0.20186
		LSD	0.15292	0.15169
40	10	LAD	0.10061	0.15998
		LSD	0.10400	0.14360
	6	LAD	0.15666	0.18604
		LSD	0.15555	0.17129
65	10	LAD	0.11479	0.14447
		LSD	0.11073	0.13048
	6	LAD	0.15084	0.15964
		LSD	0.14880	0.14895
100	10	LAD	0.11799	0.12954
		LSD	0.11559	0.12049
	6	LAD	0.13551	0.13828
		LSD	0.12614	0.12206
160	10	LAD	0.10895	0.11034
		LSD	0.10701	0.10334
	6	LAD	0.11705	0.11682
		LSD	0.09116	0.08748
250	10	LAD	0.09543	0.09467
		LSD	0.08361	0.07840
	6	LAD	0.10046	0.09831
		LSD	0.06166	0.05949
500	10	LAD	0.07177	0.06644
		LSD	0.04617	0.04218
	6	LAD	0.07518	0.07223
		LSD	0.03362	0.03134

one single-sample estimator of shrinkage given in column C4/C2 can be measured. The ratio values in column C4/C3 were obtained by dividing the ratio values in column C4/C2 by the corresponding ratio values in column C3/C2. They provide the comparison ratio values by which the drop-one single-sample estimator of shrinkage is evaluated.

For each of the five populations summarized in Tables 1–5, the ratio values in column C4/C3 are close to unity for samples with $n > 25$. The few C4/C3 values that exceed 1.0 are probably due to sampling error. It should be noted that the C4/C3 ratios tend to be less than unity for the smaller sample sizes. When $n \leq 25$, reductions from unity of the C4/C3 values are 4.5%–11% for the LAD regression model and 4.5%–15% for the LSD regression model in population 1. For populations 2–5, the corresponding reductions are 4%–10% (LAD) and 4.5%–14% (LSD), 3.5%–9.5% (LAD) and 3.5%–14% (LSD), 3%–6% (LAD) and 4.5%–11% (LSD), and 0%–5.5% (LAD) and 0%–11% (LSD), respectively. Thus, the drop-one single-sample estimator (i.e., C4/C2) is an excellent estimator of shrinkage (i.e., C3/C2), although it is conservative for very small samples. This conclusion holds for all sample sizes greater than $n = 25$, both cases (6 and 10), both regression models (LAD and LSD), and all five populations with differing degrees and amounts of data contamination.

Column C3/C1 summarizes, in ratio format, the validation fit (C3) to the true population ρ value (C1). This is sometimes referred to as ''expected skill'' (Mielke et al. 1996). In general, the C3/C1 values indicate the amount of skill that is expected relative to the true skill possible when an entire population is available. More specifically, the C3/C1 values indicate the expected reduction in fit of the y and $\tilde{y}$ values for future events (Mielke et al. 1996). A C3/C1 value that is greater than 1.0 is cause for concern since this indicates that the sample regression coefficients provide a better validation fit, on the average, than would have been possible had the actual population been available.

Inspection of column C3/C1 in Table 1 reveals that the LSD regression model consistently performs better than the LAD regression model, case 10 has lower values than case 6, and the C3/C1 values increase with increasing sample size. Table 2, with 1% moderate contamination, yields a few C3/C1 values greater than 1.0 and they all appear with the LSD regression model. Table 3, with 1% severe contamination, shows the same pattern, but the C3/C1 ratio values are somewhat higher. Table 4, with 5% moderate contamination, continues the same motif and Table 5, with 5% severe contamination, contains C3/C1 values considerably greater than 1.0 for nearly every case. It is abundantly clear that with only a small amount of moderate or severe contamination, the LSD regression model produces inflated estimates of expected skill. The LAD regression model, based on absolute deviations about the median, is relatively unaffected by even 1% severe contamination, but the LSD regression model, based on squared deviations about the mean, systematically overestimates the validation fit and yields inflated values of expected skill (i.e., C3/C1).

Since C3 (validation fit $\hat{\rho}$) values and C4 (drop-one single-sample validation fit $\hat{\rho}$) values are essentially the same for all five populations, both cases, both regression models, and all sample sizes, it is readily apparent that C4/C1 ratios would be nearly identical to the C3/C1 ratios in Tables 1–5. Consequently, caution should be exercised in using drop-one estimators with the LSD regression model as they will likely provide inflated estimates of validation fit when contaminated data are present. Because the drop-one estimate of shrinkage is equivalent to drop-one cross-validation, the same caution applies to drop-one cross-validation with an LSD regression model.

While it is abundantly evident that LSD regression systematically overestimates validation fit, the reason for the optimistic C3/C1 values is not as manifest. It is obvious that the inflated estimates of expected skill for LSD regression in Tables 1–5 are systematically related to sample size with larger sample sizes associated with C3/C1 values in excess of 1.0. This is probably due to a moderately or

severely contaminated population event occurring in a single sample. Very small samples (e.g., $n = 15$) are not likely to include a contaminated event, whereas very large samples (e.g., $n = 500$) are much more likely to include one or more contaminated events. Table 6 provides the probability values that no contaminated population event belongs to a single sample for both 1% and 5% contamination. The probability that no contaminated event belongs to a single sample with 1% moderate or severe contamination in the population is given by $(3958/3998)^n$, and the probability that no contaminated event belongs to a single sample with 5% moderate or severe contamination in the population is given by $(3958/4158)^n$ in Table 6. For 1% moderate or severe contamination, the probability of selecting no contaminated events from the population is greater than 0.50 for samples of size $n \leq 65$. For 5% moderate or severe contamination, the probability of selecting no contaminated events from the population never exceeds 0.50. Given the well-known sensitivity of LSD regression to extreme events, it is not surprising that LSD regression yields optimistic levels of expected skill for larger samples that are more likely to contain one or more moderate or severely contaminated events. It should be noted in Table 5 that neither LSD nor LAD regression is able to accommodate 5% severe contamination.

The single sample estimate of shrinkage, C4, is higher for 6 predictors than for 10 predictors in Table 1 with LAD regression and $n \leq 160$ and with LSD regression and $n \leq 250$, in Table 2 with LAD regression and $n \leq 250$, and with LSD regression and $n \leq 40$. The same relationship holds for both LAD and LSD regression in Table 3 with $n \leq 40$ and in Tables 4 and 5 with $n \leq 25$. These results are consistent with the influence of contamination since when n is small, the influence of additional noninformative predictors is mitigated because the probability of selecting a contaminated event in each sample is reduced. Clearly, regression models containing noninformative predictors should be avoided (Browne and Cudeck 1992).

The standard deviations of the 10 000 $\hat{\rho}$ values composing C2, SD($\hat{\rho}$|C2), and the standard deviations of the 10 000 drop-one $\hat{\rho}$ values composing C4, SD($\hat{\rho}$|C4), are given for each sample size ($n = 15, \ldots, 500$), case (10 and 6 predictors), and regression model (LAD and LSD) combination in Tables 7, 8, 9, 10, and 11, which correspond to the five contamination levels of Tables 1, 2, 3, 4, and 5, respectively. In particular,

$$\mathrm{SD}(\hat{\rho}) = \left[\frac{1}{M-1}\sum_{i=1}^{M}(\hat{\rho}_i - \overline{\rho})^2\right]^{1/2},$$

where

$$\overline{\rho} = \frac{1}{M}\sum_{i=1}^{M}\hat{\rho}_i,$$

$M = 10\,000$ in this study, and $\overline{\rho}$ corresponds to either C2 or C4. The standard deviations are confined to SD($\hat{\rho}$|C2) and SD($\hat{\rho}$|C4) since the associated estimable single sample $\hat{\rho}$ values exist only for C2 and C4. For all five tables, SD($\hat{\rho}$|C2) is smaller than SD($\hat{\rho}$|C4) for small sample sizes. However, SD($\hat{\rho}$|C2) and SD($\hat{\rho}$|C4) become more similar to one another with increasing sample sizes. Also for all five tables, the SD($\hat{\rho}$|C2) values are fairly similar for cases with 10 and 6 predictors; this also holds for the SD($\hat{\rho}$|C4) values. The differences between LAD and LSD regression for both SD($\hat{\rho}$|C2) and SD($\hat{\rho}$|C4) are more complex. While SD($\hat{\rho}$|C2) is smaller for LAD regression than for LSD regression with small sample sizes (15, 25, and 40), the SD($\hat{\rho}$|C2) and SD($\hat{\rho}$|C4) values are larger (perhaps slightly) for LAD regression than for LSD regression in Table 7. In Tables 8 and 9, SD($\hat{\rho}$|C2) is smaller for LAD regression than for LSD regression whereas this observation holds for SD($\hat{\rho}$|C4) only with large sample sizes (250 and 500). In Tables 10 and 11, except for SD($\hat{\rho}$|C2) with small sample sizes (15, 25, and 40), both SD($\hat{\rho}$|C2) and SD($\hat{\rho}$|C4) are larger for LAD regression than for LSD regression.

6. Summary

Mielke et al. (1996) investigated the effects of sample size, type of regression model, and noise-to-signal ratio on the degree of shrinkage in five populations containing varying amounts and degrees of data contamination. Shrinkage was measured as the ratio of the validation fit of a sample-based regression model to the retrospective fit of the same regression model where the validation fit was assessed on five independent samples from the same population. While the Mielke et al. (1996) index of shrinkage is both rigorous and comprehensive, it involves an additional five independent samples and thus is not useful in routine applications. In this paper a drop-one single-sample estimator of shrinkage is developed and evaluated on the same dataset used by Mielke et al. (1996). The drop-one single-sample estimator provides an accurate estimate of shrinkage for the five populations, both regression models, both cases, and all sample sizes, although the estimator is slightly conservative for very small sample sizes.

Finally, a caution is raised because the drop-one single-sample estimate of shrinkage is, in fact, an *estimate* of shrinkage. There is evidence that the drop-one method provides inflated estimates of validation fit for the LSD regression model when the population data is contaminated by extreme values, e.g., populations 1–4 in Tables 1–4. In population 5 (Table 5) with 5% severe contamination, both the LSD and LAD regression models provide estimates of validation fit that are too high.

Acknowledgments. This study was supported by National Science Grant ATM-9417563.

REFERENCES

Badescu, V., 1993: Use of Willmott's index of agreement to the validation of meteorological models. *Meteor. Mag.*, **122,** 282–286.
Barnston, A. G., and H. M. Van den Dool, 1993: A degeneracy in

cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977.

Browne, M. W., 1970: A critical evaluation of some reduced-rank regression procedures. Research Bulletin 70-21, Educational Testing Service, Princeton, NJ.

——, 1975a: Predictive validity of a linear regression equation. *Br. J. Math. Statist. Psychol.,* **28,** 79–87.

——, 1975b: A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *Br. J. Math. Statist. Psychol.,* **28,** 112–120.

——, and R. Cudeck, 1989: Single sample cross-validation indices for covariance structures. *Mult. Behav. Res.,* **24,** 445–455.

——, and ——, 1992: Alternative ways of assessing model fit. *Sociol. Meth. Res.,* **21,** 230–258.

Camstra, A., and A. Boomsma, 1992: Cross-validation in regression and covariance structure analysis. *Soc. Meth. Res.,* **21,** 89–115.

Copas, J. B., 1983: Regression, prediction, and shrinkage. *J. Roy. Statist. Soc.,* **45B,** 311–354.

Cotton, W. R., G. Thompson, and P. W. Mielke, 1994: Real-time mesoscale prediction on workstations. *Bull. Amer. Meteor. Soc.,* **75,** 349–362.

Efron, B., 1983: Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.,* **78,** 316–331.

Elsner, J. B., and C. P. Schmertmann, 1993: Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Wea. Forecasting,* **8,** 345–351.

——, and ——, 1994: Assessing forecast skill through cross-validation. *Wea. Forecasting,* **9,** 619–624.

Geisser, S., 1975: The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.,* **70,** 320–328.

Glick, N., 1978: Additive estimators for probabilities of correct classification. *Pattern Recog.,* **10,** 211–222.

Gray, W. M., C. W. Landsea, P. W. Mielke, and K. J. Berry, 1992: Predicting Atlantic seasonal hurricane activity 6–11 months in advance. *Wea. Forecasting,* **7,** 440–455.

Hess, J. C., and J. B. Elsner, 1994: Extended-range hindcasts of tropical-origin Atlantic hurricane activity. *Geophys. Res. Lett.,* **21,** 365–368.

Hora, S. C., and J. B. Wilcox, 1982: Estimation of error rates in several-population discriminant analysis. *J. Marketing Res.,* **19,** 57–61.

Horst, P., 1966: *Psychological Measurement and Prediction.* Wadsworth, 455 pp.

Huberty, C. J., J. M. Wisenbaker, and J. C. Smith, 1987: Assessing predictive accuracy in discriminant analysis. *Mult. Behav. Res.,* **22,** 307–329.

Kelly, F. P., T. H. Vonder Haar, and P. W. Mielke, 1989: Imagery randomized block analysis (IRBA) applied to the verification of cloud edge detectors. *J. Atmos. Oceanic Technol.,* **6,** 671–679.

Lachenbruch, P. A., 1967: An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics,* **23,** 639–645.

——, and M. R. Mickey, 1968: Estimation of error rates in discriminant analysis. *Technometrics,* **10,** 1–11.

Lee, T. J., R. A. Pielke, and P. W. Mielke, 1995: Modeling the clear-sky surface energy budget during FIFE 1987. *J. Geophys. Res.,* **100,** 25 585–25 593.

Livezey, R. E., A. G. Barnston, and B. K. Neumeister, 1990: Mixed analog/persistence prediction of seasonal mean temperatures for the USA. *Int. J. Climatol.,* **10,** 329–340.

MacCallum, R. C., M. Roznowski, C. M. Mar, and J. V. Reith, 1994: Alternative strategies for cross-validation of covariance structure models. *Mult. Behav. Res.,* **29,** 1–32.

Maltz, M. D., 1994: Deviating from the mean: The declining significance of significance. *J. Res. Crime Delinq.,* **31,** 434–463.

McCabe, G. J., and D. R. Legates, 1992: General-circulation model simulations of winter and summer sea-level pressures over North America. *Int. J. Climatol.,* **12,** 815–827.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.

Mielke, P. W., K. J. Berry, C. W. Landsea, and W. M. Gray, 1996: Artificial skill and validation in meteorological forecasting. *Wea. Forecasting,* **11,** 153–169.

Mosier, C. I., 1951: Symposium: The need and means of cross-validation, I. Problems and designs of cross-validation. *Educ. Psych. Meas.,* **11,** 5–11.

Mosteller, F., and J. W. Tukey, 1977: *Data Analysis and Regression.* Addison-Wesley, 586 pp.

Murphy, A. H., and R. L. Winkler, 1984: Probability forecasting in meteorology. *J. Amer. Statist. Assoc.,* **79,** 489–500.

Nicholls, N., 1985: Predictability of interannual variations of Australian seasonal tropical cyclone activity. *Mon. Wea. Rev.,* **113,** 1144–1149.

Picard, R. R., and R. D. Cook, 1984: Cross-validation of regression models. *J. Amer. Statist. Assoc.,* **79,** 575–583.

——, and K. N. Berk, 1990: Data splitting. *Amer. Statist.,* **44,** 140–147.

Snee, R. D., 1977: Validation of regression models: Methods and examples. *Technometrics,* **19,** 415–428.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc.,* **36B,** 111–147.

——, 1978: Cross-validation: A review. *Math. Operationsforsch. Statist., Ser. Statistics,* **9,** 127–139.

Subrahmanyam, M., 1972: A property of simple least squares estimates. *Sankhya,* **34B,** 355–356.

Toussaint, G. T., 1974: Bibliography on estimation of missclassification. *IEEE Trans. Inf. Theory,* **20,** 472–479.

Tucker, D. F., P. W. Mielke, and E. R. Reiter, 1989: The verification of numerical models with multivariate randomized block permutation procedures. *Meteor. Atmos. Phys.,* **40,** 181–188.

Watterson, I. G., 1996: Nondimensional measures of climate model performance. *Int. J. Climatol.,* **16,** 379–391.

Willmott, C. J., 1982: Some comments on the evaluation of model performance. *Bull. Amer. Meteor. Soc.,* **63,** 1309–1313.

——, S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, 1985: Statistics for the evaluation and comparison of models. *J. Geophys. Res.,* **90,** 8995–9005.