Loss Functions for Detecting Outliers in Panel Data: An Introduction

Coleman Charles

US Census Bureau

2003

LOSS FUNCTIONS FOR DETECTING OUTLIERS IN PANEL DATA: AN INTRODUCTION

Charles D. Coleman, U.S. Census Bureau, 4700 Silver Hill Rd., Stop 8800, Washington, DC 20233-8800
Email: ccoleman@census.gov

1. Introduction

In assuring data quality in forecasting, one would like to know that the data generation processes are free from anomalies. One interpretation of this is that the data do not have unexplainable outliers. In general, an outlier is an observation which departs from the norm (however defined) in a set of observations. Outliers can indicate problems with their data generation processes (i.e., anomalies) or may be true, but unusual, statements about reality.[1] In terms of Barnett and Lewis (1994, p. 37), we are testing for discordancy. This paper specializes the problem of detecting outliers to panel data, such as estimates and forecasts. Panel data are cross-sectional time series, such as a time series of population estimates for a set of areas.[2] Time may be either chronological or nominal. Nominal time indexes different sets of predictions (i.e., estimates or forecasts) for the same cross-sectional units and chronological time. Time is nominal in this context because the different predictions sets have no natural ordering. Comparing cross-sectional estimates to their true values is an instance of nominal time. The method this paper uses is to develop loss functions to identify discordant observations for further analysis. The loss functions are developed for panels of two dates and then extended to panels with arbitrary numbers of observations with arbitrary differences between dates.

Initially, the data are assumed to be positive.[3] In this context, the subject matter analyst's judgment is needed to determine the exact parametrization of the loss function, except for the special case described in Subsection 2.4.[4] The exact parametrization thus depends on the subject matter analyst and context. It is, thus, subjective. When the data can take on any real value, mathematical considerations dictate the exact parametrization.

The Population Division of the U.S. Census Bureau has been successfully using loss functions to detect outliers in the preparation of population estimates and geographic base files. Loss functions have been applied to input, intermediate and final data. Rather than use actual data, a numerical example illustrates how loss functions are used and how they avoid the pitfalls associated with taking numerical and percent differences. A map illustrates the use of loss functions with GIS and provides an illustration of the need for subject matter analyst expertise.

Section 2 develops loss functions for positive data. No distributional assumptions are made, as the natures of the data generation processes are assumed unknown and nonidentical.[5] Thus, this is an example of the nonparametric approach to outlier detection.[6] An important upshot of this approach is that data from a wide range of values are put on the same basis. This Section specifies the assumptions and develops the simplest loss function that satisfies these assumptions. Loss functions are developed for more general settings. Section 3 discusses some applications, including general usage of loss functions, parametrizing loss functions from preexisting outlier criteria and using loss functions with GIS. These examples are based on actual Census Bureau applications. Section 4 generalizes the framework to data that can take any real value. Section 5 concludes this paper.

2. The Loss Function[7]

This section describes the assumptions used to generate the loss function $L(F;B)$ and its variants, where F is the future value and B is the base period value. The loss function is the penalty, cost, or "badness" associated with the difference between F and B. Roughly speaking, the greater the difference between F and B, the greater the loss. Initially, F is assumed to be one period after B. After the necessary assumptions are made, the simplest form of L is specified. Restrictions on the values of the parameters of L which make it increase in B for a given relative difference are then specified. Subsection 2.1 axiomatically develops the simplest unsigned loss function L which satisfies these properties for data exactly

[1] This is similar to Hoaglin's (1983, p. 39-40) use of "outside cutoffs" to identify "outside values."
[2] The bidimensionality of data searched for outliers is not unique: DuMouchel (1999), Albert (1997) and Rudas, Clogg and Lindsay (1994) search for outliers in contingency tables. The contingency table approach differs in that time need not be a dimension and that parametric assumptions are made.
[3] Zeroes are permissible by adding a small constant, as discussed in Section 2 below.
[4] The subject matter analyst's judgment may already be incorporated in discrete outlier criteria. See Subsection 3.2.
[5] This obviates the use of parametric techniques, in which observations are tested for departure from a predetermined, hypothesized distribution.
[6] Barnett and Lewis (1994, pp. 107, 364-365) provide some references to nonparametric approaches in other contexts. Tukey (1977) proposed perhaps the most familiar nonparametric technique for detecting univariate outliers: the boxplot or box-and-whiskers plot. Rouseeuw, Ruts and Tukey (1999) propose the bagplot, a bivariate generalization of the boxplot.
[7] This exposition is based on Coleman, Bryan and Devine (2003, Section 2).

one period length apart. Subsection 2.2 generalizes L to situations in which F and B may not be exactly one period apart. Subsection 2.3 introduces the signed loss function for cases in which the sign of the difference is an additional important criterion. Subsection 2.4 parametrizes L for comparing two sets of estimates of the same parameters. Throughout this paper, B and F are assumed positive. Zeroes, which frequently arise in practice, are either recoded to small values or omitted from the analysis.

2.1 The Unsigned Loss Function

The unsigned loss function L is constructed by specifying three assumptions. The first assumption is that L is symmetric in the differences:

Assumption 1 (symmetry): $L(B+\varepsilon;B) = L(B-\varepsilon;B)$ for all B, $\varepsilon > 0$.

This assumption is not as innocuous as it looks. It is quite possible that, at least for some range of B, that positive and negative differences have differential impacts. However, the resulting asymmetry complicates the definition of L. Subsection 2.3 relaxes this assumption by developing the signed loss function, which allows the possibility of asymmetrically incorporating the direction of the difference ε. The symmetry of L allows us to use the equivalent notation $\lambda(\varepsilon, B) \equiv L(F, B)$ where $\varepsilon = |F - B|$.

The next assumption makes L, or, equivalently, ℓ, increasing in the difference ε:

Assumption 2 (monotonically increasing in difference): $\partial\lambda/\partial\varepsilon > 0$ for all $\varepsilon > 0$.

Note that this assumption is stated in terms of ℓ, rather than L. This assumption is quite intuitive, as it states that smaller differences are preferred to larger ones.

Finally, we want L, or, equivalently, ℓ, to decrease in B. This means that for a given value of ε, the loss associated with it decreases with its associated initial value. This has two justifications. First, for example, a difference of 500 when the initial value is 1,000 is a whopping 50%, a highly significant difference. However, the same difference, when the initial value is 1,000,000 is akin to a roundoff error. Second, when performing estimates or taking samples, the coefficient of variation, σ^2/μ^2, where σ^2 is the variance and μ is the expected value, decreases in B. This author's experience is that all areas tend to have about the same roundoff errors. Again, these are proportionately greater in small areas. We state this formally as:

Assumption 3 (monotonically decreasing in base value): $\partial\lambda/\partial B < 0$, or, equivalently $\partial L/\partial B < 0$, for all $B > 0$.

This simplest function which satisfies Assumptions 1–3 and admits Property 1 below is the Cobb-Douglas function[8]

$$L(F;B) = |F - B|B^q \tag{1a}$$

or, equivalently,

$$\lambda(\varepsilon, B) = \varepsilon B^q \tag{1b}$$

where $\varepsilon > 0$ and $q < 0$.[9]

An observed pair $(F;B)$ is an outlier whenever $L(F;B) > C$, where C is a predetermined critical value.[10] We will also refer to outliers as being critical. Additionally, we will refer to the equation $L(F;B) = C$ as the equation of criticality. The choice of q and C is an empirical matter.[11] Only a practitioner's experience with data can determine when data are suspect and incorporate these suspicions into parameters. One thing to note is that the loss function is ordinal: raising L and C to any positive power m leaves the rankings of losses unchanged.[12] It is only the rankings of losses that are important.[13] Another important quality is that loss is not necessarily interpretable. This is generally true of loss functions (Lindley, 1953, p. 46).

A desirable property of the loss function is that it increases in B for a given absolute relative difference. The absolute relative difference is:

$$|F - B|B^{-1} \tag{2}$$

Note that, in this case, $q = -1$. Choosing $q > -1$ makes the loss function increase in B, for a given absolute relative difference. We state this as Property 1:

Property 1: The loss function defined by equations (1a) and (1b) increases in B for any given absolute relative difference. This is assured whenever $q > -1$.

The reader may note that $q = 0$ turns equations (1a) and (1b) into the absolute values of the differences. Thus, values of q between 0 and –1 represent various

[8] It should be noted that an infinite number of loss functions satisfy Assumptions 1-3 and admit Property 1. This one is merely the simplest.

[9] Unlike Coleman (2000, 2002, 2003), no exponent on the difference is needed due to a Lie symmetry. See Coleman, Bryan and Devine (2003) for the explanation.

[10] Alternatively, C can also be determined from the data by taking a predetermined quantile or a multiple of the interquartile range of L (Tukey 1977).

[11] Subsection 2.4 below investigates a case in which q can be determined exactly.

[12] This is at the heart of the Lie symmetry noted in footnote 9.

[13] This is similar to the economic concept of ordinal utility. Coleman (2000, 2002, 2003) differs in using a cardinal framework: the values of the loss function can be compared to each other and operated upon arithmetically.

tradeoffs of absolute differences and absolute relative differences. Consider the product of the rth power of the absolute difference and the sth power of the absolute relative difference, where $r, s > 0$: $|F - B|^r \left(|F - B|B^{-1}\right)^s$. By the Lie symmetry invoked in footnote 9, this function is isomorphic to the loss function $|F - B|B^{-\frac{s}{r+s}}$. Thus, any value of q corresponds to an infinite number of pairs (r, s) where $q = -s / (r + s)$. Geometrically, the same loss function is generated for all (r, s) lying on the line $r = -(1 + q)\, s$.

2.2 The Time-Invariant Loss Function

Instead of considering the single set of future data, $\mathbf{F} = \{F_i\}_{i=1}^n$, where i indexes the n observations, consider the sets $\mathbf{F}_t = \{F_{it}\}_{i=1}^n$, where t is the amount of time elapsed since the base date and i indexes the cross-sectional units. We wish to develop a loss function which allows us to make comparisons across time on the same basis, by explicitly incorporating t into the loss function. One way of incorporating time-invariance is to substitute the geometric average absolute relative change

$$\left(\frac{|F_{it} - B_i|}{B}\right)^{1/t} \tag{3}$$

for the absolute relative change implicit in equation (1a) to create the time-invariant loss function[14]

$$L(F_{it}; B_i, t) = |F_{it} - B_i| B_i^{tq+t-1}. \tag{4}$$

Given this paper's framework, equation (4) should be used to make comparisons across time, as it puts the geometric average absolute relative difference on the same basis for all t. The reader can verify that $-1 < tq + t - 1 < 0$ for $t > 0$ and $0 > q > -1$.

2.3 The Signed Loss Function

At times, not only is the value of the loss function important, but also the sign of the difference. Different outlier generation processes may manifest themselves by producing predominantly positive or negative differences. We can account for these by creating the signed loss function S, which is simply the loss function L, multiplied by the signum function of the difference:

$$S(F; B) = |F - B|B^q \operatorname{sgn}(F - B) = (F - B)B^q \tag{5}$$

where $\operatorname{sgn} x = +1$ for $x > 0$, 0 for $x = 0$, and -1 for $x < 0$.

[14] For details, see Coleman, Bryan and Devine (2003), Subsection 2.3.

Using S, one can create different critical values for loss, depending on whether the difference is positive or negative. To wit, one can pick C_+, C_-, $C_+ \neq -C_-$, such that a pair $(F; B)$ is declared an outlier if either $S(F;B) < C_-$ or $S(F;B) > C_+$. Again, the choice of whether to use S and then use asymmetric critical bounds is an empirical matter.[15] For example, since, by assumption, negative values of F are impossible, then asymmetric critical bounds and/or parameters may be necessary to detect cases in which F becomes very small relative to B.

The time-invariant signed loss function is

$$S(F_{it}; B_i, t) = (F_{it} - B_i)B_i^{tq+t-1}. \tag{6}$$

2.4 Comparing Two Sets of Data: A Specialization of the Loss Function

Often, one is interested in comparing two sets of estimates of the same cross-sectional units. Suppose that the sets $\mathbf{B} = \{B_i\}$ and $\mathbf{F} = \{F_i\}$ represent two versions of estimates of the true values $\mathbf{A} = \{A_i\}$. This is an instance of nominal time. Suppose that both the B_i and F_i are unbiased estimators of the A_i and that their variances are proportionate to the A_i (i.e., $\operatorname{Var}(B_i) = \operatorname{Var}(F_i) = \sigma^2 A_i$.) One way one can think of this situation as that both B_i and F_i are constructed summing A_i jointly uncorrelated random variables with mean 1 and variance σ^2.[16] In this situation, we can use the loss functions (1a) and (1b) with $q = -½$. Since the null distributions of **B** and **F** are assumed unknown, it is impossible to do any significance testing. Moreover, since we are usually dealing with the entire population, sampling theory is not appropriate.

Of course, if the processes generating **B** and **F** are not as assumed, no theoretical guidance is available for the choice of q.

Again, the signed loss function (5) can be used with $q = -½$.

3. Applications

This section illustrates the use of loss functions by first outlining a general procedure for using loss functions in Subsection 3.1. Next, three different examples of loss functions are shown. In the first example, in Subsection 3.2, preexisting outlier criteria in terms of critical ratios by size class are transformed into a loss function. The second example, in Subsection 3.3, uses real-world data and GIS to compare two sets of real-

[15] The asymmetry need not be limited to the critical values. The signed loss function can incorporate different values of q, depending on the sign of the difference.

[16] Note that independent, identically distributed variables are a special case of this assumption.

world estimates using the $q = -½$ loss function of Subsection 2.4. The results of using absolute and absolute relative differences to evaluate differences between these two sets of estimates are discussed for comparison. Coleman et al.'s (2003, Subsection 3.4) method of using a reference variable to detect outliers is not discussed.

3.1 General Procedure for Using Loss Functions

Loss function evaluations usually begin by recoding zero base values to a small positive value,[17] (the exact value determined by the range and smallest value of the data and smaller than the smallest value) and setting $q = -0.5$. If time is chronological, the subject matter analyst then has to examine the data and the rankings of their associated losses.[18] If, in the subject matter analyst's opinion, too many observations with small changes occurring to small base values are ranked highly, then q should be increased.[19] If, on the other hand, too many observations with small changes to large base values are ranked highly, then q should be decreased. This process continues until the analyst is satisfied with the loss rankings. This author has found that changing q by increments of .1 is satisfactory. Finer increments appear to have little effect.

3.2 Creating Loss Functions From Discrete Outlier Criteria

Sometimes, discrete outlier criteria have already been developed. These discrete outlier criteria can be converted into a loss function using regression. Given a set of critical pairs (ε, B), the regression

$$\log \varepsilon = -q \log B + K + \text{error} \qquad (7)$$

is estimated. q is immediately obtained from equation (7). C is then obtained as $C = e^K$.

Often, outlier criteria do not come in discete pairs. Instead, they come in ranges $\left[\underline{B}, \overline{B}\right]$ for which an outlier is declared whenever ε / B exceeds a prescribed value. Coleman et al. (2003, Subsection 3.3) recommend using the midpoints of these ranges to form the pairs (ε, B). If an unbounded uppermost range is present, its lower bound is used.

A further complication is that the outlier criteria may be inconsistent with the assumptions used to develop a loss function. For example, two different ranges may have the same minimum ε, thereby violating Assumption 3. In these cases, the offending ranges have to be either modified or removed. They may be modified if a developer of outlier criteria can be queried to produce satisfactory criteria. If this is not possible, these ranges must be omitted from regression (7).

3.3 A Numerical Example

Table 1 presents an example of two cross-sectional series, their absolute differences and their absolute percent differences and loss functions with $q = -0.5$ using Column 'B_i' as the base. These data are presented in increasing order of B_i (or, equivalently, F_i). Normally, the data are presented to the subject matter analyst in decreasing order of loss (or absolute difference or absolute percent difference).

Table 1
Numerical Example of Loss Functions

i	B_i	F_i	Absolute Difference	Absolute Percent Difference	Loss
1	1	2	1	100	1.00
2	100	105	5	5	0.50
3	500	525	25	5	1.12
4	600	624	24	4	0.98
5	700	735	35	5	1.32
6	1000	1040	40	4	1.26
7	10000	10100	100	1	1.00

Note that the absolute difference is increasing in B (and, equivalently, in F.) If one were to use absolute difference as the measure of "outlierhood," one would generally find that the observations with the largest base values are the most likely to be outliers. Conversely, focusing on the percent absolute differences would cause the observations with the smallest base values to generally be classified as outliers. The extreme case of this is shown in the first row of Table 1. The pair (1, 2) has an absolute percent difference of 100%. Yet, in many contexts, this difference is meaningless. For example, one data source may show one birth in a county, while another shows two. If a component method is used to estimate population in that county, the two data sources will produce a difference of exactly one person. This difference is generally meaningless. For example, the difference between population estimates of 10,000 and 10,001 is meaningless, falling well within the overall error of the estimates.

The loss function effectively trades off the

[17] In some instances, this step should be omitted, as it can cause spurious identification of true zeroes as outliers. Only examination of the results can determine whether this is the case.
[18] The same can be done in nominal time. If the assumptions of Subsection 2.4 are violated, then no particular value of q is prescribed.
[19] That is, q is made closer to zero, say, –0.4.

absolute and absolute percent differences.[20] The large absolute percent difference in row 1 is severely downweighted by its small absolute difference. Likewise, the last row has a large absolute difference, but small absolute percent difference. These two cases have the same loss.

Rows 5 and 6 have similar loss. Because loss is ordinal, no meaning can be placed on this difference, other than row 5 is "worse" than row 6. Instead, the subject matter analyst examines the data process generating row 5 before examining row 6. If, in his opinion, the losses are not properly reflecting the severity of the outliers, the loss functions should be recomputed with a different value of q.

3.4 An Example Using GIS

Geographic information systems can be used with loss functions to find outliers. GIS is particularly helpful for finding geographic patterns in outliers. Map 1 at the end of this paper shows the $q = -1/2$ loss function applied to two different sets of county population estimates.[21] This is an example of nominal time. The base population is the Vintage 1998 published number obtained by the "tax method" component change model.[22] The comparison population is the county household population implied by the subcounty population estimates system, including overrides,[23] before constraining to any higher level totals.[24,25] Southern California, the Dallas-Fort Worth Metropolitan Area, northern Nevada and northern Maine stand out, among others. Most of the counties in the Great Plains that stood out on a map of absolute percentage differences[26] no longer stand out. This is because their populations are very small. Other areas stand out which do not appear on maps of absolute and absolute percent differences include the outer suburbs of Detroit and the Denver area. Northern Maine and Nevada have large enough populations to make their percentage changes stand out. In the cases of Southern California and Dallas-Fort Worth, the populations are so large that small percentage changes create large losses. This may lead the subject matter analyst to conclude that a different value of q should be used. In the other cases, it is the combination of moderate population bases and moderate percentage changes that causes high loss. In any case, the interpretation of the losses is clear: high losses indicate large divergences between the two methods. It is these areas upon which an analyst should focus his attention. By varying q and examining maps and ranked lists of outliers, the analyst can obtain an appropriate value of q, which yields the greatest information about the outliers.

4. Extending the Loss Function to All Real Pairs[27]

Sections 1 through 3 developed a loss function to find outliers in positive data. In many cases, however, data can take on any real value, such as the Census Bureau's net migration data. Thus, the arguments to the loss function are a real pair. For this problem, a new set of assumptions is required. An important difference is that the parameter q is no longer adjusted as a result of subject matter analyst's review. Instead, geometric considerations dictate the choice of q. Another difference is that the assumptions involved become more elaborate. The Census Bureau has used this loss function to find outliers in raw net migration data.

Subsection 4.1 axiomatically develops the simplest unsigned loss function L. Subsection 4.2 develops the signed loss function, similar to that developed earlier. Subsection 4.3 uses geometry to determine q.

4.1 The Unsigned Loss Function

The unsigned loss function L is constructed by making five assumptions. The first assumption is that L is defined everywhere in the real plane $\Re^2$:

Assumption 4 (unrestricted domain): For all $(F, B) \in \Re^2$, $L(F, B)$ is defined and single valued.

The next assumption is that L is symmetric in the difference between B and F:

Assumption 5 (symmetry in difference): $L(B+\varepsilon; B) = L(B-\varepsilon; B)$ and $L(F, F+\varepsilon) = L(F, F-\varepsilon)$ for all B, F and $\varepsilon \in \Re$.

Like Assumption 1, this assumption is not as innocuous as it looks. It is quite possible that, at least for some ranges of B and F, that positive and negative differences have differential impacts. However, the resulting asymmetry

[20] The discussion in the last paragraph of Subsection 2.1 formally demonstrated this.

[21] Counties with "no data" on this map are those which have no subcounty geography per the Census Bureau's Population Estimates Branch's definitions.

[22] These are contained in the Census Bureau's file *98C8_00.txt*, which was released to the public in 1999.

[23] The overrides, or administrative changes, consist of numbers obtained by special censuses, challenges and other corrections to the initial estimates.

[24] In terms of Section 2, the published populations are the B_i and the subcounty estimate-derived data are the F_i.

[25] The subcounty estimates methodology may be found at *http://www.census.gov/population/methods/e98scdoc.txt*.

[26] Coleman et. al (2003) Map 2. Map 1 of that paper displays absolute differences.

[27] This Section is based on Coleman and Bryan (2003).

complicates the definition of L. Subsection 4.2 relaxes this assumption somewhat by developing the signed loss function, which allows the possibility of incorporating the direction of the difference ε. However, as Subsection 4.2 states, this relaxation only affects the critical values used.

A desirable property is that L be symmetric with respect to its arguments. To give a concrete example, we want $L(-1,1000) = L(1000, -1)$. This stated formally as Assumption 6:

Assumption 6 (symmetry in arguments): $L(B, F) = L(F, B)$.

At this point, it useful to introduce some new notation. Let $X=|F|$ and $Y=|B|$. Let the new loss function $\lambda(\varepsilon,\Sigma) \equiv L(F,B)$, where $\varepsilon = |F-B|$ and $\Sigma = \Sigma(X,Y)$ is a function such that $\partial\Sigma/\partial X > 0$ and $\partial\Sigma/\partial Y > 0$. Assumption 6 implies that $\Sigma(X,Y) = \Sigma(Y,X)$, so that Σ is symmetric in its arguments. The remaining Assumptions are stated in terms of λ.

Assumption 2 of Section 2 is repeated to make λ (and L) increase in the difference ε:

Assumption 2 (monotonically increasing in difference): $\partial\lambda/\partial\varepsilon > 0$ for all $\varepsilon \geq 0$.

Finally, we want to create an assumption analogous to Assumption 3 of Section 2 to make λ to decrease in Σ, for similar reasons. We state this formally as:

Assumption 7 (monotonically decreasing in arguments): $\partial\lambda/\partial\Sigma < 0$ for all $\Sigma > 0$.

This simplest function which satisfies Assumptions 2 and 4–7 is (after invoking a Lie symmetry)[28]

$$\lambda(\varepsilon,\Sigma) = \begin{cases} \varepsilon\Sigma^q & \Sigma \neq 0 \\ 0 & \Sigma = 0 \end{cases} \tag{8}$$

where $q < 0$. Note that equation (8) is stated in terms of ε and Σ. The simplest form of Σ will be determined in equation (9) below. Theorem 1 of Coleman and Bryan (2003) shows that setting $\lambda(0,0) = 0$ makes λ continuous at (0,0), when $q > -1$. This way of determining $\lambda(0,0)$ avoids division by 0.

4.1.1 Determination of Σ and L

From equation (1), it is clear that $\lambda(0,\Sigma) = 0$ for all $\Sigma > 0$. We would like to define Σ so that whenever either X or $Y \neq 0$, $\Sigma > 0$. We would also like $\Sigma(0,0) = 0$. The simplest equation for Σ is:

$$\Sigma(X,Y) = X + Y = |B| + |F| \tag{9}$$

From equation (9) we can determine L to be

$$L(F,B) = \begin{cases} |F-B|\left(|F|+|B|\right)^q & B \text{ or } F \neq 0 \\ 0 & B = F = 0 \end{cases} \tag{10}$$

A desirable property of the loss function is that it rises in $|F - B|$ for a given average absolute percentage difference. The average absolute relative difference is defined as:[29]

$$|F-B|\left(|F|+|B|\right)^{-1} \tag{11}$$

Note that, in this case, $q = -1$. Choosing $q > -1$ makes the loss function rise in $|F| + |B|$, for a given average absolute relative difference. This is also required by Theorem 1 of Coleman and Bryan (2003). We state this as Property 1′:

Property 1′: The loss function defined by equations (5) increases in $|F| + |B|$ for any given average absolute percentage difference. This is assured whenever $q > -1$.

The reader may note that $q = 0$ turns equation (10) into the absolute values of the difference. Thus, values of q between 0 and –1 represent various tradeoffs between the absolute value of the difference and average absolute percentage difference. Consider the product of the rth power of the absolute difference and the sth power of the average absolute relative difference, where $r, s > 0$: $|F-B|^r \times \left(\frac{|F-B|}{|F|+|B|}\right)^s$. By Lie symmetry, this function is isomorphic to the loss function $|F-B|\left(|F|+|B|\right)^{-\frac{r}{r+s}}$. Thus, these intermediate values of q correspond to an infinite number of pairs (r, s) where $q = -r / (r + s)$. Geometrically, the same loss function is generated for all pairs (r, s) lying on the line $s = -(1 - 1/q)\, r$.

4.2 The Signed Loss Function

Again, we create the signed loss function S, which is again simply the loss function L, multiplied by the signum function of the difference:

$$S(F,B) = |F-B|\left(|F|+|B|\right)^q \operatorname{sgn}(F-B) = \begin{cases} (F-B)\left(|F|+|B|\right)^q & B \text{ or } F \neq 0 \\ 0 & B = F = 0 \end{cases} \tag{12}$$

Using S, one can create different critical values for loss, depending on whether the difference is positive or

[28] It should be noted again that an infinite number of loss functions satisfy Assumptions 1-3. This one is merely the simplest.

[29] This is obtained by taking the average of absolute relative differences formed with B and F in the denominators: $|F-B||B|^{-1}$ and $|F-B||F|^{-1}$ and assuming that $B \approx F$.

negative, similar to Subsection 2.3. Again, one can pick C_+, C_-, $C_+ \neq -C_-$, such that a pair (F, B) is declared an outlier if either $S(F;B) < C_-$ or $S(F;B) > C_+$. Again, the choice of whether to use S and then use asymmetric critical bounds is an empirical matter.[30] However, since S has been developed using strong symmetry assumptions, using asymmetric bounds is probably not worthwhile for detecting outliers. The next Section relies on geometric analysis of S to suggest the best choice for q.

4.3 Choice of Loss Function

The loss functions L and S exhibit wildly different behaviors depending on the value of q. The choice of q requires examination of plots of S for various values of q, $-1 \leq q \leq 0$, to obtain a reasonable loss function.[31] The limiting functions when $q = 0$ and $q = -1$ are of particular interest. $q = 0$ implies that $S(F,B) = F - B$. This defines a plane in $\Re^3$, which is not useful for outlier detection in this paper's framework. Setting $q = -1$ produces some strange behavior. Whenever B and F are of opposite signs, $S(F,B) = \text{sgn } F$. This can be seen by substituting $q = -1$ into equations (12) when B or F is nonzero:

$$S(F,B) = (F - B) / \left(|F| + |B|\right) \tag{13}$$

Noting that $|x| = x$ when $x > 0$ and $|x| = -x$ when $x < 0$, we can examine the behavior of S when B and F are of opposite signs. When $F > 0$ and $B < 0$, equation (13) becomes

$$\begin{aligned} S(F,B) &= (F - B) / \left(|F| + |B|\right) \\ &= \left[|F| - \left(-|B|\right)\right] \Big/ \left(|F| + |B|\right) \\ &= \left(|F| + |B|\right) \Big/ \left(|F| + |B|\right) = 1 = \text{sgn } F \end{aligned} \tag{14}$$

The reader may verify that $S(F,B) = -1 = \text{sgn } F$ when $F < 0$ and $B > 0$. These equalities easily generalize to the cases in which either B or F is zero.

Another problem occurs at the origin when $q = -1$: from the previous paragraph we can observe that S simultaneously acquires the values ± 1, which contradicts the assumption that S is single-valued.[32]

Finally, cusps exist along the axes for every $q < 0$, but are most severe for $q = -1$.[33]

Given all of the anomalies and degeneracies associated with this family of loss functions, the problem is to decide on a value of q which produces reasonable behavior, in his mind. It appears that intermediate choices of q are best behaved: these offer a good compromise between simply taking the difference between F and B ($q = 0$) and the bizarre behavior of S when q approaches -1. In particular, the value $q = -0.5$ shows the best tradeoff of the different attributes. Thus, the recommended unsigned loss function is

$$L(F,B) = \begin{cases} |F - B| \left(|F| + |B|\right)^{-0.5} & B \text{ or } F \neq 0 \\ 0 & B = F = 0 \end{cases} \tag{15}$$

with the corresponding signed loss function

$$S(F,B) = \begin{cases} (F - B) \left(|F| + |B|\right)^{-0.5} & B \text{ or } F \neq 0 \\ 0 & B = F = 0 \end{cases}. \tag{16}$$

Again, note that no subject matter analyst's judgment is used to parametrize these loss functions. Instead, the parametrization is based on an evaluation of the geometry of these functions.

6. Conclusion

This paper has used time as an explicit dimension in constructing loss functions for detecting outliers in panel data. Loss functions put all differences on the same basis so that data ranging several orders of magnitude can be compared. When the data are positive, interaction with the subject matter analyst is necessary to properly parametrize the loss function. When the data can assume any real value, geometric considerations dictate the parametrization of the loss function. Some examples have been provided.

7. Acknowledgements

This paper is based on Coleman et al. (2003) and Coleman and Bryan (2003). I would like to thank Mary H. Mulry for helpful comments and peer review.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

References

Barnett, Vic and Lewis, Toby (1994). *Outliers in*

[30] The asymmetry need not be limited to the critical values. The signed loss function can incorporate different values of q, depending on the sign of the difference. However, as Subsection 3.2 shows, there is little latitude in the choice of q.

[31] This is done in Coleman and Bryan (2003). This is a different sort of subjectivity than that of Section 2. There, the coefficient q is determined empirically, often from the data. In this Section, the subjectivity lies in the choice of the form of the loss function.

[32] This argument does not even consider approaching the origin along rays in the positive and negative orthants, which may produce yet other values for S.

[33] These can be seen in Coleman and Bryan (2003, Figures 3-11).

Statistical Data, 3rd edition, John Wiley & Sons, New York.

Coleman, Charles D., (2000). "Evaluating and Optimizing Population Projections Using Loss Functions," *Federal Forecasters Conference 2000: Papers and Proceedings*, Washington: U.S Department of Education, Office of Educational Research and Improvement, 27-32.

Coleman, Charles D., (2002). "Optimizing Population Projections Using Loss Functions When the Base Populations are Subject to Revision," *Federal Forecasters Conference 2002: Papers and Proceedings*, Washington: U.S Department of Education, Office of Educational Research and Improvement, 27-32.

Coleman, Charles D. (2003). "Loss Functions for Assessing the Accuracy of Cross-Sectional Predictions," manuscript, U.S. Census Bureau.

Coleman, Charles D. and Bryan, Thomas (2003). "Loss Functions for Detecting Outliers in Panel Data when the Data May Change Sign," manuscript, U.S. Census Bureau.

Coleman, Charles D., Bryan, Thomas and Devine, Jason (2003). "Loss Functions for Detecting Outliers in Panel Data," manuscript, U.S. Census Bureau.

DuMouchel, William (1999). "Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System," *The American Statistician* **53**, 177-188.

Hoaglin, David C. (1983). "Letter Values: A Set of Selected Order Statistics." In, Hoaglin, David C., Frederick Mosteller and John W. Tukey [eds.], *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

Lindley, D.V. (1953). "Statistical Inference," *Journal of the Royal Statistical Society*, Series B **15**, 30-76.

Rousseeuw, Peter J., Ruts, Ida and Tukey, John W. (1999). "The Bagplot: A Bivariate Boxplot," *The American Statistician* **53**, 382-387.

Rudas, T., Clogg, C. C. and Lindsay, B. G. (1994). "A New Index of Fit Based on Mixture Methods for the Analysis of Contingency Tables," *Journal of the Royal Statistical Society*, Series B **56**, 623-639.

Tukey, John W. (1977). Exploratory Data Analysis, Addison-Wesley, Reading, Massachussetts.