Local Control

SAS macros:

User’s Guide

Local Control (LC) is a unique, new, non-parametric approach to Adjustment for Treatment Selection Bias and Confounding in Large Observational Studies.

These macros were created under contract to OMOP, FNIH, and are freely distributed under the **Apache Licensing Agreement**, Version 2.0.

Guide Version 1.2, February 2010

Bob Obenchain, PhD, FASA, OMOP Methods Partner
Observational Medical Outcomes Partnership (OMOP)

Foundation for the National Institutes of Health (FNIH)
ATTN: Executive Director, OMOP
9650 Rockville Pike, Bethesda, MD 20814

Table of Contents

1. Introduction

This user's guide defines syntax and illustrates use of SAS macros that perform Local Control (LC) patient **clustering**, a new approach to analysis of observational studies, retrospective databases, patient registry data or poorly randomized (chaotic) studies. The SAS macros described here implement an analysis strategy that can be described as **post hoc blocking** of patients. Any of the traditional "unsupervised learning" algorithms implemented in SAS/STAT can be used to hierarchically cluster patients in the, say, Euclidean space defined by their baseline X-characteristics. Of course, other clustering algorithms or patient dissimilarity metrics could also be used in LC analyses; see Kaufman and Rousseeuw (1990).

The LC macros make some assumptions that are different from those generally made by other methods currently being implemented for OMOP evaluation and comparison. Specifically, the LC macros assume that the objective of the analysis of observational data is to compare two treatments head-to-head in a fair and objective way that adjusts for not only possible treatment selection bias (i.e. treatment cohort imbalance) but also for confounding of patient X-characteristics. Thus the LC macros assume that each subject in the subset of the data currently being analyzed has received only one of two alternative treatments for a single disease or condition. For example, in the two datasets used here to illustrate LC analyses, all patients have either received a PCI cardiovascular procedure (with treatments "usual care alone" or "usual care augmented with a blood thinner") or else took one of two hypothetical statins. Finally, the treatment indicator is assumed here to be a single, binary variable with the two treatments coded as 0 and 1, and at least one numerical, baseline patient X-characteristic (measure of disease severity or patient frailty) is contained in the dataset to at least partially reveal the potential extent of treatment imbalance (selection bias) and/or confounding.

The fundamental LC analysis strategy is to **make treatment comparisons only within X-space clusters of relatively well-matched patients**. The primary LC "sensitivity analysis" tactic implemented in the LC macros is to steadily increase the number of clusters, thereby forcing clusters to become small, compact and numerous. In this process, some clusters may ultimately become "pure" and "uninformative" in the sense that they contain only treatment = 1 patients or only treatment = 0 patients. An objective in LC analysis is thus to **reach a compromise** where not only [1] all clusters are small enough that the patients within each cluster can be considered well-matched in X-space but also [2] most clusters are large enough that a high percentage of patients (say, 95% or 90%) are in informative clusters (i.e. clusters that contain at least one treatment = 1 patient as well as at least one treatment = 0 patient.) A useful **upper bound on the total number of clusters** considered is that the overall average number of patients per cluster should be at least 10 to 12. For example, no more than about 900 clusters should be formed given data on 10,000 patients.

The statistic used for making within cluster treatment comparisons is the **Local Treatment Difference** in Y-outcomes:

$$LTD: \ \bar{Y}_{Treatment} - \bar{Y}_{Control} \, .$$

It is intuitively clear from first principles that the ***LTD***, being a **simple difference in mean values** within a cluster, is unbiased even when the number of treatment = 1 patients within a cluster is different from the number of treatment = 0 patients. In fact, it is easily shown that the ***LTD*** statistic is a local **Inverse Probability Weighted** (IPW) estimate, where the probabilities are the observed **local propensities** to receive treatment or control, respectively; see Robins, Hernan and Brumback (2000) or Lunceford and Davidian (2004) for IPW concepts and Rosenbaum and Rubin **(**1983, 1984) for propensity scoring fundamentals. Finally, the ***LTD*** statistic is also identical to the local **"doubly robust"** estimate, Bang and Robins (2005), for a simple Nested ANOVA model (treatment within cluster.)

For users wishing to learn much more LC, the primary published reference is "The Local Control Approach using JMP" …which is Chapter 7 of the SAS Press book: ***Analysis of Observational Health Care Data Using SAS,*** Faries, Leon, Haro and Obenchain, eds. (2010.) While the SAS macros described here have similar basic functionality as my **JMP** scripts, the macros tend to be not only less intuitive and easy-to-use but also produce visualizations will less immediate impact.

Overall LC analysis strategy is divided into four tactical phases.

Phase One: Explore

Use the LC_NCreq macro procedure to request a large number of non-hierarchical clusters or, alternatively, use the LC_Cluster, LC_LTDdist and LC_UnBias macro procedures to systematically increase the number of clusters within a given hierarchy. Note how the mean of the LTD distribution changes as the number of clusters is increased when treatment selection bias and/or confounding are/is present.

Phase Two: Confirm

Use the LC_Salient macro to confirm that a potential (candidate) clustering of patients on their baseline X-characteristics makes a "real" (clearly visible) difference in the observed LTD distribution when compared with the "artificial" LTD distribution resulting from purely random patient clustering.

Phase Three: Agonize

Use the LC macros to perform "systematic sensitivity analyses" and to identify LTD distributions that are most typical and stable   Literally redo Phase One and Two analyses in several alternative ways …say, using only subsets of the available patient baseline X-characteristics and/or different analysis options for SAS proc CLUSTER (§3.07) and proc STANDARDIZE (§3.08.)

Phase Four: Realize

Use the SAS Stat procedures of your choice to (try to) predict patient-level LTD estimates from their baseline X-characteristics. This activity is postponed until last because it is frequently quite frustrating …due to absence of law-like relationships between patient outcomes and their observed X-characteristics. In fact, it's actually "OK" if this final phase of analysis essentially fails (e.g. yields low model R-squares.) After all, the observed LTD distribution has already been shown to be “salient” …i.e. it has been meaningfully adjusted for treatment selection bias and confounding.

2. Local Control (LC) Macro Procedures

Five macro procedures are documented here for performing Local Control analyses.

Phase One LC tactics are usually implemented via:

> [1a] a single invocation of macro procedure “LC_Cluster” to construct the dendrogram “tree” for a hierarchical clustering of all patients in X-space, followed by
>
> [1b] multiple invocations of macro procedure “LC_LTDdist” for a single Y-outcome variable named by “LC_Yvar” with ever increasing values for “NCreq” = number of clusters requested, and ending with
>
> [1c] a single invocation of macro procedure “LC_UBtrace,” which sorts the LC summary dataset named by the “LC_UnBias” macro variable (on variable _FREQ_ = NCreq) and then both plots and tabulates these LC summary statistics in a PDF output file.

Unfortunately, this default approach to LC Phase One **does not scale up well** to very large numbers of patients …say, when there are more than 100,000 subjects embedded in a Euclidean X-space of more than a single dimension. However, an alternative LC Phase One computing strategy using SAS procedure FASTCLUS (method = K-means, non-hierarchical) may still be practical when the **observed patient X-vectors contain many exact matches to only a limited number of distinct patterns**.

For example, the “statin1m” dataset distributed with the LC macros contains one million patients, but all 7 patient X-characteristics are binary. Thus there cannot be more than 2^7 = 128 possible X-patterns! In fact, only 96 different X-patterns are actually expressed in these data, and 94 of these 96 X-space clusters turn out to be informative about LTDs.

This alternative, **non-hierarchical approach** also provides a strategy for implementing an “approximate” LC analysis with a very large dataset (more than 100,000 patients) containing continuous X-variable(s), which are unlikely to provide “exact” matches among patients. A preliminary step is then needed in which the user must **recode all continuous X-variables into a small number of levels** (say, 5 levels.) These recoded X-variables are then used instead of the original, continuous X-variables in LC analyzes as follows:

[1’] Invoke the alternative LC Phase One macro procedure “LC_NCreq” using a numerical value for “NCreq” = (number of clusters requested) that is large enough that **all clusters will contain only exact matches on recoded Xs.** In fact, you can specify a value of “NCreq” that is much too large because macro procedure “LC_NCreq” will automatically reset the numerical value of “NCreq” to the **number of distinct clusters actually found**. Because the LC Phase One summary dataset (named as specified by the “LC_UnBias” macro variable) will contain only a single row in this case, plotting this single “point” with macro procedure “LC_UBtrace” is unlikely to be worthwhile.

Finally, macro procedure “LC_Salient” implements **LC Phase Two Tactics**:

[2] Generate the random, “artificial” LTD (or aLTD) distribution that corresponds to an observed LTD distribution with a previously specified value of “NCreq” = (number of clusters requested), and output a PDF that compares the aLTD and observed LTD distributions using both histograms and also overlaid empirical Cumulative Distribution Functions (eCDFs.) When these two distributions are clearly different, the observed LTD distribution is said to be “salient” (meaningfully different from random.)

WARNING: Creation of the overlaid eCDF graph requires SAS version 9.2. The output PDF file will contain only histograms for the observed LTD and aLTD distributions (and the log file will contain an ERROR message) when the “LC_aLTDdist” macro procedure is run under SAS 9.1.

3. Local Control Macro Variables

All nineteen of the SAS macro variables that may be needed to specify LC analyzes are listed and explained in this section of the User’s Guide.

3.01 Macro Variable LC_Path

Variable “LC_Path” specifies a SAS libname for the location where the input dataset is stored.

Example: `LC_Path = pcidata,`

In SAS for Unix, the corresponding libname statement could be something like:

```
LIBNAME pcidata "/omop_home/bobenchain/dev";
```

In SAS for Windows, that libname statement could be simply:

```
LIBNAME pcidata "D:\data";
```

3.02 Macro Variable LC_YTXdata

Variable “LC_YTXdata” specifies the name of the SAS dataset that contains the Y-outcome variable(s), the T-treatment binary (0-1) indicator, and the patient baseline X-characteristic variables (as columns) and patients as observations (rows.)

Example: `LC_YTXdata = pci15k,`

Note that the SAS dataset filename extension [.sas7bdat] is not specified when the “LC_YTXdata” macro variable is created and/or reset to a new value.

In a UNIX environment, all SAS datasets are stored with filenames that contain only digits and lower case letters. Thus any upper case letters used to define the contents of the “LC_YTXdata” macro variable are ignored within the UNIX environment.

3.03 Macro Variable LC_Yvar

Variable “LC_Yvar” specifies the SAS variable name of the Y-outcome variable. This name must correspond to an existing variable within the SAS dataset named by the “LC_YTXdata” macro variable.

Example: `LC_Yvar = mort6mo,`

3.04 Macro Variable LC_T01var

Variable “LC_T01var” specifies the SAS variable name of the T-treatment **binary** indicator. This name must correspond to an existing variable within the SAS dataset named by the “LC_YTXdata” macro variable.

Example: `LC_T01var = trtm,`

Note that level 1 usually denotes the new “treatment” while level 0 denotes the standard treatment or “control.” In all cases, Local Treatment Differences (LTDs) are always computed as [mean Y-outcome for patients receiving treatment type 1] minus [mean Y-outcome for patients receiving treatment type 0.]

3.05 Macro Variable LC_Xvars

Variable “LC_Xvars” specifies the SAS variable names of the patient baseline X-characteristic variables. These names must correspond to existing variables within the SAS dataset named by the “LC_YTXdata” macro variable.

Example:

```
LC_Xvars = stent height female diabetic
                     acutemi ejfract ves1proc,
```

The SAS X-variables may be of mixed types. In the above example, variables stent, female, diabetic and acutemi are binary (1 => yes, 0 => no); variables height and ejfract are continuous; and variable ves1proc (number of vessels involved in the patient's first PCI) is probably best viewed as being ordinal with 6 levels (0 to 5.) However, when the default patient dissimilarity metric of Euclidean distance is used, the ves1proc variable would be interpreted as being a continuous (interval) measure.

3.06 Macro Variable LC_PatID

Variable "LC_PatID" specifies the SAS variable name of the Patient Identification variable that will be **created or overwritten** by macro LC_Cluster. This "new" ID variable simply **numbers patients sequentially**. This macro variable only needs to be set when using **hierarchical clustering** in LC (i.e. when invoking macro procedures "LC_Cluster" and "LC_LTDdist".)

Example:

```
LC_PatID = sequen_id,
```

3.07 Macro Variable LC_ClusMeth

Variable "LC_ClusMeth" must specify a valid **METHOD=** argument for SAS proc CLUSTER. This macro variable needs to be set only when using **hierarchical clustering** in LC (i.e. when using macro procedures "LC_Cluster", "LC_LTDdist" and "LC_UBtrace".)

Example:

```
LC_ClusMeth = ward,
```

The default method is WARD; viable alternatives include AVERAGE (AVE), CENTROID (CEN), COMPLETE (COM), FLEXIBLE with BETA=–0.25 (FLE), MCQUITTY (MCQ) and MEDIAN (MED). The SINGLE linkage method is not recommended for use in LC analysis, while the DENSITY, EML and TWOSTAGE clustering methods tend to be too slow for use in LC. None of these hierarchical methods scale up well to very large numbers of patients (more than, say, 100,000).

SAS proc FASTCLUS performs "K-means" clustering (non-hierarchical) and can work well with very large numbers of patients IFF the number of clusters requested is large enough (and the X-values are "coarse" enough) to identify **exact X-matches**.

3.08 Macro Variable LC_Stand

Variable “LC_Stand” must specify a valid **METHOD=** argument for SAS proc STDIZE. This macro variable needs to be set only when using **hierarchical clustering** in LC (i.e. when using macro procedures “LC_Cluster”, “LC_LTDdist” and “LC_UBtrace”.)

Example: `LC_Stand = STD,`

The default option here is STD, which causes all X-variables to be translated to have mean zero and rescaled to have variance one. Alternative values useful in LC analyses include RANGE, MIDRANGE, MAXABS, IQR and MAD.

3.09 Macro Variable LC_Tree

Variable “LC_Tree” specifies the filename for the SAS permanent dataset describing the hierarchical “tree” structure that is output by proc CLUSTER. This macro variable needs to be set only when using **hierarchical clustering** in LC (i.e. when using macro procedures “LC_Cluster”, “LC_LTDdist” and “LC_UBtrace”.)

Example: `LC_Tree = pcim15dendog,`

3.10 Macro Variable LC_LTDtable

Variable “LC_LTDtable” specifies the filename for the SAS dataset containing summary statistics for each of the “NCreq” clusters created by invoking macro procedure “LC_LTDdist.”

Example: `LC_LTDtable = pcim15tab12h,`

3.11 Macro Variable LC_LTDoutput

Variable “LC_LTDoutput” specifies the filename for the (very large) SAS dataset resulting from **merging** relevant columns of the “LC_YTXdata” dataset with the “LC_LTDtable” dataset created by macro procedure “LC_LTDdist” …after both datasets have been **sorted by cluster number.**

Example: `LC_LTDoutput = pcim15out12h,`

3.12 Macro Variable LC_UnBias

Variable “LC_Unbias” specifies the filename of the SAS output dataset (containing summary statistics averaged across clusters) that is either created by (and later augmented by) calls to macro procedure “LC_LTDdist” (hierarchical case) or else created by macro procedure “LC_NCreq” (non-hierarchical case.)

Example: `LC_UnBias = pcim15ubtr,`

3.13 Macro Variable LC_swidth

Variable “LC_swidth” specifies the half-width (in “ltdsehom” units) for the confidence band around the TRACE display of “ltdavg” versus the logarithm of the number of clusters requested (NCreq.)

Example: `LC_swidth = 2.0,`

3.14 Macro Variable LC_pdftrace

Variable “LC_pdftrace” specifies a quoted string specifying the Path, Filename, and “.pdf” extension for the output file created by invoking macro procedure “LC_UBtrace.”

Example: `LC_pdftrace = "/omop_home/bobenchain/dev/pcimtrace.pdf",`

In SAS for Windows, the above example string could be:

```
LC_pdftrace = "D:\data\pcimtrace.pdf",
```

3.15 Macro Variable LC_aLTDreps

Variable “LC_aLTDreps” specifies the number of replications to be used to simulate the “artificial LTD distribution.” In each such replication, all patients are assigned randomly to one of a fixed set of mutually exclusive and exhaustive clusters.

Example: `LC_aLTDreps = 25,`

3.16 Macro Variable LC_seed

Variable “LC_seed” specifies a numerical “initial seed” value for the SAS pseudo-random number generator.

Example: `LC_seed = 1234567,`

To reproduce the output from an earlier invocation of “LC_Salient”, it is essential to use the very **same (positive) initial seed value**. To augment the output from an earlier invocation of “LC_Salient” (essentially performing “LC_aLTDreps” additional, independent replications), it is essential to use a **different initial seed value**. A simple way to assure this is to specify **a negative value for “LC_Seed”** because this signals the SAS pseudo-random number generator to use a **random (positive) initial seed value**.

3.17 Macro Variable NCinform

Variable “NCinform” contains the **user specified value** for “number of informative clusters” that result when “NCreq” clusters are requested. This value is an input for macro procedure “LC_Salient” so that it can be displayed in the subtitle below the graphical display where the Observed LTD distribution and its corresponding aLTD distribution are to be visually compared. The appropriate pairings of values for “NCreq” and “NCinform” are displayed in the LC Phase One output from macro procedure “LC_UBtrace.” These pairings are also recorded in the SAS output dataset written by multiple invocations of macro procedure “LC_LTDdist”. The name of this dataset is contained in the “LC_UnBias” macro variable; the corresponding variable names there are “_FREQ_” and “siclust”.

Example:

```
NCinform = 1112,
```

3.18 Macro Variable LC_aLTDdist

Variable “LC_aLTDdist” specifies the name of the SAS dataset output by macro procedure “LC_Salient”.

Example:

```
LC_aLTDdist = pcim15altd,
```

3.19 Macro Variable LC_pdfaltdd

Variable “LC_pdfaltdd” specifies a quoted string specifying the Path, Filename, and “.pdf” extension for the output file created by invoking macro procedure “LC_Salient.”

Example:

```
LC_pdfaltdd = "/omop_home/bobenchain/dev/pcimaltdd.pdf",
```

In SAS for Windows, the above example string could be:

```
LC_pdfaltdd = "D:\data\pcimaltdd.pdf",
```

4. Local Control Output Datasets

This section describes the permanent SAS datasets created by the LC macro procedures and defines the names and contents of the variables in those datasets.

4.01 LC Output Dataset with name specified by the “LC_Tree” Macro Variable

This is a highly specialized permanent dataset output by SAS proc CLUSTER that describes the full hierarchical clustering tree (dendrogram) using the “LC_Xvars” variables for all patients in the “LC_YTXdata” dataset. Invocations of macro procedure “LC_LTDdist” need this dataset to call SAS proc TREE to create the dataset that assigns individual patients to “NCreq” mutually exclusive and exhaustive clusters in X-space.

4.02 LC Output Dataset with name specified by the “LC_UnBias” Macro Variable

This dataset contains LC summary statistics that are generated by multiple calls to the “LC_LTDdist” macro with the same values for the “LC_Yvar” and “LC_T01var” macro variables.

- If a dataset with the name specified by the “LC_UnBias” macro variable does not exist when macro LC_LTDdist is invoked, a dataset with one observation (row) is created.
- If the dataset named by the “LC_UnBias” macro does exist when macro LC_LTDdist is invoked, one observation (row) is appended to that existing dataset.
- To start over accumulating summary statistics via invocations of the “LC_LTDdist” macro procedure, the user must first delete (or rename) the summary dataset named by the “LC_UnBias” macro variable.

While this dataset contains a total of 18 variables (columns), the 5 most important (key) variables are as follows:

FREQ = The Number of Clusters Requested (**NCreq** value) used to generate the current row of summary statistics.

siclust = number of **Informative Clusters** when “NCreq” clusters are requested.

sicpats = total number of patients within all resulting, informative clusters.

ltdavg = overall patient-weighted mean of the LTD distribution (treatment main-effect estimate.)

ltdsehom = estimated standard error of ltdavg when variances are assumed to be homogeneous across both treatments and clusters.

The 13 additional (secondary) variables contained in the dataset named by the “LC_UnBias” macro are:

TYPE = always set to 0

svrhom = sum of variance estimator numerator terms assuming homoscedasticity.

svrden = sum of variance estimator denominator terms …sum of (n0 + n1).

late = overall average Local Average Treatment Effect. (Note that his summary statistic does not vary when NCreq changes; all clusters are informative about local late effects.)

sn0 = overall number of patients taking trtm = 0. (Also does not vary with NCreq.)

sis0 = number of clusters that are informative about variability in outcome when taking trtm = 0 (i.e. number of clusters containing at least 2 trtm = 0 patients.)

ssos0 = sum of within cluster adjusted sum-of-squares for trtm = 0 outcomes.

sn1 = overall number of patients taking trtm = 1. (Also does not vary with NCreq.)

sis1 = number of clusters that are informative about variability in outcome when taking trtm = 1 (i.e. number of clusters containing at least 2 trtm = 1 patients.)

ssos1 = sum of within cluster adjusted sum-of-squares for trtm = 1 outcomes.

sigma = overall estimate of the standard deviation in observed outcomes when variances are assumed to be homogeneous across both treatments and clusters.

sigma0 = estimated standard deviation in trtm = 0 observed outcomes when variances are assumed to be homogeneous across clusters.

sigma1 = estimated standard deviation in trtm = 1 observed outcomes when variances are assumed to be homogeneous across clusters.

4.03 LC Output Dataset with name specified by the “LC_LTDtable” Macro Variable

While this dataset also contains a total of 18 variables (columns), the 4 most important (key) variables are as follows:

CLUSTER = cluster number between 1 and the value of “NCreq” specified when macro procedure “LC_LTDdist” was invoked.

ltd = Local Treatment Difference (difference in within-cluster outcomes, trtm = 1 minus trtm = 0) for a cluster.

$$\text{LTD: } \overline{y}_1 - \overline{y}_0$$

late = Local Average Treatment Effect (average within-cluster outcome disregarding treatment choice) within a cluster.

$$\text{LATE: } \left[n_1 \overline{y}_1 + n_0 \overline{y}_0 \right] / \left(n_1 + n_0 \right)$$

ltdvrhom = estimated variance of the ltd when variances are assumed to be homogeneous across both treatments and clusters.

ltdsehet = estimated standard error of the ltd when variances are assumed to be homogeneous across clusters but different for the two treatment cohorts.

The 14 additional (secondary) variables contained in the dataset named by the “LC_LTDtable” macro variable are:

n1 = number of trtm = 1 patients contained in cluster.

ybar1 = average Y-outcome for trtm = 1 patients within cluster.

var1 = variance of the Y-outcomes of trtm = 1 patients within cluster.

vbar1 = variance of ybar1 for the trtm = 1 patients within cluster.

n0 = number of trtm = 0 patients contained in cluster.

ybar0 = average Y-outcome for trtm = 0 patients within cluster.

var0 = variance of the Y-outcomes of trtm = 0 patients within cluster.

vbar0 = variance of ybar0 for the trtm = 0 patients within cluster.

sos0 = adjusted sum-of-squares for the Y-outcomes of trtm = 0 patients within cluster.

is0 = 0-1 flag indicating whether this cluster is informative about Y-outcome variability for trtm = 0 within the cluster (i.e. n0 > 1.)

sos1= adjusted sum-of-squares for the Y-outcomes of trtm = 1 patients within cluster.

is1 = 0-1 flag indicating whether this cluster is informative about Y-outcome variability for trtm = 1 within the cluster (i.e. n1 > 1.)

iclust = 0-1 flag indicating whether this cluster is informative about the ltd (i.e. n1 > 0 and n0 > 0.)

ltdvrden = ltd variance denominator assuming homoscedasticity = n1 + n0.

4.04 LC Output Dataset with name specified by the "LC_aLTDdist" Macro Variable

This SAS dataset contains the sample LTD outcomes that characterize the full "artificial" LTD distribution simulated by macro procedure "LC_Salient". This dataset will always contain two variables (named "ltd" and "freq") and the number of observations (rows) will equal "LC_aLTDreps" times "NCinform" = Number of Informative Clusters when "NCreq" clusters are requested. (Variable "siclust" in the SAS dataset named in the "LC_UnBias" macro, described in Section **§4.02,** also contains the NCinform values corresponding to all values of _FREQ_ = NCreq.)

5. Example Use of LC macros for Hierarchical Clustering

Example dataset "pci15k.sas7bdat" contains simulated data for 15,487 patients who underwent a PCI cardiovascular procedure and were treated with either "usual care alone" (trtm = 0) or else with "usual care augmented with a hypothetical blood thinner" (trtm = 1). This "benchmark" pseudo-dataset may be freely distributed for use in research and training on methods of analysis of observational data.

The ten variables contained in this dataset consist of two Y-outcome variables, a treatment indicator variable and seven patient X-characteristic variables:

Y1: mort6mo = Binary 6-month mortality indicator.

Y2: cardcost = Cumulative 6-month cardiac related charges.

T: trtm = Binary indicator (1 => treated, 0 => control).

X1: stent = Binary indicator (1 => coronary stent deployment, 0 => no)

X2: height = Patient height rounded to the nearest centimeter.

X3: female = Binary sex indicator (1 => yes, 0 => male.)

X4: diabetic = Binary indicator (1 => diabetes mellitus, 0 => no.)

X5: acutemi = Binary indicator (1 => acute myocardial infarction within the previous 7 days, 0 => no.)

X6: ejecfrac = Left ejection fraction % rounded to integer.

X7: ves1proc = Number of vessels involved in initial PCI.

Acknowledgement: This simulation was motivated by the dataset used by Kereiakes et al. (2000). That dataset contained patient registry data and follow-up outcomes for 997 patients who received their initial PCI in 1997 or 1998 from an Interventionist associated with the Lindner Center in Cincinnati, OH. I mimicked the observed confounding among the seven patient X-characteristic variables and the treatment selection imbalance from that data in this simulation.

Table 5.1: Example Invocation of LC Macros for Hierarchical Clustering

```
***********************************************************************************
Example Dataset:  15,487 PCI patients with or without a hypothetical blood thinner.
Local Control Phase One:  Invoke macro "LC_Cluster, then make a sequence of calls
to "LC_LTDdist” for LC_Yvar = mort6mo and increasing numbers of clusters.
Finally, finish LC Phase One by invoking macro “LC_UBtrace.
Local Control Phase Two:  Invoke macro "LC_Salient" for NCreq = 1200.
***********************************************************************************;
***    Copyright (c) 2009 Foundation for the National Institutes of Health (FNIH).
***********************************************************************************;

LIBNAME pcidata  "/omop_home/bobenchain/dev";
LIBNAME pdftrace "/omop_home/bobenchain/dev/pcimtrace.pdf";
LIBNAME pdfaltdd "/omop_home/bobenchain/dev/pcimaltdd.pdf";
OPTIONS sasautos = ("/omop_home/bobenchain/dev/SAS" sasautos) mautosource;

*** Local Control Phase One (EXPLORE) **********;

%LC_Cluster(LC_Path = pcidata, LC_YTXdata = pci15k, LC_Tree = pcitree,
          LC_ClusMeth = ward, LC_Stand = STD, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id)

%LC_LTDdist(NCreq = 1, LC_LTDtable = pcimtab3h, LC_LTDoutput = pcimout3h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

*** Overwrite pcimtab3h and pcimout3h datasets until NCreq reaches 300 *****;

%LC_LTDdist(NCreq = 10, LC_LTDtable = pcimtab3h, LC_LTDoutput = pcimout3h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_LTDdist(NCreq = 50, LC_LTDtable = pcimtab3h, LC_LTDoutput = pcimout3h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_LTDdist(NCreq = 100, LC_LTDtable = pcimtab3h, LC_LTDoutput = pcimout3h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_LTDdist(NCreq = 300, LC_LTDtable = pcimtab3h, LC_LTDoutput = pcimout3h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_LTDdist(NCreq = 600, LC_LTDtable = pcimtab6h, LC_LTDoutput = pcimout6h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_LTDdist(NCreq = 900, LC_LTDtable = pcimtab9h, LC_LTDoutput = pcimout9h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_LTDdist(NCreq = 1200, LC_LTDtable = pcimtab12h, LC_LTDoutput = pcimout12h,
          LC_Path = pcidata, LC_Tree = pcitree, LC_YTXdata = pci15k,
          LC_Yvar = mort6mo, LC_T01var = trtm, LC_Xvars = stent height female diabetic
          acutemi ejfract ves1proc, LC_PatID = sequen_id, LC_UnBias = pcimubtr)

%LC_UBtrace(LC_Path = pcidata, LC_UnBias = pcimubtr, LC_swidth = 2.0,
          LC_pdftrace = “/omop_home/bobenchain/dev/pdftrace.pdf”)

*** Local Control Phase Two (CONFIRM) **********;

%LC_Salient(LC_Path = pcidata, LC_LTDoutput = pcimout12h, LC_Yvar = mort6mo,
          LC_T01var = trtm, LC_aLTDreps = 25, LC_seed = 1234567,
          LC_aLTDdist = pcimaltd12h, NCinform = 1112,
          LC_pdftrace = “/omop_home/bobenchain/dev/pdfaltdd.pdf”)

********************* END *********************************************************;
```

The PDF output files generated by the above invocations of LC macro procedures “LC_UBtrace” and “LC_Salient” when “NCreq” = 1200 for outcome “LC_Yvar” = mort6mo are collected together at the end of this User Guide. The corresponding outputs for outcome “LC_Yvar” = cardcost are also collected there for completeness.

A summary of the findings from these analyzes is as follows:

When problems with treatment cohort imbalance and X-covariate confounding within this dataset are ignored, the rate of mortality within 6 months for patients receiving trtm = 1 (hypothetical blood thinner) is 2.5% lower than that for patients receiving trtm = 0 (usual care alone.) On the other hand, the accumulated cardiac related costs within 6 months appear to be more than $500 higher for trtm = 1 than for trtm = 0.

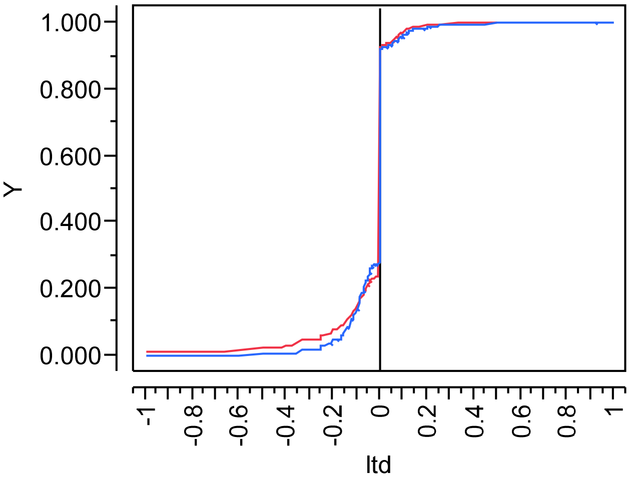
After LC adjustment for treatment cohort imbalance and X-covariate confounding, the 6 month mortality rate for trtm = 1 (hypothetical blood thinner) is seen to actually be 3.9% to 4.1% lower than that of trtm = 0 (usual care alone.) Furthermore, the accumulated cardiac related costs within 6 months are actually $60 to $150 lower for trtm = 1 than for trtm = 0. In other words, LC adjustment for treatment cohort imbalance and X-covariate confounding consistently yields outcome comparisons that are more favorable to trtm = 1 than before adjustment.

The observed LTD distributions for “NCreq” = 1200 are also clearly salient. In other words, making comparisons only within clusters of patients who are relatively well-matched on seven baseline X-covariates (stent, height, female, diabetic, acutemi, ejfract and ves1proc) has been shown to be “meaningful”; the comparisons resulting from this clustering are clearly different (less biased) from those resulting from random clusterings.

Notes: There is no need to invoke “LC_Cluster” a second time to analyze a second Y-outcome (cardcost) associated with the same baseline X-covariates on the same patients as the first Y-outcome (mort6mo.) One simply makes multiple invocations of macro procedure “LC_LTDdist” with the new value of “LC_Yvar” and increasing values of “NCreq”. Unless there are missing values in the two Y-outcome variables, the same numbers of informative clusters will result. Upon invoking macro procedure “LC_UBtrace”, the new PDF output may show (as here) that the Unbiasing Trace has different “volatility” for intermediate values of “NCreq”, but the trace should still ultimately “stabilize” at roughly the same (larger) values of “NCreq” as before.

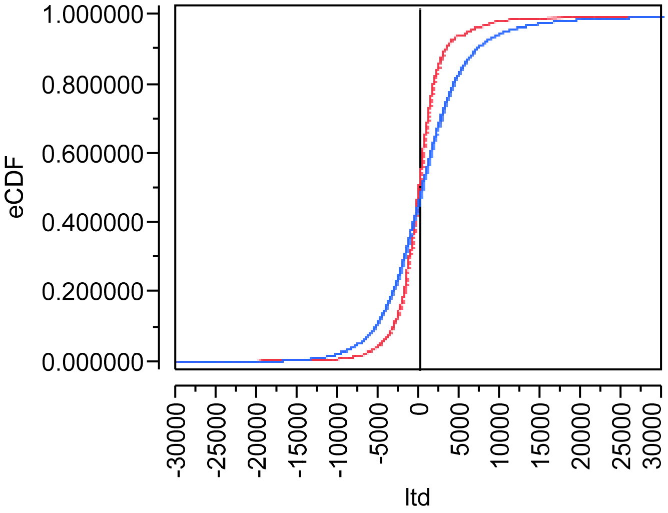
Table 5.2 Comparison of eCDFs by Importing SAS Output into JMP.

Obs. LTD and aLTD distributions of mort6mo



The **Observed LTD** distribution for mort6mo has a slightly thinner upper (right-hand) tail of outcomes unfavorable to trtm = 1 than does the **artificial LTD** distribution. The **Observed LTD** distribution also has a (mostly) thicker lower (left-hand) tail of outcomes favorable to trtm = 1 than does the **artificial LTD** distribution. The only exception here is the small area of negative LTDs before the two eCDFs cross again at LTD = –0.083. In both distributions, the probability of an **exactly zero LTD** is quite large, at least 0.65.

Obs. LTD and aLTD distributions of cardcost



While the **Observed LTD** distribution for cardcost has a smaller mean value than does the **artificial LTD** distribution, that relationship is difficult to “see” in the eCDFs at left.

What is quite easy to see here is that the **Observed LTD** distribution of cardcost has **much lower variance** than does the **artificial LTD** distribution.

WARNING: On different computing platforms, SAS proc CLUSTER may give different clusters …i.e. clusters of different sizes, patient constitution or treatment fractions. For example, when NCreq = 1200 for the above pci15k dataset, SAS 9.1 for Unix produces 1112 informative clusters, while SAS 9.1 for Windows produces only 1106.

6. Example of Use of the LC_NCreq macro for non-hierarchical Clustering

The simulated dataset “statin1m.sas7bdat” contains ***1 million patients*** supposedly taking one of two hypothetical statins. This “benchmark” pseudo-dataset may be freely distributed for use in research and training on methods of analysis of observational data.

The LC Phase One macros that perform conventional **hierarchical clustering** with SAS proc CLUSTER (e.g. Ward's method) have no hope of scaling up to a dataset with this many patients. On the other hand, this dataset consisting of only binary (0-1) variables is ideal for use of SAS proc FASTCLUS as long as the number of clusters requested is large enough. Because all 7 of the patient X-characteristics in this dataset are binary, at most 2^7 = 128 X-space clusters are possible. Requesting this many clusters assures that each non-empty cluster will contain only ***exact X-matches***!

Dataset Location: /omop_home/bobenchain/dev/statin1m.sas7bdat

Role Played by Variable	Name of Variable	Definition of Variable
Y-outcome:	CVE = Cardiovascular Event (0=No, 1=Yes) after starting on a Statin.	
Treatment Indicator:	TRTM	where 0 => Newer Statin (~210K patients), 1 => Established Statin (~790K patients)
X1:	AGE60 = Age at least 60 years, where 0 => No and 1 => Yes	
X2:	FEMALE, where 0 => No and 1 => Yes	
X3:	HYPN = Hypertension, where 0 => No and 1 => Yes	
X4:	DIAB = Diabetes, where 0 => No, 1 => Yes	
X5:	APLAT = on an Antiplatlet drug, where 0 => No and 1 => Yes	
X6:	CVPR = Prior CVE (before statin), where 0 => No and 1 => Yes	
X7:	MIPR = Prior Miocardial Infarction, where 0 => No, 1 => Yes (and CVPR = 1.)	

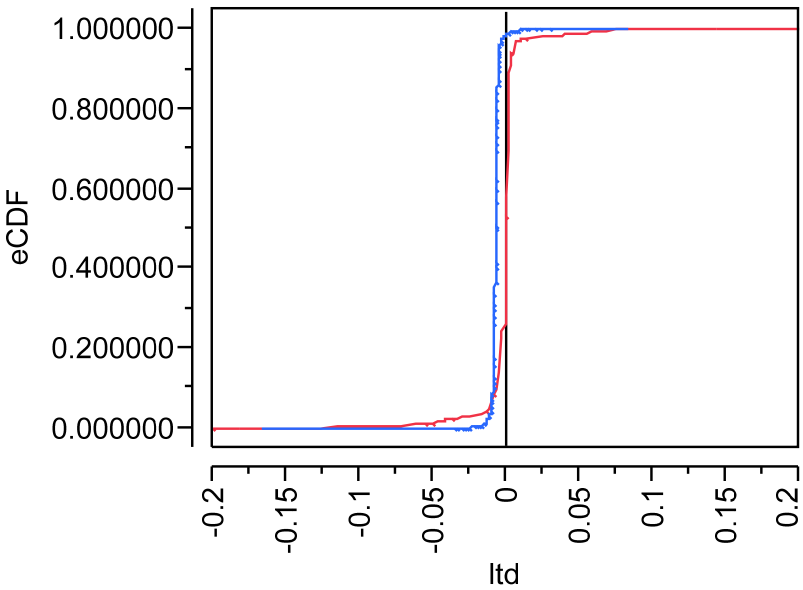
Note: The statin1m dataset actually contains only 96 non-empty clusters because MIPR = 1 is only possible when CVPR = 1. Furthermore, 94 of these 96 clusters is “informative” about a Local Treatment Difference (LTD) in the sense that each contains at least one patient taking TRTM = 1 as well as at least one patient taking TRTM = 0.

Acknowledgement: This simulation was motivated by the dataset used by Michael Gaffney and Jack Mardekian of Pfizer in their talk “Propensity Scores in Observational Studies,” given at the Joint Statistical Meetings in Washington, DC, on August 2, 2009. Their dataset contained retrospective data on 83 thousand patients, and I mimicked the observed confounding among the seven patient X-characteristic variables and the treatment selection imbalance from that data in this simulation.

Table 6.1: Example Invocation of the LC Macros in a UNIX Environment

```
***********************************************************************************
Local Control Phase One:  Invoke macro "LC_NCreq" for non-hierarchical clustering.
Local Control Phase Two:  Invoke macro "LC_Salient"
***********************************************************************************;
***    Copyright (c) 2009 Foundation for the National Institutes of Health (FNIH).
***********************************************************************************;

LIBNAME st1mdata "/omop_home/bobenchain/dev";
OPTIONS sasautos = ("/omop_home/bobenchain/dev/SAS" sasautos) mautosource;

%LC_NCreq(LC_Path = st1mdata, LC_YTXdata = statin1m, LC_LTDoutput = st1mout,
          LC_T01var = TRTM, LC_Yvar = CVE, LC_Xvars = AGE60 FEMALE HYPN DIAB APLAT
          CVPR MIPR, LC_LTDtable = st1mtab, LC_Unbias = st1mubtr, NCreq = 128)

proc print data = st1mdata.st1mubtr;
  title "LC Summary Statistics";
run;

%LC_Salient(LC_Path = st1mdata, LC_LTDoutput = st1mout, LC_T01var = TRTM,
          LC_Yvar = CVE, LC_aLTDreps = 5, LC_seed = 33, LC_aLTDdist = st1maltd94,
          NCinform = 94, LC_pdfaltdd = "/omop_home/bobenchain/dev/st1maltdd.pdf")

********************* END *********************************************************;
```

Table 6.2 Comparison of eCDFs by Importing SAS Output into JMP.



It is easy to see here that LC adjustment for treatment cohort imbalance and X-covariate confounding has essentially shifted the **Observed LTD** distribution of CVE to the right relative to the (unadjusted) **artificial LTD** distribution. This shift makes it rather clear that there are no meaningful differences in cardiovascular (adverse) event rates between these two hypothetical statins.

7. Summary: Advantages of Local Control over Traditional Modeling

The Local Control (LC) approach follows Arnold Zellner's (1991) "KISS" principle:

Keep It Sophisticatedly Simple.

7.1 The LC approach is not esoteric or mysterious; it is easy to explain and illustrate (graphically) to nontechnical audiences.

The overall LC strategy is to use well-established clustering methodologies to identify subgroups of patients who are relatively well matched on their baseline ***x***-characteristics and to make treatment ***y***-outcome comparisons only "within" these clusters. In the limit as clusters become small, compact and numerous, the "LC Trace" display shows (rather dramatically) that treatment comparisons thereby become less biased and more and more "fair" (objective, scientific.)

7.2 The LC approach reveals and quantifies the full distribution of patient differential response to treatment.

Estimation of only treatment "**main effects**" is a woefully inadequate strategy when so much more information is badly needed for comparative effectiveness analyses, targeted therapeutics, and practice of evidence based medicine. Almost no patients are "average" in all senses relevant to their health. Health care providers must start asking themselves: "Which treatment is better for THIS patient?" See Kent and Hayward (2007a, 2007b.) Fortunately, LC strategy provides some new "steps" in this important "direction!"

The LTD distribution of outcome differences due to treatment can be used and interpreted much like a Bayesian posterior distribution.

7.3 LC results are demonstrably more robust than those from traditional global, over-smooth parametric models.

An observed LTD distribution resulting from small, compact, and numerous clusters is comprised of many, many estimates from a simple Nested ANOVA model (treatment within cluster.) This statistical model makes so few assumptions, all of which are frequently reasonable or even realistic, that an LC analysis essentially ends up being non-parametric.

Oscar Kempthorne (197?): Analysis of Variance (ANOVA) is a Scientific Method;
Analysis of Covariance (a General Linear Model) is NOT …due to too
many, too strong assumptions about continuous predictor variables.
George Box (1979): All models are WRONG; some (robust models) are USEFUL.

7.4 Final Notes

The LC method of adjustment for treatment selection bias (patient channeling) and confounding in human health care studies is based upon patient clustering (unsupervised learning.) This is a form of "post hoc blocking" that makes treatment comparisons only among relatively well-matched patients.

The theoretical basis for LC is that cluster membership is guaranteed to become a "balancing score" that is finer (more detailed) than the unknown true propensity score in the limit as clusters become small, compact and numerous.

The dual overall strategies of the LC approach are (a) to use systematic sensitivity analyses to validate the observed Local (within cluster) Treatment Difference (LTD) distribution and (b) to use resampling (simulation) methods to show that this LTD distribution is "salient" (clearly different from what results from purely random patient clustering.)

8. References

Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. ***J Amer Stat Assoc*** 1996; 91: 444-472.

Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. ***Biometrics*** 2005; 61(4): 962–973.

Barlow HB. “Unsupervised learning.” ***Neural Computation*** 1989; 1: 295-311.

Box GEP. “Robustness is the Strategy of Scientific Model Building.” ***Robustness in Statistics.*** R.L. Launer and G.N. Wilkinson, eds. (Quote on page 202.) New York: Academic Press, 1979.

Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. ***Biometrics*** 1968; 24: 205-213.

D’Agostino RB Jr. Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. ***Stat Med*** 1998; 17, 2265-2281.

Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting clinical trial results to inform providers, payers and consumers: the need to assess benefits and harms for lower versus higher risk patients. ***Health Affairs*** 2003; 24: 1571–1581.

Kaufman L, Rousseeuw PJ. ***Finding Groups in Data. An Introduction to Cluster Analysis.*** New York: John Wiley and Sons. 1990.

Kempthorne O. (1919-2000; Iowa State, Statistics 1947-1989.) Kempthorne was widely known for his early work on randomization and design of experiments. In talks I heard him give at ASA meetings in the 1970s, I thought he sounded rather skeptical about the “objectivity” of extensions of ANOVA models that incorporate continuous predictor variables.

Kent DM, Hayward RA. When averages hide individual differences in clinical trials. ***American Scientist*** 2007; 95: 60–68.

Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. ***JAMA*** 2007; 298: 1209–1212.

Kereiakes DJ, Obenchain RL, Barber BL, Smith A, McDonald M, Broderick TM, Runyon JP, Shimshak TM, Schneider JF, Hattemer CH, Roth EM, Whang DD, Cocks DL, Abbottsmith CW. Abciximab provides cost effective survival advantage in high volume interventional practice. ***Am Heart J*** 2000; 140: 603-610.

Lunceford, JK, Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. ***Stat Med*** 2004; 23: 2937–2960.

McClellan M, McNeil BJ, Newhouse JP. Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables. ***JAMA*** 1994; 272: 859-866.

Obenchain RL. Nearest Neighbors Analysis for PRRAP, the Probable Report Rate Analysis Plan. **Bell Laboratories** TM-79-1711-4. 1979. Holmdel, NJ.

Obenchain RL. Unsupervised Propensity Scoring: NN and IV Plots. **2004 Proceedings of the American Statistical Association** (on CD.) 8 pages.

Obenchain RL. "USPS_1_01.zip: An R package for unsupervised and supervised propensity score adjustment." 2006b. http://www.math.iupui.edu/~indyasa/download.htm.

Obenchain RL. "The Local Control Approach using JMP." ***Analysis of Observational Health Care Data Using SAS*** (Chapter 7.) Faries DE, Leon AC, Maria Haro J, Obenchain RL eds. Cary, NC: SAS Press. February 2010.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. ***Epidemiology*** 2000; 11: 550–560.

Rosenbaum PR, Rubin RB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. ***Biometrika*** 1983; 70: 41-55.

Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on a Propensity Score. ***J Amer Stat Assoc*** 1984; 79: 516-524.

Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. ***Amer Statist*** 1985; 39: 33-38.

Rosenbaum PR. Optimal matching in observational studies. ***J Amer Stat Assoc*** 1989; 84: 1024-1032.

Rosenbaum PR. "Multivariate matching methods." In: Kotz S, Read CR, Banks D, eds. ***Encyclopedia of Statistical Sciences***, Update Volume 2. New York: J Wiley 1998: 435-438.

Rosenbaum PR. ***Observational Studies, Second Edition.*** New York: Springer-Verlag 2002.

Rubin DB. Bias reduction using Mahalanobis metric matching. ***Biometrics*** 1980; 36: 293-298.

Zellner A. American Statistical Association (ASA) Presidential Address. Atlanta, GA. 1991.

9. Appendix: The Patient-Weighted Average LTD and its Variance.

This Appendix will use the following notation:

y_{tjk} = Observed outcome on treatment t (0 or 1) for the k^{th} patient within cluster j.

$n_{.j}$ = Total number of patients within cluster j = $n_{1j} + n_{0j}$.

Δ_j = j^{th} Local Treatment Difference **(LTD)** = $\overline{y}_{1j} - \overline{y}_{0j}$ **(A1)**

= **difference in outcome mean values** (t = 1 minus t = 0.)

Note that this LTD is not well defined unless n_{1j} and n_{0j} are both at least one. In this case, the j^{th} cluster is said to be **informative.** Let the number of informative clusters be denoted by the symbol I, while the total number of clusters is denoted by the symbol K (which is $\geq I$.)

This appendix will treat only the case where observed outcomes are assumed to be independent and have constant variance (homoscedasticity) within each treatment cohort of patients (t = 1 or t = 0.) These two variances will be denoted here by σ_1^2 and σ_0^2, respectively, and are assumed to be estimated "locally" …i.e. measuring only variability about local mean outcomes within both clusters and treatment cohorts. In other words, n_{tj} needs to be at least 2 to provide information about σ_t^2. Furthermore, outcome variability information can be provided by clusters un-informative about LTDs. As a result, the total number of degrees-of-freedom for estimation of σ_1^2 will be at least $\Sigma\, n_{1j} - K$ while the degrees-of-freedom for estimation of σ_0^2 will be at least $\Sigma\, n_{0j} - K$. Finally, when σ_1^2 and σ_0^2, are assumed to be equal, their estimators can be pooled to provide an estimate of the common variance, σ^2. This combined estimator would have degrees-of-freedom equal to (total number of patients – K – I.)

Using the above notation, the variance of the j^{th} LTD can be written as:

$$V(\Delta_j) = \frac{\sigma_1^2}{n_{1j}} + \frac{\sigma_0^2}{n_{0j}} \qquad \textbf{(A2)}$$

$$= \sigma^2 \left(\frac{n_{1j} + n_{0j}}{n_{1j} \times n_{0j}} \right)$$

…assuming common cohort variances.

Since the remainder of our discussion here concerns combining information from only informative clusters, the range of the $\boldsymbol{j}$ subscript can be assumed, without loss of generality, to be $\boldsymbol{j} = 1, 2, \ldots, \boldsymbol{I}$.

The ordinary, unweighted mean if the LTDs is rarely of interest because **informative clusters can be of very different sizes**. However, we give the following expression for the common cohort variances case here, simply for completeness:

$$V(\text{Unweighted Mean}\,\overline{\Delta}) = \frac{\sigma^2}{I^2}\sum\left(\frac{n_{1j}+n_{0j}}{n_{1j}\times n_{0j}}\right)$$

If all informative clusters had $n_{1j} = n_{0j} = n$, the above variance would be $\mathbf{2\sigma^2/(I\times \mathit{n})}$.

The following weighted, across-cluster mean of the LTDs is the most appropriate LC estimate of the "**Main Effect of Treatment**." Again, because **informative clusters can be of very different sizes**, this estimate treats the total number of patients in a cluster, $n_{1j} + n_{0j}$, as the **frequency** of the LTD estimate for that cluster in the overall LTD distribution.

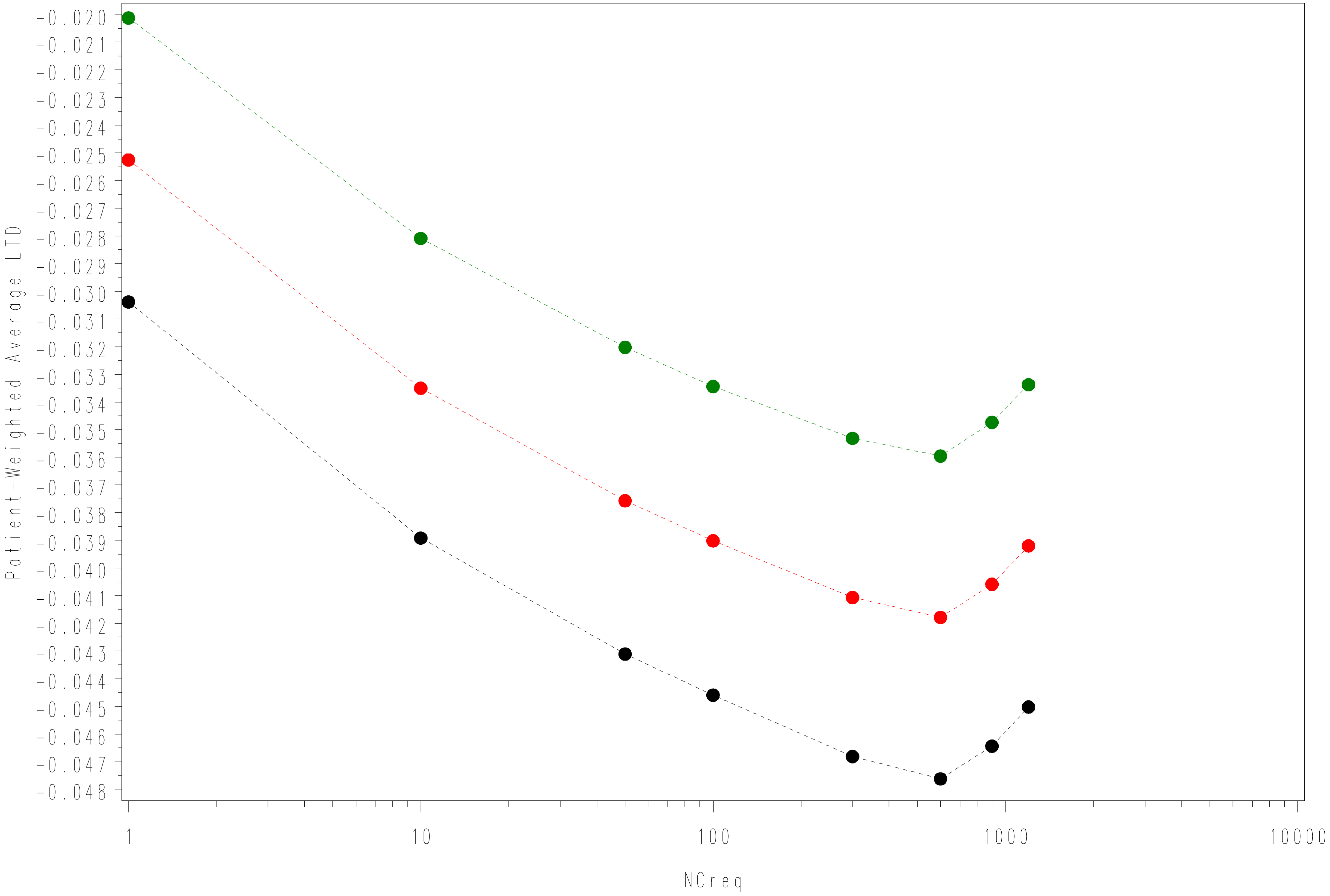
$$\overline{\overline{\Delta}}\cdot = \sum\left(n_{1j}+n_{0j}\right)\times\Delta_j \;/\; \sum\left(n_{1j}+n_{0j}\right). \qquad \textbf{(A3)}$$

The corresponding variance for the common cohort variances case is then:

$$V\left[\overline{\overline{\Delta}}\cdot\right] = \sigma^2\sum\frac{\left(n_{1j}+n_{0j}\right)^3}{\left(n_{1j}\times n_{0j}\right)} \;/\; \left[\sum\left(n_{1j}+n_{0j}\right)\right]^2. \qquad \textbf{(A4)}$$

Note that, if all informative clusters had $n_{1j} = n_{0j} = n$, this weighted variance would again reduce to $\mathbf{2\sigma^2/(I\times \mathit{n})}$. However, this case of equal cluster sizes with a fixed 1:1 ratio of $\boldsymbol{t} = 1$ and $\boldsymbol{t} = 0$ patients within each cluster is **highly unlikely** to occur in actual practice.

LC Unbiasing Trace Display



NCreq = number of clusters requested

Obs	NCreq	siclust	sicpats	sicppct	ltdavg	lolim	uplim	ltdsehom
1	1	1	15487	100.000	-0.025251	-0.030382	-0.020121	.002565305
2	10	10	15487	100.000	-0.033501	-0.038915	-0.028087	.002707121
3	50	50	15487	100.000	-0.037568	-0.043109	-0.032026	.002770952
4	100	100	15487	100.000	-0.039017	-0.044595	-0.033439	.002788902
5	300	299	15470	99.890	-0.041062	-0.046813	-0.035312	.002875078
6	600	592	15385	99.341	-0.041785	-0.047617	-0.035952	.002916156
7	900	873	15221	98.282	-0.040587	-0.046437	-0.034736	.002925217
8	1200	1112	14857	95.932	-0.039199	-0.045023	-0.033376	.002911735

NCreq = number of clusters requested

Comparison of LTD Distributions

The UNIVARIATE Procedure
Variable: ltd

Freq: freq
oa = aLTD

Moments			
N	371425	**Sum Weights**	371425
Mean	-0.0249565	**Sum Observations**	-9269.4573
Std Deviation	0.09830434	**Variance**	0.00966374
Skewness	-1.2616013	**Kurtosis**	26.4275223
Uncorrected SS	3820.67921	**Corrected SS**	3589.34626
Coeff Variation	-393.90321	**Std Error Mean**	0.0001613

Basic Statistical Measures			
Location		**Variability**	
Mean	-0.02496	**Std Deviation**	0.09830
Median	0.00000	**Variance**	0.00966
Mode	0.00000	**Range**	2.00000
		Interquartile Range	0.04348

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-154.72	**Pr > |t|**	<.0001
Sign	**M**	-37327	**Pr >= |M|**	<.0001
Signed Rank	**S**	-2.453E9	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	1.0000000
99%	0.2424242
95%	0.0769231
90%	0.0000000
75% Q3	0.0000000
50% Median	0.0000000
25% Q1	-0.0434783
10%	-0.1250000
5%	-0.1666667

Number of Informative Clusters = 1112

The UNIVARIATE Procedure
Variable: ltd
Freq: freq
oa = aLTD

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	-0.3333333
0% Min	-1.0000000

Extreme Observations					
Lowest			**Highest**		
Value	**Freq**	**Obs**	**Value**	**Freq**	**Obs**
-1	9	42622	1	13	40242
-1	8	42604	1	12	40796
-1	5	42269	1	22	41221
-1	6	41880	1	11	41363
-1	2	41542	1	7	41825

Number of Informative Clusters = 1112

The UNIVARIATE Procedure
Variable: ltd

Freq: freq
oa = obsLTD

Moments			
N	14857	**Sum Weights**	14857
Mean	-0.0391992	**Sum Observations**	-582.3826
Std Deviation	0.13839181	**Variance**	0.01915229
Skewness	-3.8805967	**Kurtosis**	21.7494339
Uncorrected SS	307.3554	**Corrected SS**	284.526464
Coeff Variation	-353.04748	**Std Error Mean**	0.00113539

Basic Statistical Measures			
Location		**Variability**	
Mean	-0.03920	**Std Deviation**	0.13839
Median	0.00000	**Variance**	0.01915
Mode	0.00000	**Range**	1.50000
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-34.5249	**Pr > |t|**	<.0001
Sign	**M**	-1232.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	-3472192	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	0.500000
99%	0.200000
95%	0.062500
90%	0.000000
75% Q3	0.000000
50% Median	0.000000
25% Q1	0.000000
10%	-0.142857
5%	-0.250000

Number of Informative Clusters = 1112

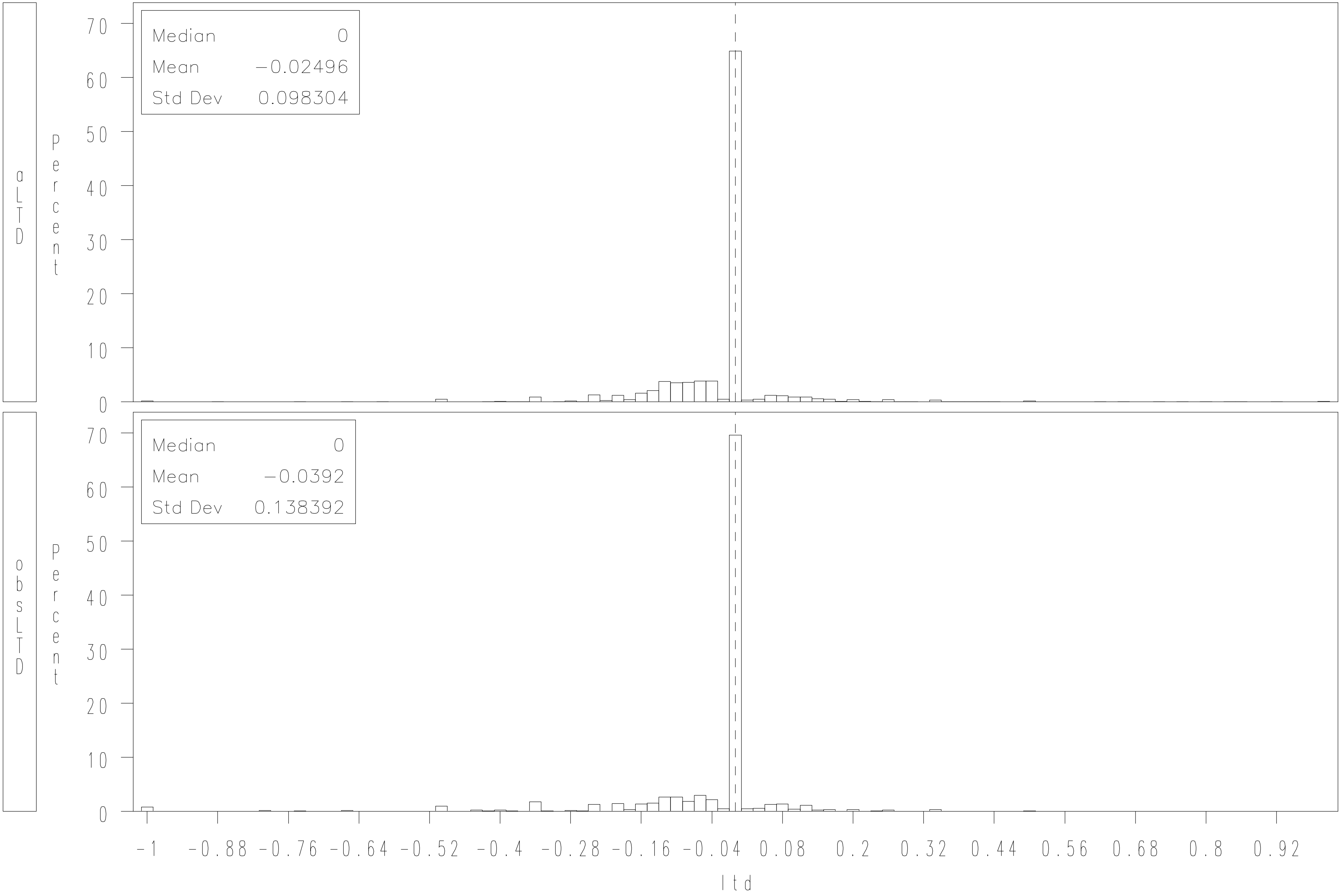
The UNIVARIATE Procedure
Variable: ltd
Freq: freq
oa = obsLTD

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	-0.800000
0% Min	-1.000000

Extreme Observations					
Lowest			**Highest**		
Value	**Freq**	**Obs**	**Value**	**Freq**	**Obs**
-1	1	14854	0.5	1	2721
-1	1	14853	0.5	1	2722
-1	1	14852	0.5	1	2723
-1	1	14851	0.5	1	2724
-1	1	14673	0.5	1	2725

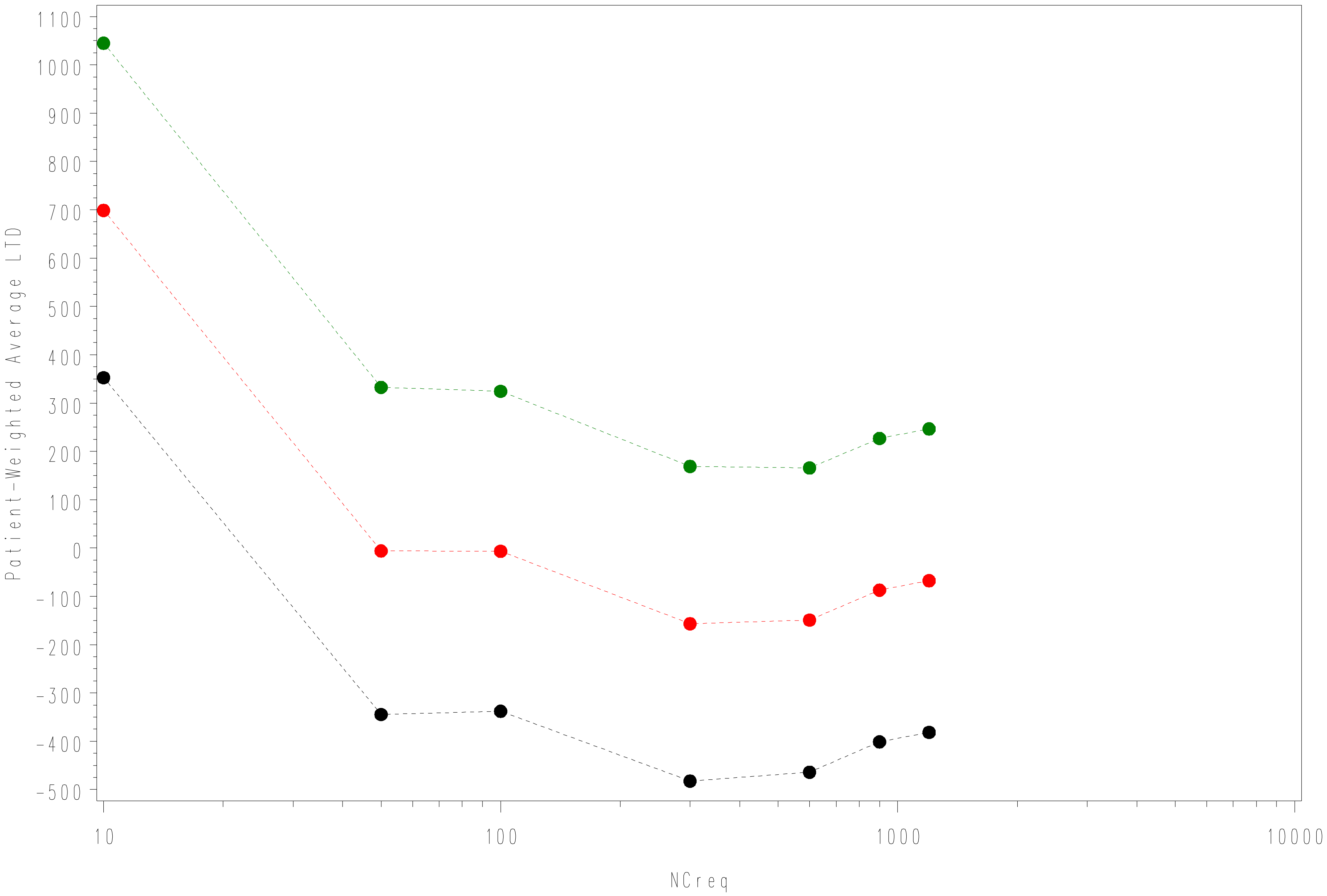
Number of Informative Clusters = 1112

Comparison of LTD Distributions



Number of Informative Clusters = 1112

LC Unbiasing Trace Display



NCrea = number of clusters reauested

LC Unbiasing Trace Display

Obs	NCreq	siclust	sicpats	sicppct	ltdavg	lolim	uplim	ltdsehom
1	10	10	15487	100.000	698.832	352.743	1044.92	173.044
2	50	50	15487	100.000	-5.980	-344.543	332.58	169.281
3	100	100	15487	100.000	-6.671	-337.843	324.50	165.586
4	300	299	15470	99.890	-156.721	-482.334	168.89	162.807
5	600	592	15385	99.341	-149.067	-463.939	165.81	157.436
6	900	873	15221	98.282	-87.108	-401.120	226.90	157.006
7	1200	1112	14857	95.932	-67.474	-381.539	246.59	157.033

NCreq = number of clusters requested

The UNIVARIATE Procedure
Variable: ltd

Freq: freq
oa = aLTD

Moments			
N	371425	**Sum Weights**	371425
Mean	526.730937	**Sum Observations**	195641038
Std Deviation	7016.05284	**Variance**	49224997.4
Skewness	2.90326867	**Kurtosis**	54.8534706
Uncorrected SS	1.83864E13	**Corrected SS**	1.82833E13
Coeff Variation	1331.99938	**Std Error Mean**	11.5121724

Basic Statistical Measures			
Location		**Variability**	
Mean	526.73	**Std Deviation**	7016
Median	144.95	**Variance**	49224997
Mode	-9718.19	**Range**	277360
		Interquartile Range	5845

Note: The mode displayed is the smallest of 25 modes with a count of 59.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	45.75426	**Pr > |t|**	<.0001
Sign	**M**	5157.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1.8401E9	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	163424.600
99%	22162.964
95%	9955.339
90%	6753.265
75% Q3	3139.507
50% Median	144.954
25% Q1	-2705.888
10%	-5734.612
5%	-7936.579

Number of Informative Clusters = 1112

The UNIVARIATE Procedure
Variable: ltd
Freq: freq
oa = aLTD

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	-15049.049
0% Min	-113935.348

Extreme Observations					
Lowest			**Highest**		
Value	**Freq**	**Obs**	**Value**	**Freq**	**Obs**
-113935.3	2	19302	155872	6	28127
-105643.3	9	41554	160532	7	37222
-102931.4	10	24005	161055	9	38377
-83781.0	4	41414	161864	4	39289
-76888.2	9	28904	163425	11	22703

Number of Informative Clusters = 1112

The UNIVARIATE Procedure
Variable: ltd

Freq: freq
oa = obsLTD

Moments			
N	14857	**Sum Weights**	14857
Mean	-67.473793	**Sum Observations**	-1002458.1
Std Deviation	5305.73807	**Variance**	28150856.5
Skewness	4.35279332	**Kurtosis**	74.3943759
Uncorrected SS	4.18277E11	**Corrected SS**	4.18209E11
Coeff Variation	-7863.4057	**Std Error Mean**	43.5291558

Basic Statistical Measures			
Location		**Variability**	
Mean	-67.474	**Std Deviation**	5306
Median	-192.911	**Variance**	28150857
Mode	112.675	**Range**	113537
		Interquartile Range	2941

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-1.55008	**Pr > |t|**	0.1211
Sign	**M**	-491.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	-5526493	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	77544.765
99%	14954.457
95%	5724.911
90%	2921.691
75% Q3	1217.166
50% Median	-192.911
25% Q1	-1723.619
10%	-3415.171
5%	-5130.331

Number of Informative Clusters = 1112

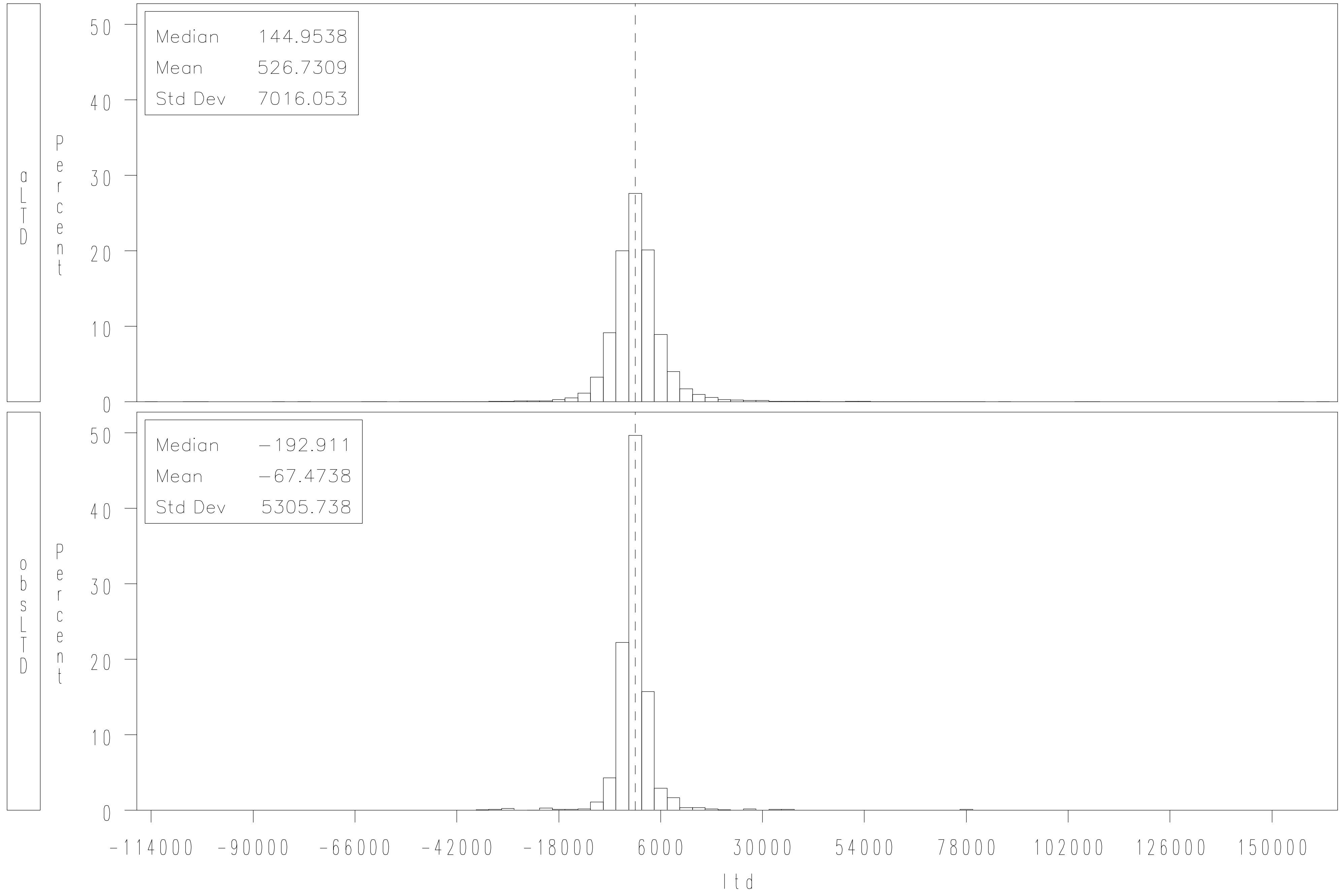
The UNIVARIATE Procedure
Variable: ltd
Freq: freq
oa = obsLTD

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	-10625.692
0% Min	-35992.286

Extreme Observations					
Lowest			**Highest**		
Value	**Freq**	**Obs**	**Value**	**Freq**	**Obs**
-35992.3	1	159	77544.8	1	76
-35992.3	1	158	77544.8	1	77
-35992.3	1	157	77544.8	1	78
-35992.3	1	156	77544.8	1	79
-35992.3	1	155	77544.8	1	80

Number of Informative Clusters = 1112

Comparison of LTD Distributions



Number of Informative Clusters = 1112

The UNIVARIATE Procedure
Variable: ltd

Freq: freq
oa = aLTD

Moments			
N	4999640	**Sum Weights**	4999640
Mean	-0.0071887	**Sum Observations**	-35941.076
Std Deviation	0.0027482	**Variance**	7.55259E-6
Skewness	-1.3181951	**Kurtosis**	80.8260676
Uncorrected SS	296.131012	**Corrected SS**	37.7602144
Coeff Variation	-38.229231	**Std Error Mean**	1.22908E-6

Basic Statistical Measures			
Location		**Variability**	
Mean	-0.00719	**Std Deviation**	0.00275
Median	-0.00715	**Variance**	7.55259E-6
Mode	-0.00888	**Range**	0.13889
		Interquartile Range	0.00179

Note: The mode displayed is the smallest of 5 modes with a count of 212938.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-5848.89	**Pr > |t|**	<.0001
Sign	**M**	-2432320	**Pr >= |M|**	<.0001
Signed Rank	**S**	-6.15E12	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	0.05555556
99%	0.00175439
95%	-0.00460200
90%	-0.00534510
75% Q3	-0.00645311
50% Median	-0.00714682
25% Q1	-0.00824140
10%	-0.00914754
5%	-0.00978416

Number of Informative Clusters = 94

The UNIVARIATE Procedure
Variable: ltd
Freq: freq
oa = aLTD

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	-0.01443841
0% Min	-0.08333333

Extreme Observations					
Lowest			**Highest**		
Value	**Freq**	**Obs**	**Value**	**Freq**	**Obs**
-0.0833333	120	1E6	0.0277778	48	1E6
-0.0833333	36	1E6	0.0347222	168	1E6
-0.0833333	72	1E6	0.0357143	120	1E6
-0.0833333	120	1E6	0.0416667	72	1E6
-0.0833333	36	1E6	0.0555556	48	999960

Number of Informative Clusters = 94

The UNIVARIATE Procedure
Variable: ltd

Freq: freq
oa = obsLTD

Moments			
N	999928	**Sum Weights**	999928
Mean	-0.0011262	**Sum Observations**	-1126.1422
Std Deviation	0.01575236	**Variance**	0.00024814
Skewness	-5.1707701	**Kurtosis**	158.996017
Uncorrected SS	249.38697	**Corrected SS**	248.118682
Coeff Variation	-1398.6888	**Std Error Mean**	0.00001575

Basic Statistical Measures			
Location		**Variability**	
Mean	-0.00113	**Std Deviation**	0.01575
Median	-0.00063	**Variance**	0.0002481
Mode	-0.00081	**Range**	0.76190
		Interquartile Range	0.00244

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	-71.493	**Pr > |t|**	<.0001
Sign	**M**	-31349	**Pr >= |M|**	<.0001
Signed Rank	**S**	-7.404E9	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	0.333333333
99%	0.052631579
95%	0.006226159
90%	0.002522870
75% Q3	0.001611419
50% Median	-0.000631103
25% Q1	-0.000829438
10%	-0.004349959
5%	-0.008531678

Number of Informative Clusters = 94

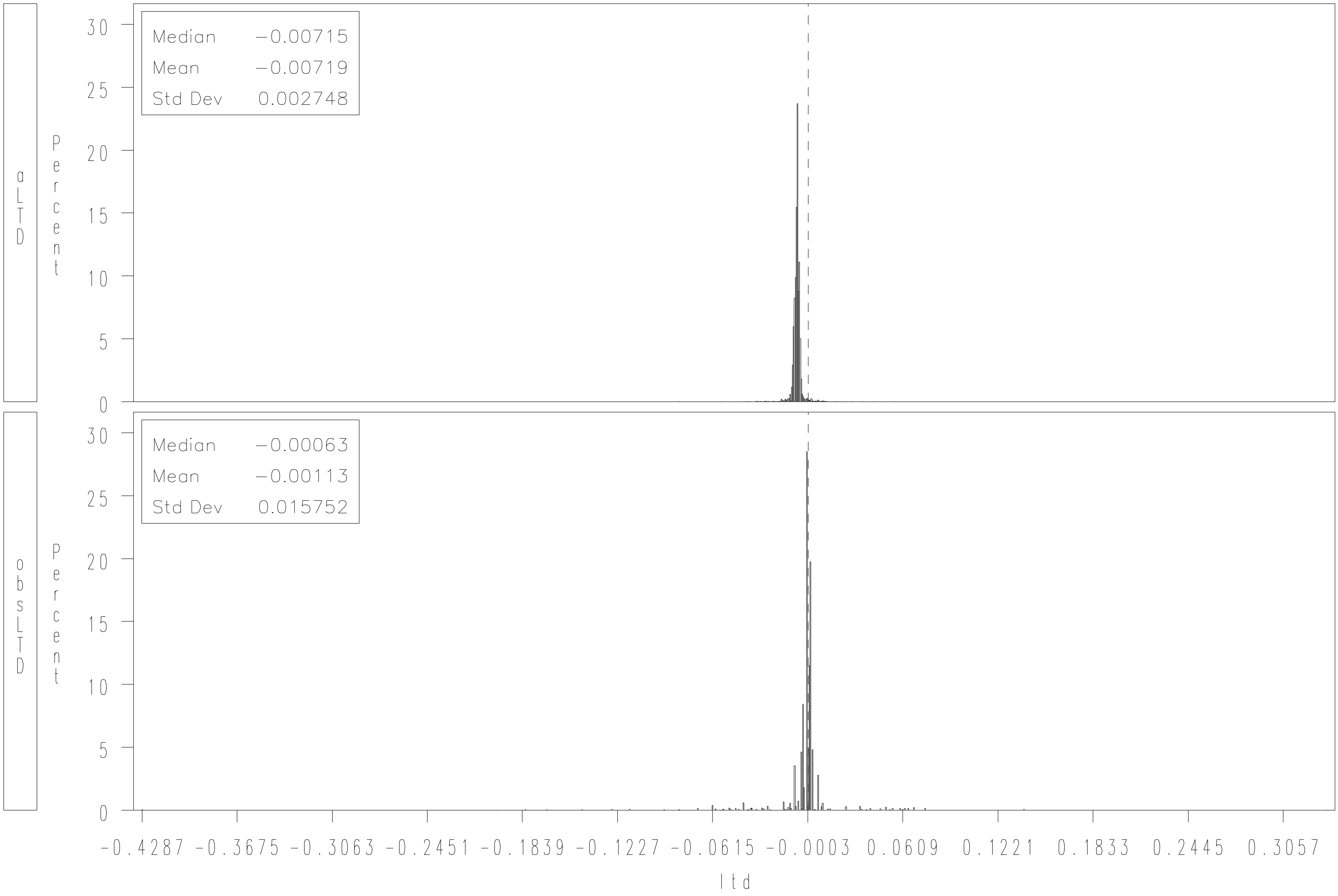
The UNIVARIATE Procedure
Variable: ltd
Freq: freq
oa = obsLTD

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	-0.055124892
0% Min	-0.428571429

Extreme Observations					
Lowest			**Highest**		
Value	**Freq**	**Obs**	**Value**	**Freq**	**Obs**
-0.428571	1	57369	0.333333	1	56405
-0.428571	1	57368	0.333333	1	56406
-0.428571	1	57367	0.333333	1	56407
-0.428571	1	57366	0.333333	1	56408
-0.428571	1	57365	0.333333	1	56409

Number of Informative Clusters = 94

Comparison of LTD Distributions



Number of Informative Clusters = 94