Identifying Meaningful Patient Subgroups via Clustering - Sensitivity Graphics

Robert L. Obenchain, Outcomes Research, US Medical
Eli Lilly and Company, Indianapolis, IN 46285

Abstract

When examining a dataset for evidence of differential patient response to treatment, comparisons among treated or untreated patients who are well matched on their observed X-covariates are more relevant than those between arbitrary (e.g. random) subgroups of patients. Local Treatment Differences (LTDs) within subgroups of most similar patients are differences in outcome most clearly attributable to treatment, such as [average y outcome when treated] minus [average y outcome when untreated.] We argue that LTDs within subgroups become more meaningful when they are numerically different from LTDs within subgoups immediately adjacent in X-space. Furthermore, the overall main effect of treatment is taken to be the mean of the LTD distribution across subgroups. When forming subgroups and quantifying their implied LTD distribution, two potentially conflicting objectives are [1] patients within any meaningful subgroup need to be more similar to each other than to patients from other subgroups and [2] the LTDs corresponding to any sub-subgroups within a meaningful subgroup are similar (i.e. form a tight, unimodal distribution.) We end up proposing that meaningful patient subgroups should not contain any meaningful sub-subgroups. We then conclude by demonstrating that inferences about LTD distributions are better expressed using confidence and tolerance intervals than by the traditional p-value for the statistical significance of the LTD main effect.

Keywords: Local Treatment Differences (LTDs), LTD Distributions, LTD Main Effects, Artificial LTD Distributions, Random Patient Clusterings, Meaningful Subgroups, Covariate Imbalance, Sensitivity Analyses.

1. Introduction

Local Control (LC), Obenchain(2004, 2005, 2006), is a new approach to estimation of treatment effects in randomized or non-randomized studies that is based upon a treatment-within-cluster nested ANOVA model that is frequently less restrictive and more robust than traditional Covariate Adjustment (CA) models. Because LC systematically forms subgroups, compares subgroups and (following overshooting) recombines subgroups that are not "meaningful" (as defined here), LC searches proceed via built-in sensitivity analyses. Specifically, the analyst may view graphical displays that identify the more effective treatment within each patient subgroup as well as averages across subgroups, where subgroups are generated by varying the number of clusters, the clustering metric and/or the clustering (unsupervised learning) algorithm.

The LC approach uses generalized definitions of treatment effect distributions and their main effects. In LC, treatment effects are distributions composed of (heteroskedastic) Local Treatment Difference (LTD) estimates, and the overall main effect of treatment is the unknown true mean of this distribution. It then becomes natural to ask whether [a] an observed LTD distribution could be a mixture of two or more well defined sub-distributions and whether [b] it is possible to predict the numerical size or sign of any LTDs directly from the baseline patient X-characteristics used to define clusters (subgroups.)

1.1 Nested Treatment-within-Subgroup ANOVA

Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	C = Number of Clusters	Local Average Treatment Effects (LATEs) are Cluster Means
Treatment within Cluster	Number of "Informative" Clusters ≤ C	Local Treatment Differences (LTDs)
Error	≥ Number of Patients – 2C	Uncertainty

Although a nested model can, technically, be a "wrong" model, these models are sufficiently versatile to almost always be "useful" models as the number of clusters increases. As the number of clusters is forced to increase, some clusters will become "uninformative" because they contain only treated patients or only untreated patients. As in the propensity scoring (PS) approaches of Rosenbaum and Rubin (1983) and Rosenbaum (2002), power can be lost in the LC process of focusing attention only upon more-and-more relevant patient comparisons (smaller and smaller subgroups.) After all, some treated and/or untreated patients simply cannot be well "matched" in PS-space …let alone in X-space.

The LC approach actually ignores all information within the first (Clusters) row of the above nested ANOVA table. The approach of McClellan, McNeil and Newhouse (1994) and the LATE estimates of Imbens and Angrist (1994) can use this information by making the very strong assumption that all X-variables are Instrumental Variables (IVs) that determine treatment selection but not outcome (except through treatment.) While these pure IV approaches do avoid (essentially) doubling the variance of point estimates implied by not forming treatment differences, they in turn ignore all information within the (always relevant) second row (Treatment within Cluster) of the nested ANOVA table. This is indeed a very high price to pay.

Rather than making strong, potentially unrealistic assumptions via generalized linear CA models that are too simplistic, the LC approach can provide robust yet powerful insights into all sorts of head-to-head treatment comparisons …processing information from sources ranging from massive administrative claims databases to highly restrictive, well-controlled clinical trials.

1.2 Basic Notation and Unbiased Estimation

y = observed outcome variable(s)
x = observed baseline covariate(s)
t = observed, binary treatment assignment
z = unobserved explanatory variable(s)

Now define a hypothetical binary indicator variable, δ_i, that is independent of outcome and equals 1 when treatment is selected for the i^{th} patient but equals 0 otherwise. This is the usual "counterfactual" situation where only the treated outcome for the i^{th} patient, y_{1i}, or else only the untreated outcome, y_{0i}, is actually observed. The observed outcome for the i^{th} patient is then of the form $y_i, = \delta_i y_{1i} + (1-\delta_i)y_{0i}$. Finally, assume that the corresponding propensity score for the i^{th} patient is strictly positive,

$$\Pr(\delta_i = 1) = E(\delta_i) = p_i > 0 .$$

Consider now the "shrinkage" estimator defined by the product of δ_i times y_i divided by p_i, as considered in Bang and Robbins(2005). There are then 2 terms in the expectation of this statistic, and the second term for $\delta_i = 0$ is always zero because $p_i > 0$.

$$E\left(\frac{\delta_i y_i}{p_i}\right) = E(\text{observed } y_{1i}) .$$

It then follows that this estimator (patient outcome weighted inversely to propensity score) is unbiased.

Now, note that a nested ANOVA model weights the outcome of each patient in exactly this same way within each cluster. The natural propensity score estimate within the j^{th} cluster is simply the observed proportion of patents selected for treatment within that cluster,

$$\hat{p}_j = n_{1j} / \left(n_{0j} + n_{1j}\right) \propto n_{1j} \text{ and } (1-\hat{p}_j) \propto n_{0j} ,$$

where n_{1j} is the number of treated patients within the j^{th} cluster and n_{0j} is the number of untreated patients within the j^{th} cluster The corresponding nested ANOVA estimate of the LTD within the j^{th} cluster is then

$$\textit{Estimated } LTD_j = \frac{\sum(\text{outcome for a treated patient})}{n_{1j}} - \frac{\sum(\text{outcome for an untreated patient})}{n_{0j}}$$

Thus the outcome for each patient is being weighted inversely proportional to the estimate of within-cluster probability, p_i or $(1-p_i)$, of receiving the treatment actually selected. In other words, if a CA model includes, say, only terms for main effects of treatment (rather than a full nested ANOVA structure), that CA model almost surely provides biased estimates at least in the sense that outcomes from different patients have not been appropriately weighted (i.e. inversely proportional to their local prevalence.)

2. Artificial LTD Distributions

Artifical LTD distributions are formed by ignoring the observed X-characteristics of all patients and then clustering them randomly. My algorithms for this, Obenchain(2005, 2006), generate several sets of independent bivariate normal pseudo-random Xs for each patient and then use K-Means clustering within each such set (simulation replicate.) This assures that the resulting simulated artificial LTD distribution includes the less and least relevant comparisons of outcomes between treated and untreated patients as well as the more and most relevant comparisons.

If some patient X-characteristics are highly predictive of outcome, the artificial LTD distribution can be over-dispersed even for a randomized study in which good balance happened to be achieved. When the data are from a nonrandomized study subject to treatment selection bias (imbalance), the artificial LTD

distribution can also be badly biased as well as over-dispersed.

One of our proposed "discovery" tactics will be to compare observed LTD distributions with their artificial counterparts. In fact, we will argue that rather obvious differences between these two types of distributions are commonplace.

3. Patient Subgroups

Subgroups of patients can be considered meaningful only if given patient characteristics (X-variables) are much more similar within subgroups than between subgroups.

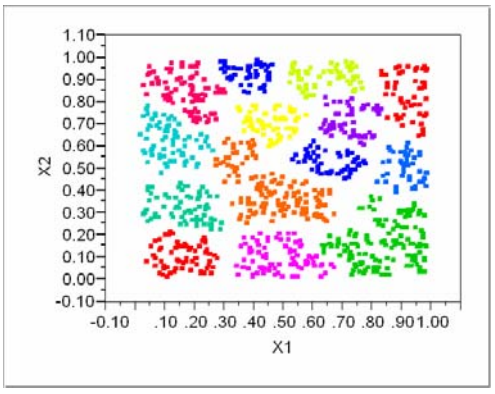
Furthermore, when making head-to-head treatment comparisons, subgroups of patients remain meaningful only if their LTD distributions of local (within subgroup) differences in outcome attributable to treatment, (Y outcome when treated) minus (Y outcome when untreated), also differ across subgroups.

Given enough data, meaningful subgroups should not be tiny. Meaningful subgroups usually contain sub-subgroups that, in turn, make these LTD sub-distributions estimable.

In other words, identification of LTD distributions involves "overshooting" by considering sub-subgroups that are too small. When these sub-subgroups fail to provide meaningfully different LTD sub-distributions, the analyst becomes confident that these sub-subgroups should be recombined.

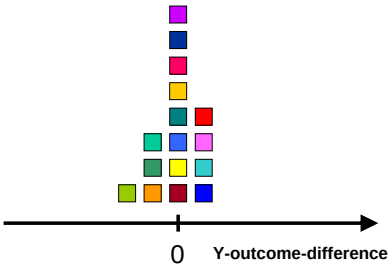
The key principle is simply that meaningful subgroups can only contain sub-subgroups that do not have meaningfully different LTD sub-distributions.

Consider the following illustration of 16 "informative" clusters (i.e. each cluster contains both treated and untreated patients) in a two dimensional X-space:



An unfortunate characteristic of the above color graphic is that three pairs of distinct but adjacent clusters are depicted with only very-slightly-different shades of green, turquoise or orange. Also, in this X-space, the 16 subgroups look to be rather "forced" rather than widely separated (or "natural".) On the other hand, it is still clearly true that patients are more similar within the 16 clusters than between clusters (subgroups.)

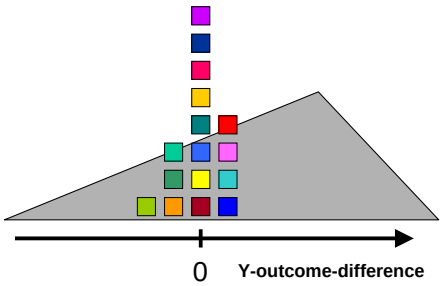
Let us now visualize the corresponding LTD distribution as a histogram of 16 estimated LTD main effects, each colored the same as the cluster that generated that estimate. In the hypothetical histogram illustrated below, our 16 "informative" clusters may not be "meaningful" (different.) After all, this LTD distribution has turned out to be quite peaked and unimodal. In fact, this LTD distribution could easily depict just a little bit of random noise about a central value that suggests that the overall main effect of treatment is zero!



Furthermore, the above 16 LTDs suggest no patient differential response to treatment (no interaction with patient X1 or X2 characteristics.)

On the other hand, if these 16 clusters have "informative" sub-clusters (at least one patient on each treatment), many more than 16 LTD main effects could be computed and displayed in a histogram or histograms. Would these distributions then look "different" and will they still be centered at zero?

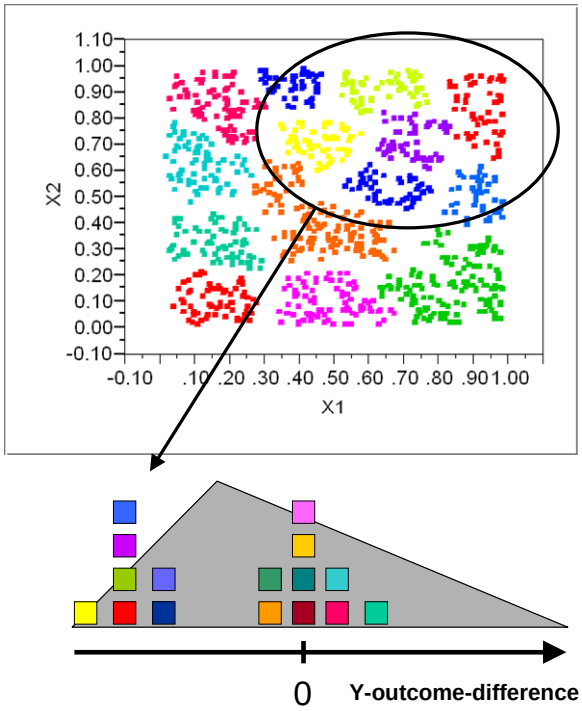
But wait. How does the artificial LTD distribution for 16 random clusters (represented by a grey pyramid below) compare with our 16 hypothetical LTDs?



Clear treatment selection bias has been revealed by the above differences in means and shapes between the two alternative distributions. In other words, the 16 cluster LTD distribution is "meaningful" because, relative to the artificial distribution for 16 random subgroups, bias has indeed been removed and precision has indeed been increased. But these 16 sub-groups are not meaningfully different from each other!

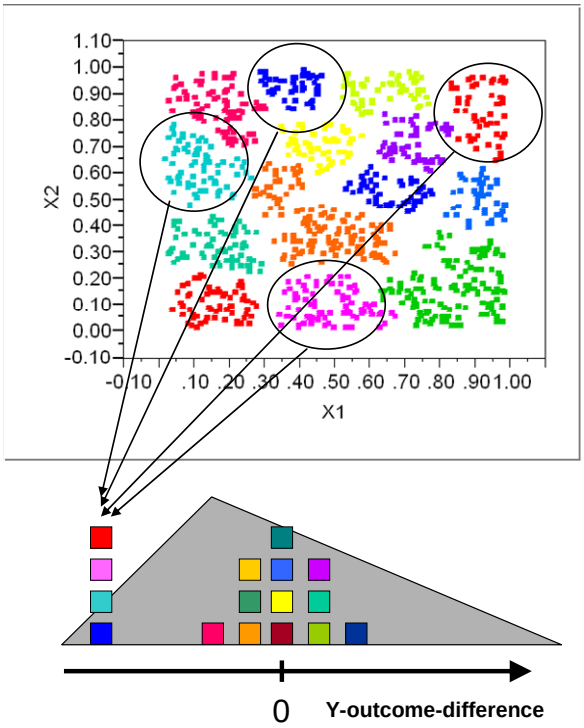
For data from Randomized Clinical Trials (RCTs), the observed LTD distribution is expected (asymptotically) to have the same overall mean effect as its artificial distribution. But that is definitely not the case in nonrandomized studies with imbalance because the LTD distribution displays only the more and most highly relevant comparisons between treated and untreated patients.

Here's a different situation. Suppose that the original 16 clusters generate the following pair of observed and artificial LTD distributions, where one "local mode" is attributable to "adjacent" sub-clusters in X-space. What might this mean?



The above LTD distribution suggests that a meaningful super-cluster can be formed by recombining adjacent subclusters.

On the other hand, clusters could fail to remain meaningful if the "local mode" in the observed LTD distribution corresponds to treatment effects on outcomes from widely dispersed clusters as illustrated next. Occam's razor then suggests that the patients from these widely separated clusters may have something in common (perhaps some unobserved patient characteristic) that is causing them all to have the same LTD estimate.



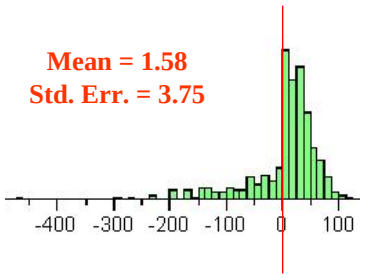
Again, if these 16 clusters have "informative" sub-clusters (at least one patient on each treatment), many more than 16 LTD main effects should be computed and displayed in a histogram to clarify plausible interpretations.

4. LTD Confidence and Tolerance Intervals

When an estimated LTD distribution is not clearly unimodal and/or symmetric, it usually reveals patient level differential responses to treatment. The overall mean (main effect) of such a LTD distribution is then not particularly interesting because it is not a very representative measure of central tendency of individual LTD main effect estimates.

The figure at the top of the next page illustrates a highly skewed LTD distribution with a mean value (overall LTD main effect) that is not significantly different from zero.

What strikes me as a clearly more interesting summary of this sort of estimated LTD distribution is the implied confidence one has that, say, at least 2 out of 3 patients (or 3 out of 4 patients) will have better outcomes when treated than when untreated.



Confidence intervals are random intervals designed to "cover", with stated probability, the unknown value of a single parameter (like the mean or a specified percentage point) of the distribution being sampled. Tolerance intervals are designed to "cover", with stated probability, ψ, at least a specified, positive minimum proportion, φ, of the distribution being sampled. However, a one-sided tolerance interval is also an unbounded confidence interval for a percentage point of the distribution being sampled.

A tolerance interval is said to be "distribution-free" if the population being sampled is continuous (rather than discrete) and the end-points of the interval coincide with a pair of the observed order statistics within the random sample. Via the "probability integral" transformation, any continuous distribution can be converted into the uniform distribution on [0, 1]. As a result, the coverage probabilities of intervals formed using any given pair of order statistics from all continuous distributions are identical to those for the uniform distribution, as tabulated in Natrella(1963).

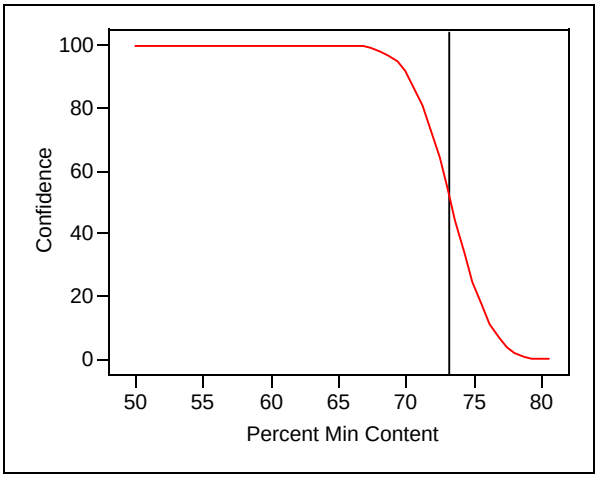
Today, the necessary Incomplete Beta calculations can be made using, say, the SAS® PROBBETA(φ, A, B) function, which is equivalent to Beta Distribution[φ, A, B] in JMP® and pbeta(φ, A, B) in R. For example, the probability, ψ, that at least a minimum proportion, φ, of an unknown, absolutely continuous distribution will lie between order statistics R and S within a random sample of size N (with $R < S \leq N+1$) is

$$\psi = 1.0 - \text{PROBBETA}(\varphi, A, B)$$
where $A = S-R$ and $B = N-S+R+1$.

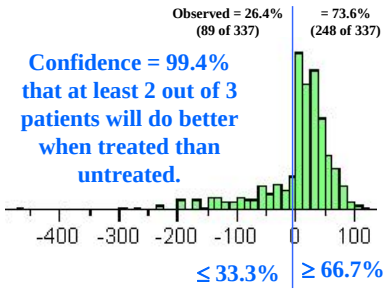
Now, consider the special case where an LTD distribution has been computed from patient outcomes within K informative clusters, and let λ_o denote the numerical value of the smallest of the K observed LTD order statistics that is strictly positive. Furthermore, let R = 1 + (the total number of treated or untreated patients within clusters yielding strictly non-positive LTD main effect estimates), let N = the total number of treated or untreated patients within K informative clusters, and let S = N+1 so that the upper limit of the tolerance interval will be $+\infty$. This is the case where A = N+1–R and B = R in the above formula relating confidence percentage, $100\times\psi$, to minimum content percentage, $100\times\varphi$, for an unbounded tolerance interval of the form $[\lambda_o, +\infty)$.

Alternatively, for simplicity, one could say that $(0, +\infty)$ is then an "approximate" tolerance interval with fixed endpoints and stated confidence in any stated minimum content, which is a conservative statement because $\lambda_o > 0$.

The figure below plots one's confidence in minimum content of 50% to 80% for $(0, +\infty)$ when N=337 and R=89, a situation where 73.6% of the numerical estimates in the observed LTD distribution were strictly positive. Note that confidence in this observed content (or even more content) is only 50%.



Also for simplicity, instead of presenting a plot, one could simply state the implied confidence at only 1 or 2 key values of minimum content. For example, in the example illustrated here, one has 99.4% confidence that at least 2 out of 3 patients will have better outcomes when treated than untreated.



One has only 24.8% confidence that at least 3 out of 4 patients will have better outcomes when treated than untreated because 75% content is larger than the

observed percentage of strictly positive LTD estimates of 73.6%.

The above sort of confidence statements strike me as being more meaningful summaries of this skewed LTD distribution than just stating its (non-significant) mean value. On the other hand, the rather long left-hand tail of negative LTD estimates (especially those less than, say, −100) definitely needs to be scrutinized and carefully described! What are the characteristics of this minority of patients who do so much better when untreated? Identifying this sort of differential response to treatment is essential to evidence based medicine.

5. Summary

My open source computing algorithms for R and JMP®, Obenchain(2005, 2006), implement many facets of the LC approach described here. They help automate the otherwise relatively tedious process of performing systematic sensitivity analyses to reveal the effects of alternative choices for X-covariates, numbers of clusters, etc. My JMP® script for LC also, optionally, inserts detailed patient level information about subgroup membership and LTD estimates back into the original dataset. The power of JMP® menus and interactive pointer/hand/brush/lasso functionality can then be unleashed to visualize how patient X-characteristics vary between clusters and whether they are predictive of either non-parametric LTD estimates or at least their numerical signs.

Acknowledgements

I wish to thank my Lilly colleagues Doug Faries, Gerhardt Pohl and Joe Johnson as well as participants at the SAMSI/NISS Summer 2006 Workshop on Subgroups in and Reproducibility of human studies for many helpful discussions of randomization, propensity scoring and local control (blocking) concepts.

References

Bang H, Robins JM. "Doubly Robust Estimation in Missing Data and Causal Inference Models." ***Biometrics*** 2005; 61: 962-972.

Fraley C, Raftery AE. "Model-based clustering, discriminant analysis, and density estimation." ***J Amer Stat Assoc*** 2002; 97: 611-631.

Imbens GW, Angrist JD. "Identification and Estimation of Local Average Treatment Effects." ***Econometrica*** 1994; 62: 467-475.

Johnson NL, Kotz S. ***Distributions in Statistics: Discrete Distributions.*** [Chapter 3: Binomial Distribution.] New York: John Wiley and Sons. 1969.

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

Kereiakes DJ, Obenchain RL, Barber BL, et al. "Abciximab provides cost effective survival advantage in high volume interventional practice." ***Am Heart J*** 2000; 140: 603-610.

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.

Natrella MG. ***Experimental Statistics.*** National Bureau of Standards Handbook 91. [Table A-23, pages T-41,44.] U.S. Government Printing Office. 1963. (1966 reprint with corrections.)

Obenchain RL. "Unsupervised Propensity Scoring: NN & IV Plots." ***2004 Proceedings of the American Statistical Association***, Health Policy Statistics Section [CD-ROM]. Alexandria, VA: American Statistical Association. 1899-1906.

Obenchain RL. ***USPS:*** An R package for Unsupervised and Supervised Propensity Score (and Instrumental Variable) adjustment for bias. http://www.r-project.org 2005.

Obenchain RL. JMP Scripts for "Local Control" and "Artificial LTD Distribution" calculations. http://www.math.iupui.edu/~indyasa 2006.

Rosenbaum PR. ***Observational Studies, 2nd Edition***. New York: Springer-Verlag. 2002.

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