Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Rogus et al., J Biom Biostat 2012, S7
DOI: 10.4172/2155-6180.S7-022

Research Article Open Access

Efficient Monte Carlo Resampling of Continuous Data with a Dichotomous Treatment Variable

John J Rogus[1,2]*, Shu-Fang Lin[1] and Eduardo K Lacson Jr[1]

[1]*Fresenius Medical Care North America, Waltham, MA, USA*
[2]*Renal Research Institute, New York, NY, USA*

Abstract

Introduction: Methods such as generalized least squares regression and linear mixed models have traditionally been used for analyzing repeated measurement data. However, the computational burden for these procedures can be prohibitively high for large data sets. We propose an efficient, non-parametric method for the analysis of a continuous outcome variable with intrapatient correlation and a dichotomous predictor variable.

Methods: The patient-level values of the dichotomous variable of interest are randomized to generate sets of equally likely permutations of the data under the null hypothesis. For each replication, the test statistic for the dichotomous variable is calculated and the collection of all such test statistics forms an empirical reference distribution used to assign a p-value to the actual test statistic from the original data. Efficient calculation of the reference distribution is possible by operating on the level of sufficient statistics for the outcome variable, as the dichotomous nature of the predictor variable then allows for rapid recalculation of the tests statistic at each replicate. An example based on 629,452 measurements of systolic blood pressure in 39,313 dialysis patients is used for illustration.

Results: The Monte Carlo p-value for a decrease in systolic blood pressure following a decrease in dialysate sodium was 0.04. Other computationally feasible, but inefficient, approaches such as data aggregation and year-over-year comparisons were unable to find a significant association.

Discussion: Monte Carlo simulation offers a valid approach to analyze a continuous outcome variable with intrapatient correlation and a dichotomous predictor of interest. This method can accommodate other predictors through a two-step procedure involving an initial regression analysis. Future work is needed to characterize the power of this approach relative to other methods and to study whether weighting strategies may be helpful in the situation where not all patients contribute the same number of data points to the analysis.

Keywords: Monte carlo simulation; Longitudinal analysis

Introduction

A fundamental assumption of linear regression is that the error terms are uncorrelated. When repeated measurements are made over time on a fixed cohort of patients, this assumption is likely to be violated. Ordinary least squares regression on this type of data will typically result in p-values that are artificially (and possibly dramatically) low, since the repeat measurements are not providing independent information about the true underlying relationship between the predictor and outcome variables. Various methods can be used to analyze quantitative outcomes measured repeatedly over time on a set of study subjects. Generalized least squares overcomes the issue of correlation by modeling the general form of the covariance structure [1]. Linear mixed models offer another approach by treating the repeat measurements as random effects [2]. While these methods are appealing, they are computationally intensive and, therefore, may not be practical for large data sets. Faced with this dilemma, researchers may resort to inefficient solutions such as aggregating data into a single summary measure for each subject, possibly by time period (e.g., mean of all values prior to treatment vs. mean of all values following treatment).

In light of these challenges, we set out to develop a computationally friendly approach that would allow us to keep the underlying repeated data structure intact. We describe an approach for assigning a valid p-value to the parameter estimate obtained by ordinary least squares regression, when the predictor variable of interest is dichotomous (e.g., 0/1, yes/no, etc.). Unlike typical inference in a regression framework, our p-values are empirical, estimated from a reference distribution formed by repeatedly reevaluating the test statistic for the dichotomous variable after randomly reassigning the values of the dichotomous variable across the study subjects. With large data sets, this process can also be computationally intensive due to the calculations that take place for linear regression in its most general form. However, for the case of a single dichotomous predictor variable, the parameter estimate and its standard error can be expressed as a simple function of sums and sums of squares of the outcome variable. We illustrate how to exploit this fact to efficiently estimate the empirical reference distribution.

Additional predictor variables of any type (dichotomous, categorical, quantitative, etc.) are easily accommodated by our approach by a simple initial regression procedure that we describe. It is also possible to extend our method to any test statistic based only on sums and sums of squares of the outcome variable. Through an example taken from a large study of dialysis patients, we demonstrate this by calculating an empirical p-value for an interaction term that reflects the differential change (post minus pre intervention) among a set of cases relative to a set of controls.

***Corresponding author:** John J Rogus, Clinical Science, Epidemiology, and Research Fresenius Medical Care, North America, Tel: (781) 699-2975; Fax: (781) 699-9383; E-mail: john.rogus@fmc-na.com

Received March 31, 2012; **Accepted** September 25, 2012; **Published** September 30, 2012

Citation: Rogus JJ, Lin SF, Lacson Jr EK(2012) Efficient Monte Carlo Resampling of Continuous Data with a Dichotomous Treatment Variable. J Biom Biostat S7:022. doi:10.4172/2155-6180.S7-022

Copyright: © 2012 Rogus JJ, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Materials and Methods

Under the null hypothesis, the values of a predictor variable (e.g., the labels of "case" or "control") provide no information about the outcome variable [3]. Thus, non-parametric analysis of a dichotomous predictor variable and a continuous outcome variable may be made by comparing a true test statistic for the parameter of interest to a reference distribution formed by a resampling procedure that randomizes the data labels among subjects. A "permutation test" considers all rearrangements of the labels followed by recalculation of the test statistic to construct the reference distribution. For large data sets, complete enumeration is impractical, but Monte Carlo sampling of the rearrangements offers an asymptotically equivalent alternative. Formally, it is the property of "exchangeability" (i.e., that each possible permutation is equally likely under the null hypothesis) that allows the assignment of a p-value based on the proportion of reference distribution replicates exceeding the actual test statistic.

The test statistic in our case is the Z-statistic corresponding to a dichotomous variable from a regression analysis ignoring intrapatient correlation. This value is easily calculated even for large data sets using standard procedures such as bdLm in Splus. However, the prospect of carrying out the regression procedure repeatedly on the simulated data sets may be daunting. For example, at 5 seconds per replicate, 10,000 replicates would take almost 14 hours to run and 100,000 replicates would take 10 times that amount. However, by taking advantage of the relatively simply form of the regression equations for the case of a single dichotomous predictor variable, a much quicker resampling algorithm can be applied. Specifically, the regression coefficient for a dichotomous predictor variable is equal to $\Sigma Y_{1i} /n_1 - \Sigma Y_{0j} /n_0$, where *i* and *j* index all n_1 observed outcomes in case subjects and all n_0 observed outcomes in control subjects. The variance terms is $(\Sigma Y_{1i}^2 + \Sigma Y_{0j}^2 - (\Sigma Y_{1i})^2 /n_1 - (\Sigma Y_{0j})^2/n_0)/n_0\ n_1$. From the perspective of the resampling procedure, the sufficient statistics for these two expressions are sums and sums of squares of the outcomes for each subject. Thus, an efficient resampling algorithm involves the following steps:

1. Calculate the sum of outcomes and the sum of squared outcomes for each subject.
2. Use the expressions above to calculate the observed test statistic as the regression coefficient divided by the square root of its variance.
3. Apply Monte Carlo simulation by randomizing the labels of the dichotomous variable (at the subject level) and recalculating a reference distribution of the test statistics under the null hypothesis.
4. Assign a p-value to the observed test statistic equal to the proportion of times a more extreme test statistic occurs in the reference distribution.

This same procedure can be carried out for any other parameter estimates that involve only sums and sums of squares of outcomes. In our example, we are interested in comparing whether values in a case group decreases from baseline more than that in a control group. Thus, the key parameter is the group by time period (baseline vs. follow up) interaction term. Through matrix algebra, it can be shown that the parameter estimate for the interaction term is $(\Sigma Y_{0Bi} / n_{0B} - \Sigma Y_{0Fj} / n_{0F}) - (\Sigma Y_{1Bk} / n_{1B} - \Sigma Y_{1Fl} / n_{1F})$ with variance $(1/n_{0B} + 1/n_{0F} + 1/n_{1B} + 1/n_{1F})(Q - R_{0B} - R_{0F} - R_{1B} - R_{1F})/(N - 4)$, where $N = n_{0B} + n_{0F} + n_{1B} + n_{1F}$, Q is the sum of all N squared outcomes, and $R_{0B} = (\Sigma Y_{0Bi})^2/n_{0B}$ and R_{0F}, R_{1B}, and R_{1F} are defined analogously.

Adjustment for additional explanatory variables is easily accommodated by performing a single initial regression of the outcome variable on all covariates except the dichotomous variable of primary interest. The residuals from this regression serve as the outcome variable in the 4-step regression procedure described above. Provided the initial regression is done only on subjects with non-missing values at the dichotomous variable of interest, the results (i.e., parameter estimate and test statistic for the dichotomous variable of interest) from this two stage procedure will be exactly the same as those from a single multiple regression model.

To illustrate the Monte Carlo procedure, we apply it to a data set assembled to investigate the effect of a treatment change in patients undergoing hemodialysis. For patients with end-stage renal disease, dialysis is necessary to compensate for the loss of kidney function. Hemodialysis accomplishes this via a chemical solution called dialysate that interacts with the blood through a semi-permeable filter in a specialized machine called a dialyzer. One component of the dialysate solution is sodium, a cation known to increase blood pressure. The sample data set comprises 765 facilities of Fresenius Medical Care, North America. In 581 of these facilities, sodium decreased by 3 mEq/L during January 2009—June 2009 ("transition period") while the rest remained stable. For our purposes, we will focus on 629,452 pre-dialysis measurements of systolic blood pressure in 39,313 subjects (28,211 from changer ("case") facilities and 11,102 from non-changer ("control") facilities). Given the switch to lower dialysate sodium, we test whether case facilities experience a larger decrease in systolic blood pressure (values during 6 month baseline vs. values during 24 month follow up) than control facilities.

Results

Monthly mean values for systolic blood pressure are shown for case facilities and control facilities (Figure 1). One striking feature of this data is the clear pattern of seasonality, with highest mean systolic blood pressure during December/January and lowest during July. One approach to deal with this phenomenon and avoid the complication of intra-subject correlation is to look at a series of year over year comparisons (e.g., Jul 2008 vs. Jul 2009, etc). Using the data in this inefficient manner, we begin to see a consistently larger drop of systolic blood pressure in the case group relative to the control group, presumably due to the decrease in dialysate sodium, but the results

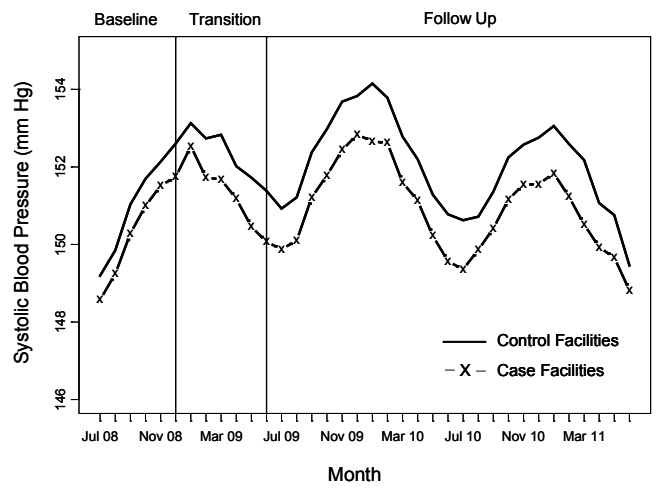

Figure 1: Monthly mean systolic blood pressure values for case facilities and control facilities.

are not statistically significant (Table 1). Similarly, no significant association was found if we aggregated data for each patient into a single mean baseline and single mean follow up value.

If we ignore the issue of intrasubject correlation and naively include all monthly values into a single regression analysis (with pre-adjustment for seasonality), we obtain an overall parameter estimate of -0.45, which is quite reasonable given the year-over-year results in Table 1. The corresponding Z-statistic of -3.59 would ordinarily result in a highly significant p-value, but the underlying assumption of independent errors is most likely violated due to intrasubject correlation. Thus, the Z-statistic must be judged according to a modified reference distribution, in our case, from Monte Carlo simulation.

The Monte Carlo reference distribution is nothing more than an empirical distribution of test statistics under the null hypothesis, conditional on the observed pattern of outcome data. If error terms are indeed independent, then the Monte Carlo reference distribution is equivalent to a random sample from a standard normal distribution. To show this, we constructed a reference distribution for comparing cases and controls in a single month (i.e., to completely avoid the intrasubject correlation issue) based on 10,000 replicates. We found 5% and 2.5% lower tail thresholds to be -1.647 and -1.963, very similar to the familiar -1.645 and -1.960 cutoffs derived from the standard normal distribution.

The Monte Carlo reference distribution for the case group effect when all monthly values are considered has much greater variation than a standard normal distribution (Figure 2). Now, test statistics of -1.645 and -1.960 correspond to empirical p-values of 0.20 and 0.16, respectively (as opposed to 0.05 and 0.025), illustrating that traditional thresholds are not at all applicable to this correlated data. From this reference distribution, we find that the global parameter estimate of -0.45 with Z-statistic of -3.59 has an empirical p-value of 0.04. Thus, statistical significance, elusive in the year-over-year comparisons as well as with data aggregation, was achieved using the Monte Carlo procedure.

Discussion

Large data sets with extensive longitudinal follow up can be a double edged sword for statisticians. On one hand, "longer" data sets (i.e., more subjects) enable the detection of smaller effects and "wider" data sets (i.e., repeated measurements) further enhance this ability and also allow for a fuller appreciation of the relationship over time. On the other hand, such data sets can also complicate the computational

Year-over-year	Case Group Effect*	p-value
Jul 08 - Jul 09	-0.43	0.28
Aug 08 - Aug 09	-0.50	0.21
Sep 08 - Sep 09	-0.41	0.32
Oct 08 - Oct 09	-0.50	0.23
Nov 08 - Nov 09	-0.63	0.14
Dec 08 - Dec 09	-0.13	0.76
Jul 08 - Jul 10	-0.65	0.14
Aug 08 - Aug 10	-0.25	0.58
Sep 08 - Sep 10	-0.19	0.67
Oct 08 - Oct 10	-0.38	0.41
Nov 08 - Nov 10	-0.41	0.38
Dec 08 - Dec 10	-0.34	0.47

* A negative case group effect indicates that the case group decreased by a greater amount than the control group for the 1-year period under consideration.

Table 1: Effect of being a case relative to being a control for year-over-year comparisons.

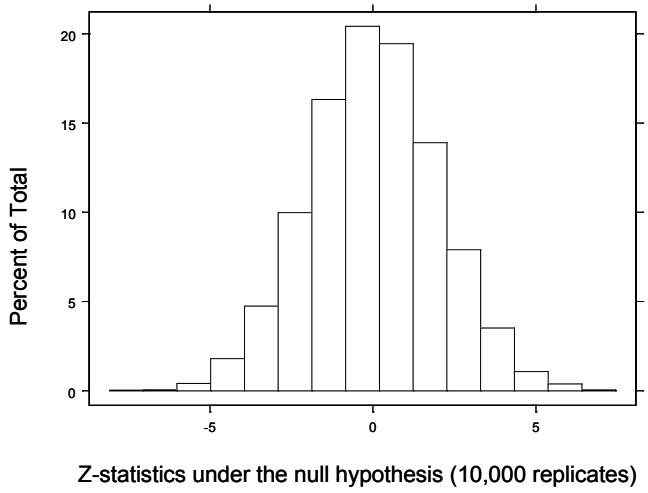

Figure 2: Monte Carlo reference distribution of Z-statistics for systolic blood pressure example.

aspects of the analysis. For example, while the bigdata library in Splus provides functionality for ordinary linear regression, it does not include the ability to run generalized least squares regression or linear mixed models. Other platforms (e.g., SAS) may overcome this technical challenge in some instances, but trial runs with the data set described above were unsuccessful.

Rather than embarking on an exhaustive quest to solve these computational limitations through different software, a bigger/faster computer, or better optimization, we decide to explore an alternative approach, one that assigns a valid p-value not through modeling a covariance matrix or treating repeated measurements as random effects, but rather through the generation of a null reference distribution using a common non-parametric technique. Specifically, we developed an algorithm that repeatedly randomizes the dichotomous predictor variable of interest (on a subject level) so that a set of representative values under the null hypothesis can be assembled. Importantly, we show how to accomplish this type of Monte Carlo simulation efficiently by working with sufficient statistics for each subject that can quickly manipulated for each replicate of the simulation. Without this simplification, the time needed to derive an empirical p-value would be prohibitively large.

We focused only on the case of a dichotomous predictor variable of interest, a critical assumption for the simplification described. However, our algorithm can be generalized to include any statistic that is a function of only sums and sums of squares of the outcome variable. Overall tests of categorical predictors as well as specific contrasts of such variables fall into this category. Although not ideal, a continuous predictor variable could be transformed to fit this paradigm by categorizing it into a discrete set of values (e.g., transforming age into <40, 40-50, 50-60, and >60).

We also considered only the case of a continuous outcome variable. For a dichotomous outcome variable, exact logistic regression is theoretically possibly [4], but our simplified randomization algorithm based on sums and sums of squares cannot be easily adapted. We are currently developing an alternative procedure based on the Mantel-Haenszel test statistic when predictor are categorical variables (or can be grouped as such). In this setting, the labels that define each of the substrata can be randomized and if working on the scale of log-

transformed odds ratio, all estimates and their variances are simple functions of counts within substrata [5].

In the case of a continuous outcome, adjusting for other variables, whether continuous, dichotomous, or categorical, is easily accommodated by a simple initial regression procedure including all of these variables. Owing to the simple closed-form equations in linear regression, using the residuals from this model in a second regression with the dichotomous variable of interest results in inference that is identical to the one which has all variables been added in a single regression model. Thus, the sufficient statistics are now sums and sums of squares of the residuals for the original regression and these need only be calculated one time.

Our application of a non-parametric technique was motivated by our large data sets that could not easily be handled by standard programs. Often, non-parametric methods are chosen for the opposite reason, because data sets are too small to rely on asymptotic approximations. Indeed, our approach is fully applicable to small data sets, although the simplification of working with sufficient statistics would probably be unnecessary. This approach could be useful if certain model assumptions were in doubt or if outliers are exerting a large amount of leverage on the regression results.

The Monte Carlo procedure that we have described will give a valid p-value, due to the property of exchangeability with regard to the permutation of dichotomous labels. Nevertheless, future work is required to determine the power of this non-parametric approach relative to other methods such as generalized least squares regression, linear mixed models, or other non-parametric regression methods for correlated errors [6]. For example, the Monte Carlo procedure described may be criticized because individuals with incomplete data will contribute fewer observations to the overall test statistic. The linear mixed model, by contrast, models a single mean value for each subject and uses the repeat measurements to model the degree of variation within a subject. Further work is needed to understand how strongly this aspect of the Monte Carlo procedure impacts power in the presence of unbalanced data. We speculate, however, that a weighting scheme that assigns weights that are inversely proportional to the number of observations per subject may help to increase power in these situations.

References

1. Kariya T, Kurata H (2004) Generalized Least Squares. John Wiley & Sons Ltd., San Francisco.
2. Verbeke G, Molenberghs G (2000) Linear Mixed Models for Longitudinal Data. Springer-Verlag, New York.
3. Good PI (2006) Resampling Methods. (3rdedn), Birkhauser, Boston.
4. Mehta, CR, Patel NR (1995) Exact logistic regression: theory and examples. Stat Med 14: 2143–2160.
5. Silcocks P (2005) An easy approach to the Robins-Breslow-Greenland variance estimator. Epidemiol Perspect Innov 2: 9.
6. Opsomer J, Wang Y, Yang Y (2001) Nonparametric regression with correlated errors. Statist Sci 16: 134-153.

This article was originally published in a special issue, **Medical statistics: Clinical and experimental research** handled by Editor(s). Dr. Herbert Pang, Duke University, USA.