DOCUMENT RESUME

ED 369 815 TM 021 454

AUTHOR Halperin, Si; Jorgensen, Randall
TITLE The Use of Control in Non-Randomized Designs.
PUB DATE Apr 94
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Control Groups; *Psychological Studies; Research Design; *Research Methodology; Scores; Statistical Analysis; Statistical Bias; Statistical Studies
IDENTIFIERS Biostatistics; *Nonrandomized Design; Residuals (Statistics); Simulated Data Sets; *Subclassification on Propensity Score

ABSTRACT

The concept of control is fundamental to comparative research. In research designs where randomization of observational units is not possible, control has been exercised statistically from a single covariate by a process of residualization. The alternative, known as subclassification on the propensity score, was developed primarily for biostatistical applications. The illustration included (comparing the physiological variables of 34 subjects from a home with a family history of hypertension with 46 subjects with no history of hypertension) demonstrates applicability to psychological research as well. Subclassification on the propensity score has several advantages over residualizing as a means of control. First, it allows control on many covariates. Second, it allows one to assess how well bias in the covariates has been reduced. Third, interactions with covariates can be followed by simple effects more readily than with residualization. Finally, subclassification on the propensity score requires fewer assumptions than residualization. Four tables. (Contains 19 references.) (Author/SLD)

The Use of Control in Non-randomized Designs

Si Halperin
Randall Jorgensen

Syracuse University
Department of Psychology

April 7, 1994

Correspondence: Si Halperin
Department of Psychology
430 Huntington Hall
Syracuse University
Syracuse, NY 13244-2340
315 443-2705
HALPERIN@MAILBOX.SYR.EDU

Key words and phrases. propensity score, non-randomized designs

Paper presented at the annual meeting of the American Educational Research Association - Special Interest Group / Educational Statisticians, April 1994, New Orleans, LA.

Abstract

The concept of control is fundamental to comparative research. In research designs where randomization of observational units is not possible, control has been exercised statistically from a single covariate by a process of residualization. In this paper, we consider an alternative to control by statistical adjustment. The alternative, known as subclassification on the propensity score, was developed primarily for biostatistical applications. The illustration included in the paper will demonstrate applicability to psychological research as well. Subclassification on the propensity score has several advantages over residualizing as a means of control. First, it allows control on many covariates. Second, it allows one to assess how well bias in the covariates has been reduced. Third, interactions with covariates can be followed by simple effects more readily than with residualization. Finally, subclassification on the propensity score requires fewer assumptions than residualization.

Introduction

The concept of control is fundamental to comparative research. In his chapter on the introduction to basic concepts in experimental design, Kirk (1982) discussed the control of nuisance variables. "Nuisance variables are undesired sources of variation in an experiment that may affect the dependent variable" (p. 6). Cochran (1983) recommended control as a means to reduce bias and increase precision in non-randomized designs. It is the primary objective of this paper to review how control may be applied to non-randomized designs. A research design will be included to provide an extended illustration.

"Four approaches can be used to control nuisance variables. One approach is to hold the nuisance variable constant for all subjects" (Kirk, 1982, p. 6). A second approach is to randomize the assignment of subjects to treatment. A third approach is to subclassify on important covariates and include the subclasses in the research model. A fourth approach is to perform a statistical adjustment such as is done in analysis of covariance (e.g., Benjamin, 1967; Huitema, 1980).

The first approach to control tends to limit the generalizability of research results. Approach two requires the ability to randomize treatment assignment. There are two broad classes of research design where such randomization is not possible. Observational studies (Cochran, 1983), or quasi-experimental designs (Campbell and Stanley, 1963), represent a class of designs where interest centers on the effect of intervention in the absence of randomization. According to Holland (1986), the units in these designs are "potentially exposable" to the treatments. The study of causal inference requires that units be potentially exposable.

In some designs, a subject's characteristic rather than an intervention defines his or her group membership. In these designs, units are not potentially exposable and causal inference cannot be examined. Even in designs with no prospect for causal inference, the use of control may be desired (cf., Benjamin, 1967; Wilder, 1968).

Incorporating Many Covariates

When many sources of potential bias threaten the interpretation of a non-randomized design, both subclassification on covariates and statistical adjustment become difficult to implement. Until statistical procedures capable of incorporating many covariates were recently developed, investigators controlled on a limited number of the most important covariates.

Rosenbaum and Rubin (1984) demonstrated how a large number of covariates can be controlled in the comparison of two non-randomized treatments. They recommended use of the "propensity score" to reduce potential confounding due to uncontrolled covariates. "The propensity score is the conditional probability that a unit with vector x of *observed* covariates will be assigned to treatment 1" (Rosenbaum and Rubin, 1984, p. 516). They show that "subclassification on the population propensity score will balance x, in the sense that within subclasses that are homogeneous in the propensity score, the distribution of x is the same for treated and control units."

In practice, propensity scores must be estimated, and subclassification on the estimated propensity score usually balances the covariates only approximately. Propensity scores may be estimated by posterior probabilities from any classification procedure. In practice, logistic regression provides an appropriate means to estimate propensity scores because it can effectively utilize both quantitative and qualitative covariates (Press & Wilson, 1978). Sub-classification using more than 4 or 5 subclasses usually gains very little (Cochran, 1983). However, there are instances where estimated propensity scores can be used to form matched pairs, thus turning a "between blocks" variable into a "within block" variable. Rosenbaum and Rubin (1985) provide an example of forming matched pairs from estimated propensity scores. Another interesting application is found in Rosenbaum (1986).

There are several advantages of subclassifying on the estimated propensity score rather than performing a statistical adjustment using some of the more important covariates. First, the design and model are simpler. If four subclasses are used in a two group design, the design becomes a 4 x 2 factorial

model. Cell means within subsclasses are readily compared and reduction in error variance is readily assessed. Second, fewer assumptions are required for subclassification than for statistical adjustment. Third, the effectiveness of subclassification to balance all the covariates is readily assessed by examining a 4 x 2 analysis of variance for each covariate. By doing so, inadequacy of estimated propensity scores can be found and often corrected.

Research Design

The study used to illustrate the application of the propensity score was designed to compare physiological variables of 34 subjects coming from a home with a family history of hypertension with 46 subjects who had no history of hypertension. Since the distinction between the two groups is drawn on the basis of a demographic characteristic of the subjects and is not potentially exposable, this is not an observational study, nor a quasi-experimental design. Although causal inference is inappropriate to this study, it is clearly a non-randomized design. Making comparisons between those with and without a parental history of hypertension will be improved by balancing on a number of baseline covariates.

A number of cardiovascular measures were taken during a 30 minute resting baseline period. The variables included minute by minute estimates of diastolic and systolic blood pressure, mean arterial pressure, and heart rate. These measures, plus a body mass index, provided a pool of variables from which the covariates were to be selected.

On a separate day one week later, physiological measures were observed during three treatment periods. Subjects were observed during a pre-task anticipation period, during a period where subjects participated in a digits backward task, and during a period of mental arithmetic. During each period, measures of systolic and diastolic blood pressure were collected to be used as the response variables.

Selection of Covariates

The pool of potential covariates supplied by the 30 minute resting period was well in excess of 100 variables. With only 80 subjects, clearly not all variables available from the resting period could be utilized in the development of the propensity score. In addition to the large number of variables in the covariate pool, missing data due to equipment failure created an additional complication.

To cope with the issues of missing data and large numbers of variables, we decided to consolidate the variables in the pool by calculating selected statistics for the resting measures of diastolic and systolic blood pressure, mean arterial pressure, and heart rate. Two different sets of statistics were examined. For each of the four cardiovascular measures, Covariate Set 1 consisted of the following quantiles: the median, the interquartile range, the 5th percentile, and the 95th percentile. These 16 variables, plus the index of body mass, provided the 17 covariates from which the propensity score for Covariate Set 1 would be estimated and assessed

Covariate Set 2 was inspired by growth curve modelling. It included coefficients derived from regressing each set of 30 resting measures on a quadratic function of the serial order of the measure. The correlations among the coefficients were reduced by using a hierarchy of models: (1) an intercept only model; (2) a model linear in serial order; and (3) a model quadratic in serial order. The highest level coefficient from each of the hierarchy was selected as a covariate. A similar method is used in the creation of orthogonal polynomial coefficients. These 12 coefficents, plus the mean of minutes 21-25 for each resting measure, produced the set of 16 variables in Covariate Set 2. The mean of minutes 21-25 for each resting measure of cardiovascular activity was selected due to analyses showing that a nadir resting value occurs at this point (Jorgensen, Schreer, & Gelling, 1990).

Rosenbaum and Rubin (1984) selected covariates to include in their logistic regression estimate of the propensity score by use of "an inexpensive stepwise discriminant analysis." We had fewer covariates, and they were consistent with a classification of variables suggested by Mosteller and Tukey (1977) for variable selection in multiple regression. "Key carriers" are covariates "that we want to include in any regression" (p. 393); "promising carriers ... deserve somewhat special attention." "The haystack (is) a

motley collection of other carriers that deserve limited attention." Following the advice of Mosteller and Tukey (1977), while making use of the relationship between discriminant analysis and multiple regression (Flury and Riedwyl, 1985), we proceeded with variable selection as follows.

For each Covariate Set, we defined key and promising carriers. To select from among the promising carriers, all-subset regression (Hocking, 1976) was used with binary coded family history as the response variable, and the key carriers forced into the model. A second regression analysis forced the key and selected promising carriers into the model, and used stepwise procedures to examine the haystack variables. The final logistic regression model consisted of the main effects selected and, following the procedures of Rosenbaum and Rubin (1984), their two-way interactions. Care should be exercised not to include interactions without the corresponding main effects (Peixoto, 1990).

For the quantiles in Covariate Set 1, median resting systolic blood pressure was selected to be the single key carrier. Median resting diastolic blood pressure, median arterial pressure, median heart rate, and body mass served as the promising carriers. Using the all-subset based on Mallows' C_p from PROC REG (SAS, 1990), two promising carriers were added to the regression model: median resting distolic blood pressure and median arterial pressure. Stepwise regression on the haystack of remaining variables failed to add to the set of three carriers.

For Covariate Set 2, mean resting systolic and diastolic blood pressure were used as key carriers. The promising carriers were means of resting systolic blood pressure, diastolic blood pressure, and arterial pressure for minutes 21-25. Using all-subset regression again, only mean resting arterial pressure for minutes 21-25 was added to the model. Stepwise regression on the haystack of remaining variables again failed to add to the set of carriers selected previously.

To choose between the two covariate sets, canonical correlation analyses were performed, where the left-hand variables were the carriers in a covariate set (three main effect variables and their two-way interactions) and the right-hand variables were the six response variables. For each analysis, two statistically significant canonical correlation were found (for Set 1: .76, .70; for Set 2: .75, .70). Redundancy analysis results were similar for the two sets also, so, for this illustration, canonical correlation analysis was of little help in choosing between the covariate sets. Because existing evidence (e.g., Jorgensen et al., 1990) indicates that at least 20 minutes are required to achieve an accurate resting blood pressure level, we decided to proceed with Covariate Set 2.

Subclassification on the Propensity Score

Using PROC LOGISTIC (SAS, 1990), the parameters of a logistic regression model were estimated from the carriers in Covariate Set 2. Propensity scores were estimated by evaluating the logistic regression function for each subject. A schematic plot of estimated propensity scores for the two family history groups is displayed in Figure 1. Some separation between the family history groups may be observed for the propensity scores, suggesting the need to improve the balance between the two groups on the covariates.

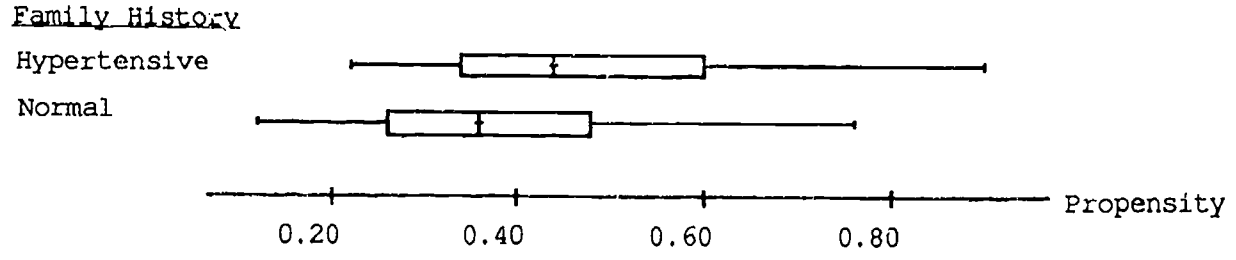

Figure 1. Schematic plot of propensity scores from Covariate Set 2.

Four subclasses of equal size were formed from the estimated propensity scores. To assess the effectiveness of subclassification to balance the groups on the 16 carriers of Covariate Set 2, two analyses were performed for each carrier. First, a two sample t-test was performed on each carrier, using family history as a two-level blocking variable. Second, a 4 x 2 analysis of variance was performed on each carrier, using the two levels of family history and the four subclass levels as the blocking variables. Table 1 illustrates the effect of subclassification for one covariate - systoloic blood pressure. Table 2 summarizes the p-values for all the covariates. To achieve balance after subclassification, each covariate should display a non-significant main effect due to family history, and a non-significant interaction between family history and subclassification.

Table 1.
Evidence of Balance Between Family History Groups for Mean Sytolic Blood Pressure

prior to subclassification	Family History	
	Hypertensive	Normal
weighted means	119.5	116.1

after subclassification	Family History	
Subclass	Hypertensive	Normal
0	114.2	111.5
1	114.7	114.3
2	119.5	119.8
3	123.9	123.8
unweighted means	118.0	117.4

Table 2.
Evidence of Balance Between Family History Groups for all Covariates

Covariate	p-value x 100		
	before subclassification	for main effect after subclassification	for interaction after subclassification
mean systolic blood pressure	2	56	81
mean diastolic blood pressure	11	74	18
mean arterial pressure - minutes 21-25	31	77	64
mean arterial pressure	19	83	89
mean heart rate	45	58	79
linear coefficient - diastolic blood pressure	67	77	63
linear coefficient - systolic blood pressure	97	89	23
linear coefficient - arterial pressure	77	49	15
linear coefficient - heart rate	71	78	39
quadratic coefficient - diastolic blood pressure	56	91	88
quadratic coefficient - systolic blood pressure	42	38	32
quadratic coefficient - arterial pressure	68	66	71
quadratic coefficient - heart rate	31	85	45
mean diastolic blood pressure - minutes 21-25	14	73	8
mean systolic blood pressure - minutes 21-25	4	68	57
mean heart rate - minutes 21-25	37	52	92

Tables 1 and 2 provide strong evidence that subclassification on the propensity score has balanced the 16 covariates. Had lack of balance occurred, the opportunity exists for fine tuning the logistic regression model by including extra covariates. This step was unneccessary for the illustration.

Statistical Analysis

Having demonstrated balance from the subclassification, the primary statistical analysis may proceed. Each response variable (diastolic and systolic blood pressure) was measure during each of three treatment periods: pre-task anticipation, digits backward, and mental arithmetic. A split-plot analysis of variance was performed, using family history and subclassification as between-subjects variables and treatment period as a within-subjects variable. The split-plot analysis of variance source table is presented in Table 3. For comparative purposes, the split-plot analysis without the subclassification variable is also included in Table 3.

Table 3.
Effect of Subclassification on the Analysis of Variance.

3a. Diastolic	Subclassification Included			Subclassification Excluded		
Source	MS	F	p-value	MS	F	p-value
Between						
History	0.0	0.00	1.00	168.2	1.16	0.29
Subclass	448.1	3.40	0.02			
HxS	82.5	0.63	0.60			
Error	131.8			145.2		
Within						
Treatment	2616.2	113.92	0.00	2721.4	118.30	0.00
TxH	29.0	1.26	0.29	22.3	0.97	0.38
TxS	17.7	0.77	0.60			
TxHxS	20.5	0.89	0.50			
Error	23.0			23.0		

3b. Systolic	Subclassification Included			Subclassification Excluded		
Source	MS	F	p-value	MS	F	p-value
Between						
History	20.6	0.12	0.72	239.9	1.04	0.31
Subclass	12140.5	7.50	0.00			
HxS	550.5	3.33	0.02			
Error	165.4			230.5		
Within						
Treatment	3335.7	88.85	0.00	3717.6	94.35	0.00
TxH	169.3	4.51	0.29	135.8	3.45	0.03
TxS	46.1	1.23	0.60			
TxHxS	49.7	1.32	0.50			
Error	37.5			39.4		

Two conclusions are evident from Tables 1,2, and 3. First, subclassification on the estimated propensity score can reduce bias attributable to covariates. Second, subclassification can improve the precision of an analysis. The second point is apparent when the error MS with subclassification included are compared to those with subclassification excluded. The error MS either stays the same or decreases. Sometimes the decrease is dramatic, as happened with the "Between-Subjects" error term for systolic blood pressure. Here, the MS error decreased by 28%, which represents an impressive gain in precision due to including additional variables in the model. Of course, since the covariates are "between-subjects" measures, they have little impact on the "within-subjects" error mean square.

Alternative Analyses

The analysis above, incorporating subclasses into the model, can be viewed as an alternative to more traditional methods based on statistical adjustment of the response variable. Such procedures include the analysis of covariance and approximations to the analysis of covariance based upon the use of regression residuals (Cochran, 1957).

The analysis of residualized gains reported in Table 4 may be compared to those of Table 3. In the analysis of residualized gains, the blood pressure measurements from each of the three treatment periods are regressed on a comparable blood pressure measurement from minutes 21-25 of the resting baseline period. The identity of metric from baseline covariate to treatment response is the basis for the label "residualized gains."

Residualization does not require commensurable covariate and response measurements. An interesting variation on the traditional method is the residualization of response measurements from the covariate of propensity score. Residualizing based on the propensity score has the advantage that all measurements from the Covariate Set are being controlled, not just a covariate that represents the same metric as the response variable. An analysis based upon residuals of blood pressure on propensity score is included in Table 4 for comparative purposes.

Table 4.
Analysis of Variance Based Upon Residuals.

4a. Diastolic	Residualized Gains			Residuals from Propensity Scores		
Source	MS	F	p-value	MS	F	p-value
Between						
History	15.7	0.14	0.71	18.8	0.16	0.69
Error	109.0			117.8		
Within						
Treatment	2732.0	119.94	0.00	2688.2	118.05	0.00
TxH	15.3	0.67	0.51	50.0	2.20	0.11
Error	22.8			22.8		

4b. Systolic	Residualized Gains			Residuals from Propensity Scores		
Source	MS	F	p-value	MS	F	p-value
Between						
History	0.2	0.00	0.98	79.6	0.47	0.50
Error	176.7			169.4		
Within						
Treatment	3710.1	94.24	0.00	3654.1	95.64	0.00
TxH	144.0	3.66	0.03	226.3	5.92	0.00
Error	39.4			38.2		

As measures of precision, error mean squares can be compared for the two types of analyses presented in Table 4. This comparison yields no clear advantage in precision for either analysis. The benefits of the regression adjustments used in Table 4 can be seen by comparing those analyses with the unadjusted analysis found in the right-hand side of Table 3. Sizable reductions in the "between-subjects" error mean squares are attributable to the regression adjustments, especially for systolic blood pressure.

The regression adjustments used in Table 4 are intended not only to improve precision, but also to reduce bias. Family history means are "balanced" on the covariate measure by the adjustment in the residualized gains analysis, . One could speculate that family history means are being balanced for all the measurements in the Covariate Set by adjusting on the propensity score. The authors know of no way to corroborate this speculation, however. The ability to confirm that family history means are at least

approximately balanced by subclassification on the propensity score (Tables 1 and 2) represents an important advantage of that analysis over analyses that perform a statistical adjustment.

Another advantage of subclassifying on the propensity score can be observed in Tables 3 and 4. In Table 3, we assess not only the main effect due to subclass, but also the interaction of subclass with history and treatment. For systolic blood pressure, a History x Subclass interaction was found. Analysis of simple main effects (Kirk, 1982) is easily performed and may provide additional insights into the manner in which differences in family history depend upon covariate differences.

No main effect for the covariate, nor interactions between the covariate and other effects, are present in Table 4. When analysis proceeds by statistical adjustment, it is assumed that no interactions with the covariate exist. Because analysis of simple main effects is complex in this case, it is seldom performed. Johnson and Neyman (1936) derived an analysis of simple main effects when a covariate adjustment is used.

Conclusions

Subclassification on estimated propensity scores provides a simple and effective mean to reduce confounding in research design where treatments cannot be randomized. By balancing all known sources of potential bias, the case for causal inference can be strengthened. However, as seen in the illustration, subclassifying on estimated propensity scores has real value in designs where causal inference is not at issue. In designs where a demographic characteristic (e.g., coronary artery disease or family history of hypertension) defines the groups, there is great value in making comparisons between two samples within subclassifications that are known to be well balanced with regard to all most of the available covariates.

The use of subclassification on estimated propensity scores shares two of the advantages of analysis of residualized gains.

- The precision of the analysis, as reflected by the error mean squares, is improved.
- The bias, as measured by the difference in the means of a covariate, is eliminated.

Two additional advantages of subclassifying on estimated propensity scores are not readily available for analysis of residualized gains.

- The ability to investigate how effectively all the covariates have been balanced.
- The ability to investigate interactions with the covariates and easily proceed with an analysis of simple main effects.

References

Benjamin, L.S. (1967). Facts and artifacts in using analysis of covariance to "undo" the law of initial values. Psychophysiology, 4, 187-206.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

Cochran, W.G. (1983). Planning & analysis of observational studies. New York: Wiley.

Cochran, W.G. (1957). Analysis of covariance: Its nature and uses. Biometrics, 13, 261-281.

Flury, B., & Riedwyl, H. (1985). T^2 tests, the quadratic two-group discriminant function, and their computation by quadratic regression. The American Statistician, 39, 20-25.

Hocking, R.R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

Holland, P.W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945-960.

Huitema, B.E. (1980). The analysis of covariance and alternatives. New York: Wiley.

Johnson, P.O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. Statistical Research Memoirs, 1, 57-93.

Jorgensen, R.S., Schreer, G.E., & Gelling, P.D. (1990). Pretask "Baseline" cardiovascular activity: Pretask anticipation as a possible contributor to pretask cardiovascular activity. Psychophysiology, 27 (Suppl.), S43 (abstract).

Kirk, R.E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Belmont, CA: Brooks / Cole.

Mosteller, F., & Tukey, J.W. (1977). Data analysis and regression. Reading, MA: Addison-Wesley.

Peixoto, J.L. (1990). A property of well-formulated polynomial regression models. The American Statistician, 44, 26-30.

Press, S.J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, 73, 699-705.

Rosenbaum, P.R. (1986). Dropping out of high school in the United States: An observational study. Journal of Educational Statistics, 11, 207-224.

Rosenbaum, P.R., & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, 79, 516-524.

Rosenbaum, P.R., & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling that incorporate the propensity score. The American Statistician, 39, 33-38.

SAS Institute (1990). SAS/STAT user's guide, version 6, fourth edition (Volume 2). Cary, NC: SAS Institute.

Wilder, J.F. (1968). Stimulus and response: The law of initial values. Baltimore: Williams & Wilkins.