Along with those three equations is the additional equation

4. $\pi_{61} + \pi_{62} + \pi_{63} = 1$.

An arithmetic manipulation of these four equations results in numerical solutions for the three unknowns: $\pi_{61} = \pi_{62} = \pi_{63} = 1/3 = .33$. These stationary probabilities indicate that Harry would be at any of the three locations with equal likelihood after many steps in the random walk.2

Conclusion

If there is a sequence of random events such that a future event is dependent only on the present event and not on past events, then the sequence is likely a Markov chain, and the work of Markov and others may be used to extract useful information from an analysis of the sequence. The topic of the Markov chain has become one of the most captivating, generative, and useful topics in probability and statistics.

William M. Bart and Thomas Bart

See also Matrix Algebra; Probability, Laws of

Further Readings

Bremaud, P. (2008). *Markov chains*. New York: Springer.

Feller, W. (1968). *An introduction to probability theory and its applications: Volume 1* (3rd ed.). New York: Wiley.

Hoel, P., Port, S., & Stone, C. (1972). *Introduction to stochastic processes*. Boston: Houghton Mifflin.

Kemeny, J., & Snell, J. (1960). *Finite Markov chains*. Princeton, NJ: D. Van Nostrand.

MATCHING

The term *matching* refers to the procedure of finding for a sample unit other units in the sample that are closest in terms of observable characteristics. The units selected are usually referred to as *matches*, and after repeating this procedure for all units (or a subgroup of them), the resulting subsample of units is called the *matched sample*. This idea is typically implemented across subgroups of a given sample, that is, for each unit in one subgroup, matches are found among units of another subgroup. A matching procedure requires defining a notion of distance, selecting the number of matches to be found, and deciding whether units will be used multiple times as a potential match. In applications, matching is commonly used as a preliminary step in the construction of a matched sample, that is, a sample of observations that are similar in terms of observed characteristics, and then some statistical procedure is computed with this subsample. Typically, the term *matching estimator* refers to the case when the statistical procedure of interest is a point estimator, such as the sample mean. The idea of matching is usually employed in the context of observational studies, in which it is assumed that selection into treatment, if present, is based on observable characteristics. More generally, under appropriate assumptions, matching may be used as a way of reducing variability in estimation, combining databases from different sources, dealing with missing data, and designing sampling strategies, among other possibilities. Finally, in the econometrics literature, the term *matching* is sometimes used more broadly to refer to a class of estimators that exploit the idea of selection on observables in the context of program evaluation. This entry focuses on the implementation of and statistical inference procedures for matching.

Description and Implementation

A natural way of describing matching formally is in the context of the classical potential outcomes model. To describe this model, suppose that a random sample of size n is available from a large population, which is represented by the collection of random variables (Y_i, T_i, X_i), $i = 1, 2, \ldots, n$, where $T_i \in \{0, 1\}$,

$$Y_i = \begin{cases} Y_{0i} & \text{if} \quad T_i = 0 \\ Y_{1i} & \text{if} \quad T_i = 1 \end{cases}$$

and X_i represents a (possibly high-dimensional) vector of observed characteristics. This model aims to capture the idea that while the set of characteristics X_i is observed for all units, only one of the two random variables (Y_{0i}, Y_{1i}) is observed for each unit, depending on the value of T_i. The underlying random variables Y_{0i} and Y_{1i} are usually referred to as potential outcomes because they represent the two potential states for each unit.

For example, this model is routinely used in the program evaluation literature, where T_i represents treatment status and Y_{0i} and Y_{1i} represent outcomes without and with treatment, respectively. In most applications the goal is to establish statistical inference for some characteristic of the distribution of the potential outcomes such as the mean or quantiles. However, using the available sample directly to establish inference may lead to important biases in the estimation whenever units have selected into one of the two possible groups ($T_i = 0$ or $T_i = 1$). As a consequence, researchers often assume that the selection process, if present, is based on observable characteristics. This idea is formalized by the so-called *conditional independence assumption:* conditionally on X_i, the random variables (Y_{0i}, Y_{1i}) are independent of T_i. In other words, under this assumption, units having the same observable characteristics X_i are assigned to each of the two groups ($T_i = 0$ or $T_i = 1$) independently of their potential gains, captured by (Y_{0i}, Y_{1i}). Thus, this assumption imposes random treatment assignment conditional on X_i. This model also assumes some form of *overlap* or *common support:* For some $c > 0$, $c \leq \mathbb{P}(T_i = 1|X_i) \leq 1 - c$. In words, this additional assumption ensures that there will be observations in both groups having a common value of observed characteristics if the sample size is large enough. The function $p(X_i) = \mathbb{P}(T_i = 1|X_i)$ is known as the *propensity score* and plays an important role in the literature. Finally, it is important to note that for many applications of interest, the model described above employs stronger assumptions than needed. For simplicity, however, the following discussion does not address these distinctions.

This setup naturally motivates matching: observations sharing common (or very similar) values of the observable characteristics X_i are assumed to be free of any selection biases, rendering the statistical inference that uses these observations valid. Of course, matching is not the only way of conducting correct inference in this model. Several parametric, semiparametric, and nonparametric techniques are available, depending on the object of interest and the assumptions imposed. Nonetheless, matching is an attractive procedure because it does not require employing smoothing techniques and appears to be less sensitive to some choices of user-defined tuning parameters.

To describe a matching procedure in detail, consider the special case of matching that uses the Euclidean distance to obtain $M \geq 1$ matches with replacement for the two groups of observations defined by $T_i = 0$ and $T_i = 1$, using as a reservoir of potential matches for each unit i the group opposite to the group this unit belongs to. Then, for unit i the mth match, $m = 1, 2, \ldots, M$ is given by the observation having index $j_m(i)$ such that

$$T_{j_m(i)} \neq T_i \text{ and}$$
$$\sum_{j=1}^{n} \mathbb{1}\{T_j \neq T_i\}\mathbb{1}\{\|X_j - X_i\| \leq \|X_{j_m(i)} - X_i\|\} = m.$$

(The function $\mathbb{1}\{\cdot\}$ is the indicator function and $\|\cdot\|$ represents the Euclidean norm.) In words, for the ith unit, the mth match corresponds to the mth nearest neighbor among those observations belonging to the opposite group of unit i, as measured by the Euclidean distance between their observable characteristics. For example, if $m = 1$, then $j_1(i)$ corresponds to the unit's index in the opposite group of unit i with the property that $\|X_{j_1(i)}, -X_i\| \leq \|X_j - X_i\|$ for all j such that $T_j \neq T_i$, that is, $X_{j_m(i)}$ is the observation closest to X_i among all the observations in the appropriate group. Similarly, $X_{j_1(i)}, X_{j_2(i)}, \ldots, X_{j_M(i)}$ are the second closest, third closest, and so forth, observations to X_i, among those observations in the appropriate subsample. Notice that to simplify the discussion, this definition assumes existence and uniqueness of an observation with index $j_m(i)$. (It is possible to modify the matching procedure to account for these problems.)

In general, the always observed random vector X_i may include both discrete and continuous random variables. When the distribution of (a subvector of) X_i is discrete, the matching procedure may be done exactly in large samples, leading to so-called *exact matching*. However, for those components of X_i that are continuously distributed, matching cannot be done exactly, and therefore in any given sample there will be a discrepancy in terms of observable characteristics, sometimes called the *matching discrepancy*. This discrepancy generates a bias that may affect inference even asymptotically.

The M matches for unit i are given by the observations with indexes $J_M(i) = \{j_1(i), \ldots, j_M(i)\}$, that is, $\left(Y_{j_1(i)}, X_{j_1(i)}\right), \ldots, \left(Y_{j_M(i)}, X_{j_M(i)}\right)$. This procedure is repeated for the appropriate subsample of units to obtain the final matched sample. Once the matched sample is available, the statistical procedure of interest may be computed. To this end, the first step is to "recover" those counterfactual variables not observed for each unit, which in the context of matching is done by imputation. For example, first define

$$\hat{Y}_{0i} = \begin{cases} Y_i & \text{if} \quad T_i = 0 \\ \dfrac{1}{M} \displaystyle\sum_{j \in J_M(i)} Y_j & \text{if} \quad T_i = 1 \end{cases} \text{ and}$$

$$\hat{Y}_{1i} = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in J_M(i)} Y_j & \text{if} \quad T_i = 0 \\ Y_i & \text{if} \quad T_i = 1 \end{cases}$$

that is, for each unit the unobserved counterfactual variable is imputed using the average of its M matches. Then simple matching estimators are easy to construct: A matching estimator for $\mu_1 = \mathbb{E}[Y_{1i}]$, the mean of Y_{1i} is given by $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_{1i}$, while a matching estimator for $\tau = \mu_1 - \mu_0 = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$, the difference in means between both groups, is given by $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$, where $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_{0i}$. The latter estimand is called the *average treatment effect* in the literature of program evaluation and has received special attention in the theoretical literature of matching estimation.

Matching may also be carried out using estimated rather than observed random variables. A classical example is the so-called *propensity score matching,* which constructs a matched sample using the estimated propensity score (rather than the observed X_i) to measure the proximity between observations. Furthermore, matching may also be used to estimate other population parameters of interest, such as quantiles or dispersion measures, in a conceptually similar way. Intuitively, in all cases a matching estimator imputes values for otherwise unobserved random variables using the matched sample. This imputation procedure coincides with an M nearest neighbor ($M - NN$) nonparametric regression estimator.

The implementation of matching is based on several user-defined options (metric, number of matches, etc.), and therefore numerous variants of this procedure may be considered. In all cases, a fast and reliable algorithm is needed to construct a matched sample. Among the available implementations, the so-called *genetic matching,* which uses evolutionary genetic algorithms to construct the matched sample, appears to work well with moderate sample sizes. This implementation allows for a generalized notion of distance (a reweighted Euclidean norm that includes the Mahalanobis metric as a particular case) and an arbitrary number of matches with and without replacement.

There exist several generalizations of the basic matching procedure described above, a particularly important one being the so-called *optimal full matching*. This procedure generalizes the idea of pair or M matching by constructing multiple submatched samples that may include more than one observation from each group. This procedure encompasses the simple matching procedures previously discussed and enjoys certain demonstrable optimality properties.

Statistical Inference

In recent years, there have been important theoretical developments in statistics and econometrics concerning matching estimators for average treatment effects under the conditional independence assumption. These results establish the validity and lack of validity of commonly used statistical inference procedures involving simple matching estimators.

Despite the fact that in some cases, and under somewhat restrictive assumptions, exact (finite sample) statistical inference results for matching estimators exist, the most important theoretical developments currently available have been derived for large samples and under mild, standard assumptions. Naturally, these asymptotic results have the advantage of being invariant to particular distributional assumptions and the disadvantage of being valid only for large enough samples.

First, despite the relative complexity of matching estimators, it has been established that these estimators for averages with and without replacement enjoy root-*n* consistency and asymptotic normality under reasonable assumptions. In other words, the estimators described in the previous section (as well as other variants of them) achieve the parametric rate of convergence having a Gaussian limiting distribution after appropriate centering and rescaling. It is important to note that the necessary conditions for this result to hold include the restriction that at most one dimension of the observed characteristics is continuously distributed, regardless of how many discrete covariates are included in the vector of observed characteristics used by the matching procedure. Intuitively, this restriction arises as a consequence of the bias introduced by the matching discrepancy for continuously distributed observed characteristics, which turns out not to vanish even asymptotically when more than one continuous covariate are included. This problem may be fixed at the expense of introducing further bias reduction techniques that involve nonparametric smoothing procedures, making the "bias corrected" matching estimator somehow less appealing.

Second, regarding the (asymptotic) precision of matching estimators for averages, it has been shown that these estimators do not achieve the minimum possible variance, that is, these estimators are inefficient when compared with other available procedures. However, this efficiency loss is relatively small and decreases fast with the number of matches to be found for each observation.

Finally, in terms of uncertainty estimates of matching estimators for averages, two important results are available. First, it has been shown that the classical *bootstrap* procedure would provide an inconsistent estimate of the standard errors of the matching estimators. For this reason, other resampling techniques must be used, such as *m out of n bootstrap* or *subsampling,* which do deliver consistent standard error estimates under mild regularity conditions. Second, as an alternative, it is possible to construct a consistent estimator of the standard errors that does not require explicit estimation of nonparametric parameters. This estimator uses the matched sample to construct a consistent estimator of the asymptotic (two-piece) variance of the matching estimator.

In sum, the main theoretical results available justify asymptotically the use of classical inference procedures based on the normal distribution, provided the standard errors are estimated appropriately. Computer programs implementing matching, which also compute matching estimators as well as other statistical procedures based on a matched sample, are available in commonly used statistical computing software such as MATLAB, R, and Stata.

Matias D. Cattaneo

See also Observational Research; Propensity Score Analysis; Selection

Further Readings

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74,* 235–267.

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76,* 1537–1557.

Abadie, A., & Imbens, G. W. (2009, February). *A martingale representation for matching estimators.* National Bureau of Economic Research, working paper 14756.

Diamond, A., & Sekhon, J. S. (2008, December). *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.* Retrieved February 12, 2010, from http://sekhon.berkeley.edu

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15,* 609–627.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47,* 5–86.

Rosenbaum, P. (2002). *Observational studies.* New York: Springer.

MATRIX ALGEBRA

James Joseph Sylvester developed the modern concept of matrices in the 19th century. For him a matrix was an array of numbers. He worked with systems of linear equations; matrices provided a convenient way of working with their *coefficients,* and matrix algebra was to generalize number operations to matrices. Nowadays, matrix algebra is used in all branches of mathematics and the sciences and constitutes the basis of most statistical procedures.

Matrices: Definition

A matrix is a set of numbers arranged in a table. For example, Toto, Marius, and Olivette are looking at their possessions, and they are counting how many balls, cars, coins, and novels they each possess. Toto has 2 balls, 5 cars, 10 coins, and 20 novels. Marius has 1, 2, 3, and 4, and Olivette has 6, 1, 3, and 10. These data can be displayed in a table in which each row represents a person and each column a possession: