Balance Tests for Matching Estimators*

DO NOT QUOTE OR DISTRIBUTE

Jasjeet S. Sekhon †

10/29/2004 (05:17)

*Matching and balance test software can be downloading from `http://jsekhon.fas.harvard.edu/matching`. All errors are my responsibility.

†Associate Professor, Department of Government, Harvard University. 34 Kirkland Street, Cambridge, MA, 02138. `jasjeet_sekhon@harvard.edu`, `HTTP://jsekhon.fas.harvard.edu/`, Voice: 617.496.2426, Fax: 617.496.5149.

1 A Consistent Kolmogorov-Smirnov Test when Nuisance Parameters Must be Estimated and the Data have Point Masses

The bootstrap is used account for the sampling distribution of nuisance parameters and another bootstrap is used to construct via simulation the correct test level when the data contains point masses. Both problems arise when the hypothesis being considered is if the distribution of the estimated probabilities from a logistic regression are equal for two populations—e.g., treated and control. The parameters in the logic regression are nuisance parameters. Point masses often arise in the estimated probabilities when the covariates used in the logistic regression are themselves not entirely continuous (such as income with a spike at $0) or categorical variables.

We first outline the algorithm for calculating Kolmogorov-Smirnov p-values which are consistent even when there are point masses. In the next section we discuss the complete algorithm which corrects for both point masses and nuisance parameters. In the penultimate section Monte Carlo evidence is provided for the accuracy of the algorithm.

1.1 Point Mass

Let Y be the $n \times 1$ data vector of interest. Let Y_1 be the first sample of Y and Y_2 the second each of which has, respectively, length n_1 and n_2. Let ks_s denote the usual KS test statistic, and let ks_p denote the probability of observing ks_s as determined by the usual algorithm such as that provided by `ks.test()` in *R* and *Splus*.

Then the point mass algorithm is:

Step 1 Calculate the KS statistic in the original (full) sample using Y_1 and Y_2; denote this $\hat{ks}_s^f$.

Step 2 Resample n observations from Y with replacement B times. Denote a given resample by Y^b. Divide Y^b into two samples equal in size to n_1 and n_2, denoted Y_1^b and Y_2^b. Compute ks_s using Y_1^b and Y_2^b, denote this statistic as $\hat{ks}_s^b$.

Step 3 Calculate ks_{mc} which is the Monte Carlo KS p-values as: $\hat{ks}_{mc} = \sum_{b=1}^{B} 1 \left\{ \hat{ks}_s^b >= \hat{ks}_s^f \right\} / B$

$\hat{ks}_{mc}$ is the statistic returned by the `ks.boot()` function in the "Matching" package. For a proof of the validity of this algorithm see (Abadie 2002).

1.2 Nuisance Parameters

The bootstrap is used to integrate over the distribution of the parameters when estimating nuisance parameters (Hall 1992). In this case, we want to know the p-value for the KS test applied to $\hat{Y}_1$ and $\hat{Y}_2$. The algorithm is:

Step 1 Calculate the KS statistic in the original (full) sample using $\hat{Y}_1$ and $\hat{Y}_2$; denote this $\hat{ks}_s^f$.

Step 2 Calculate the Monte Carlos KS test in the full sample; denote this as $\hat{ks}_{mc}^f$.

Step 3 Take B samples from the multivariate distribution of the nuisance parameters. If this is a logistic regression, this would be draws from the multivariate normal distribution of the coefficient vector where each draw is denoted by $\hat{\beta}^b$. For each bootstrap draw calculate $\hat{Y}^b$. In the case of logistic regression this would be $\hat{\mu} = X\hat{\beta}^b$. And $\hat{Y}^b = exp(\hat{\mu})/(1 + exp(\hat{\mu})$.

Step 4 For each $\hat{Y}^b$ calculate ks_{mc} denoted as $\hat{ks}_{mc}^b$.

Step 5 The bootstrap p-value is calculated as: $\hat{ks}_{bs} = \sum_{b=1}^{B} 1 \left\{\hat{ks}_{mc}^b >= \hat{ks}_{mc}^f\right\} /B$

$\hat{ks}_{bs}$ has taken into account both the sampling distribution of the nuisance parameters and calculated the empirical p-value in each Monte Carlos step so the underlying data may contain point masses.

1.3 Monte Carlo Evidence

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effect in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457): 284–292.

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.