Matching Methods for Causal Inference

Gary King
Institute for Quantitative Social Science
Harvard University

Talk at University of Georgia, 3/3/2011
Problem: Model dependence (review)
Solution: Matching to preprocess data (review)
Problem: Many matching methods & specifications
Solution: The Space Graph helps us choose
Problem: The most commonly used method can increase imbalance!
Solution: Other methods do not share this problem
(Coarsened Exact Matching is usually best)
Lots of insights revealed in the process
Model Dependence Example
Replication: Doyle and Sambanis, APSR 2000

- **Data:** 124 Post-World War II civil wars
- **Dependent variable:** peacebuilding success
- **Treatment variable:** multilateral UN peacekeeping intervention (0/1)
- **Control vars:** war type, severity, duration; development status; etc.
- **Counterfactual question:** UN intervention switched for each war
- **Data analysis:** Logit model
- **The question:** How *model dependent* are the results?
## Two Logit Models, Apparently Similar Results

<table>
<thead>
<tr>
<th>Variables</th>
<th>Original “Interactive” Model</th>
<th>Modified Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>SE</td>
</tr>
<tr>
<td>Wartype</td>
<td>−1.742</td>
<td>.609</td>
</tr>
<tr>
<td>Logdead</td>
<td>−.445</td>
<td>.126</td>
</tr>
<tr>
<td>Wardur</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>Factnum</td>
<td>−1.259</td>
<td>.703</td>
</tr>
<tr>
<td>Factnum2</td>
<td>.062</td>
<td>.065</td>
</tr>
<tr>
<td>Trnsfcap</td>
<td>.004</td>
<td>.002</td>
</tr>
<tr>
<td>Develop</td>
<td>.001</td>
<td>.000</td>
</tr>
<tr>
<td>Decade</td>
<td>−.299</td>
<td>.169</td>
</tr>
<tr>
<td>Treaty</td>
<td>2.124</td>
<td>.821</td>
</tr>
<tr>
<td>UNOP4</td>
<td>3.135</td>
<td>1.091</td>
</tr>
<tr>
<td>Wardur*UNOP4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Constant</td>
<td>8.609</td>
<td>2.157</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Log-likelihood</th>
<th>Pseudo $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>122</td>
<td>-45.649</td>
<td>.423</td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>-44.902</td>
<td>.433</td>
</tr>
</tbody>
</table>

Gary King (Harvard, IQSS)
What to do?

- **Preprocess I**: Eliminate extrapolation region
- **Preprocess II**: Match (prune bad matches) within interpolation region
- **Model remaining imbalance**
Matching within the Interpolation Region
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

Matching reduces model dependence, bias, and variance
What Matching Does

- **Notation:**
  - $Y_i$: Dependent variable
  - $T_i$: Treatment variable (0/1)
  - $X_i$: Pre-treatment covariates

- **Treatment Effect for treated ($T_i = 1$) observation $i$:**
  \[
  TE_i = Y_i(T_i = 1) - Y_i(T_i = 0)
  \]
  \[
  = \text{observed} - \text{unobserved}
  \]

- **Estimate $Y_i(0)$ with $Y_j$ from matched ($X_i \approx X_j$) controls**
  \[
  \hat{Y}_i(0) = Y_j(0) \text{ or a model } \hat{Y}_i(0) = \hat{g}_0(X_j)
  \]

- **Prune unmatched units to improve balance** (so $X$ is unimportant)

- **QoI: Sample Average Treatment effect on the Treated:**
  \[
  \text{SATT} = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} TE_i
  \]

- **or Feasible Average Treatment effect on the Treated:** $\text{FSATT}$
Method 1: Mahalanobis Distance Matching

1. **Preprocess (Matching)**
   - Distance($X_i, X_j$) = $\sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance $>$ *caliper*

2. **Estimation** Difference in means or a model
Method 2: Propensity Score Matching

1. **Preprocess (Matching)**
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance($X_i, X_j$) = $|\pi_i - \pi_j|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper

2. **Estimation** Difference in means or a model
Method 3: Coarsened Exact Matching

1 Preprocess (Matching)
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
   - Apply exact matching to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned

2 Estimation Difference in means or a model
   - Need to weight controls in each stratum to equal treateds
   - Can apply other matching methods within CEM strata (inherit CEM’s properties)
The Bias-Variance Trade Off in Matching

- **Bias** (& model dependence) = $f(\text{imbalance, importance, estimator})$
  \[ \leadsto \text{we measure imbalance instead} \]
- **Variance** = $f(\text{matched sample size, estimator})$
  \[ \leadsto \text{we measure matched sample size instead} \]
- **Bias-Variance trade off** \(\leadsto\) **Imbalance-\(n\) Trade Off

**Measuring Imbalance**
- Classic measure: Difference of means (for each variable)
- Better measure (difference of multivariate histograms):
  \[
  \mathcal{L}_1(f, g; H) = \frac{1}{2} \sum_{\ell_1 \cdots \ell_k \in H(X)} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}|
  \]
Comparing Matching Methods

- MDM & PSM: Choose matched $n$, match, check imbalance
- CEM: Choose imbalance, match, check matched $n$
- Best practice: iterate
- Choose matched solution & matching method becomes irrelevant
- Our idea: Compute lots of matching solutions, identify the frontier of lowest imbalance for each given $n$, and choose a matching solution
A Space Graph: Real Data
King, Nielsen, Coberley, Pope, and Wells (2011)

Healthways Data

- ○ Raw Data
- - - Random Pruning
- ▲ PSM
- × MDM
- + CEM

N of Matched Sample ("variance")
L1 ("bias")
A Space Graph: Real Data

Called/Not Called Data

L1 ("bias")

N of Matched Sample ("variance")
A Space Graph: Real Data

FDA Data

N of Matched Sample ("variance")
L1 ("bias")

Gary King (Harvard, IQSS)
Talk at University of Georgia, 3/3/2011
A Space Graph: Real Data

Lalonde Data Subset

N of Matched Sample ("variance")

L1 ("bias")

Gary King (Harvard, IQSS)
A Space Graph: Simulated Data — CEM

Gary King (Harvard, IQSS)
A Space Graph: Simulated Data — Propensity Score

PSM: 1 Covariate

PSM: 2 Covariates

PSM: 3 Covariates

Gary King (Harvard, IQSS)
Data where PSM Works Reasonably Well — PSM & MDM

Unmatched Data: $L_1 = 0.685$

PSM: $L_1 = 0.452$

MDM: $L_1 = 0.448$

Gary King (Harvard, IQSS)
Data where PSM Works Reasonably Well — CEM

Bad CEM: L1 = 0.661
100% of the treated units

Better CEM: L1 = 0.188
100% of the treated units

Even Better CEM: L1 = 0.095
72% of the treated units
CEM Weights and Nonparametric Propensity Score

CEM Weight: \[ w_i = \frac{m_i^T}{m_i^C} \] (Unnormalized)

CEM Pscore: \[ \hat{Pr}(T_i = 1|X_i) = \frac{m_i^T}{m_i^T + m_i^C} \]

≈ CEM:
- Gives a better pscore than PSM
- Doesn’t match based on crippled information
PSM Approximates Random Matching in Balanced Data

![Diagram](image)

- **PSM Matches**
- **CEM and MDM Matches**

Gary King (Harvard, IQSS)
Destroying CEM with PSM’s Two Step Approach

CEM Matches
CEM-generated PSM Matches

Gary King (Harvard, IQSS)
Conclusions

- **Propensity score matching:**
  - The problem:
    - Imbalance can be worse than original data
    - Can increase imbalance when removing the worst matches
    - Approximates random matching in well-balanced data
      (Random matching increases imbalance)
  - The Cause: unnecessary 1st stage dimension reduction
  - Implications:
    - Balance checking required
    - Adjusting for potentially irrelevant covariates *with PSM* is a mistake
    - Adjusting experimental data *with PSM* is a mistake
    - Reestimating the propensity score after eliminating noncommon support may be a mistake

- In four data sets and many simulations:
  - CEM > Mahalanobis > Propensity Score
  - *(Your performance may vary)*
  - CEM and Mahalanobis do not have PSM’s problems
  - You can easily check with the Space Graph
http://GKing.Harvard.edu/cem