Covariate Balancing Propensity Score

Kosuke Imai

Princeton University

June 8, 2012

Causal Inference Symposium
(ASA New Jersey Chapter)

Joint work with Marc Ratkovic

Motivation

- Causal inference is a central goal of scientific research
- Randomized experiments are not always possible
  $\Longrightarrow$ Causal inference in observational studies
- Experiments often lack external validity
  $\Longrightarrow$ Need to generalize experimental results
- Importance of statistical methods to adjust for confounding factors

Overview of the Talk

1. **Review:** Propensity score
   - conditional probability of treatment assignment
   - propensity score is a balancing score
   - matching and weighting methods
2. **Problem:** Propensity score tautology
   - sensitivity to model misspecification
   - adhoc specification searches
3. **Solution:** **Covariate balancing propensity score**
   - Estimate propensity score so that covariate balance is optimized
4. **Evidence:** Reanalysis of two prominent critiques
   - Improved performance of propensity score weighting and matching
5. **Extensions:**
   - Non-binary treatment regimes
   - Longitudinal data
   - Generalizing experimental and instrumental variable estimates

Propensity Score of Rosenbaum and Rubin (1983)

- Setup:
  - $T_i \in \{0, 1\}$: binary treatment
  - X_i: pre-treatment covariates
  - $(Y_i(1), Y_i(0))$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcomes
- Definition: conditional probability of treatment assignment

$$\pi(X_i) = \Pr(T_i = 1 \mid X_i)$$

- Balancing property:

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

- Assumptions:
  1. Overlap: $0 < \pi(X_i) < 1$
  2. Unconfoundedness: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i$
- The main result:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \pi(X_i)$$

Matching and Weighting via Propensity Score

- Propensity score reduces the dimension of covariates
- But, propensity score must be estimated (more on this later)
- Simple nonparametric adjustments are possible

- Matching
- Subclassification
- Weighting:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right\}$$

- Doubly-robust estimators (Robins *et al.*):

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\hat{\mu}(1,X_i) + \frac{T_i(Y_i - \hat{\mu}(1,X_i))}{\hat{\pi}(X_i)}\right\} - \left\{\hat{\mu}(0,X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}(0,X_i))}{1-\hat{\pi}(X_i)}\right\}\right]$$

- They have become standard tools for applied researchers

Propensity Score Tautology

- Propensity score is unknown
- Dimension reduction is purely theoretical: must model T_i given X_i
- Diagnostics: covariate balance checking
- In practice, adhoc specification searches are conducted
- Model misspecification is always possible

- Theory (Rubin *et al.*): ellipsoidal covariate distributions
  $\Longrightarrow$ equal percent bias reduction
- Skewed covariates are common in applied settings

- Propensity score methods can be sensitive to misspecification

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified
- Setup:
  - 4 covariates X_i^*: all are *i.i.d.* standard normal
  - Outcome model: linear model
  - Propensity score model: logistic model with linear predictors
  - Misspecification induced by measurement error:
    - $X_{i1} = \exp(X_{i1}^*/2)$
    - $X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)$
    - $X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3$
    - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$
- Weighting estimators to be evaluated:
  1. Horvitz-Thompson
  2. Inverse-probability weighting with normalized weights
  3. Weighted least squares regression
  4. Doubly-robust least squares regression

Weighting Estimators Do Fine If the Model is Correct

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(1) Both models correct					
$n = 200$	HT	-0.01	0.68	13.07	23.72
	IPW	-0.09	-0.11	4.01	4.90
	WLS	0.03	0.03	2.57	2.57
	DR	0.03	0.03	2.57	2.57
$n = 1000$	HT	-0.03	0.29	4.86	10.52
	IPW	-0.02	-0.01	1.73	2.25
	WLS	-0.00	-0.00	1.14	1.14
	DR	-0.00	-0.00	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	-0.32	-0.17	12.49	23.49
	IPW	-0.27	-0.35	3.94	4.90
	WLS	-0.07	-0.07	2.59	2.59
	DR	-0.07	-0.07	2.59	2.59
$n = 1000$	HT	0.03	0.01	4.93	10.62
	IPW	-0.02	-0.04	1.76	2.26
	WLS	-0.01	-0.01	1.14	1.14
	DR	-0.01	-0.01	1.14	1.14

Weighting Estimators Are Sensitive to Misspecification

Sample size	Estimator	Bias GLM	Bias True	RMSE GLM	RMSE True
(3) Outcome model correct					
$n = 200$	HT	24.72	0.25	141.09	23.76
	IPW	2.69	−0.17	10.51	4.89
	WLS	−1.95	0.49	3.86	3.31
	DR	0.01	0.01	2.62	2.56
$n = 1000$	HT	69.13	−0.10	1329.31	10.36
	IPW	6.20	−0.04	13.74	2.23
	WLS	−2.67	0.18	3.08	1.48
	DR	0.05	0.02	4.86	1.15
(4) Both models incorrect					
$n = 200$	HT	25.88	−0.14	186.53	23.65
	IPW	2.58	−0.24	10.32	4.92
	WLS	−1.96	0.47	3.86	3.31
	DR	−5.69	0.33	39.54	3.69
$n = 1000$	HT	60.60	0.05	1387.53	10.52
	IPW	6.18	−0.04	13.40	2.24
	WLS	−2.68	0.17	3.09	1.47
	DR	−20.20	0.07	615.05	1.75

Smith and Todd (2005, *J. of Econometrics*)

- LaLonde (1986; *Amer. Econ. Rev.*):
  - Randomized evaluation of a job training program
  - Replace experimental control group with another non-treated group
  - Current Population Survey and Panel Study for Income Dynamics
  - Many evaluation estimators didn't recover experimental benchmark
- Dehejia and Wahba (1999; *J. of Amer. Stat. Assoc.*):
  - Apply propensity score matching
  - Estimates are close to the experimental benchmark
- Smith and Todd (2005):
  - Dehejia & Wahba (DW)'s results are sensitive to model specification
  - They are also sensitive to the selection of comparison sample

Propensity Score Matching Fails Miserably

- One of the most difficult scenarios identified by Smith and Todd:
  - LaLonde experimental sample rather than DW sample
  - Experimental estimate: \$886 (s.e. $=$ 488)
  - PSID sample rather than CPS sample
- Evaluation bias:
  - Conditional probability of being in the experimental sample
  - Comparison between experimental control group and PSID sample
  - "True" estimate $= 0$
  - Logistic regression for propensity score
  - One-to-one nearest neighbor matching with replacement

Propensity score model	Estimates
Linear	-835
	(886)
Quadratic	-1620
	(1003)
Smith and Todd (2005)	-1910
	(1004)

Covariate Balancing Propensity Score

- Recall the dual characteristics of propensity score
  1. Conditional probability of treatment assignment
  2. Covariate balancing score
- Implied moment conditions:
  1. Score equation:

$$\mathbb{E}\left\{\frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)}\right\} = 0$$

  2. Balancing condition:
     - For the Average Treatment Effect (ATE)

$$\mathbb{E}\left\{\frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)}\right\} = 0$$

     - For the Average Treatment Effect for the Treated (ATT)

$$\mathbb{E}\left\{T_i \widetilde{X}_i - \frac{\pi_\beta(X_i)(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)}\right\} = 0$$

where $\widetilde{X}_i = f(X_i)$ is any vector-valued function

Generalized Method of Moments (GMM) Framework

- Over-identification: more moment conditions than parameters
- GMM (Hansen 1982):

$$\hat{\beta}_{\text{GMM}} = \underset{\beta \in \Theta}{\text{argmin}} \; \bar{g}_\beta(T,X)^\top \Sigma_\beta(T,X)^{-1} \bar{g}_\beta(T,X)$$

where

$$\bar{g}_\beta(T,X) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\begin{pmatrix} \frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)} \\ \frac{T_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)} \end{pmatrix}}_{g_\beta(T_i, X_i)}$$

- "Continuous updating" GMM estimator with the following Σ:

$$\Sigma_\beta(T,X) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(g_\beta(T_i,X_i) g_\beta(T_i,X_i)^\top \mid X_i)$$

- Newton-type optimization algorithm with MLE as starting values

Specification Test

- GMM over-identifying restriction test (Hansen)
- Null hypothesis: propensity score model is correct
- J statistic:

$$J = N \cdot \left\{ \bar{g}_{\hat{\beta}_{\text{GMM}}}(T, X)^\top \Sigma_{\hat{\beta}_{\text{GMM}}}(T, X)^{-1} \bar{g}_{\hat{\beta}_{\text{GMM}}}(T, X) \right\} \xrightarrow{d} \chi^2_{L+M}$$

- Failure to reject the null does not imply the model is correct
- An alternative estimation framework: empirical likelihood

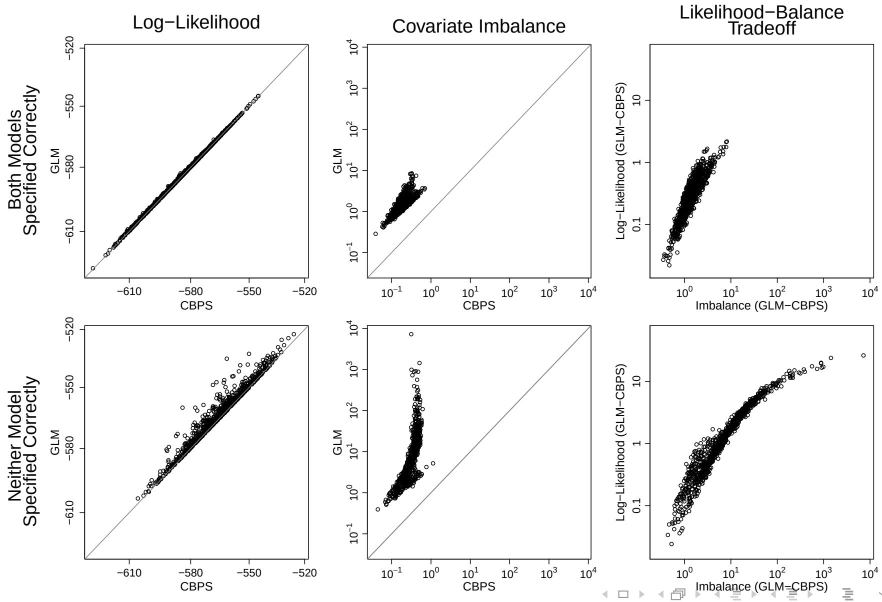
Revisiting Kang and Schafer (2007)

		Bias				RMSE			
Sample size	Estimator	GLM	Balance	CBPS	True	GLM	Balance	CBPS	True
(1) Both models correct									
$n = 200$	HT	−0.01	2.02	0.73	0.68	13.07	4.65	4.04	23.72
	IPW	−0.09	0.05	−0.09	−0.11	4.01	3.23	3.23	4.90
	WLS	0.03	0.03	0.03	0.03	2.57	2.57	2.57	2.57
	DR	0.03	0.03	0.03	0.03	2.57	2.57	2.57	2.57
$n = 1000$	HT	−0.03	0.39	0.15	0.29	4.86	1.77	1.80	10.52
	IPW	−0.02	0.00	−0.03	−0.01	1.73	1.44	1.45	2.25
	WLS	−0.00	−0.00	−0.00	−0.00	1.14	1.14	1.14	1.14
	DR	−0.00	−0.00	−0.00	−0.00	1.14	1.14	1.14	1.14
(2) Propensity score model correct									
$n = 200$	HT	−0.32	1.88	0.55	−0.17	12.49	4.67	4.06	23.49
	IPW	−0.27	−0.12	−0.26	−0.35	3.94	3.26	3.27	4.90
	WLS	−0.07	−0.07	−0.07	−0.07	2.59	2.59	2.59	2.59
	DR	−0.07	−0.07	−0.07	−0.07	2.59	2.59	2.59	2.59
$n = 1000$	HT	0.03	0.38	0.15	0.01	4.93	1.75	1.79	10.62
	IPW	−0.02	−0.00	−0.03	−0.04	1.76	1.45	1.46	2.26
	WLS	−0.01	−0.01	−0.01	−0.01	1.14	1.14	1.14	1.14
	DR	−0.01	−0.01	−0.01	−0.01	1.14	1.14	1.14	1.14

CBPS Makes Weighting Methods Work Better

		Bias				RMSE			
Sample size	Estimator	GLM	Balance	CBPS	True	GLM	Balance	CBPS	True
(3) Outcome model correct									
$n = 200$	HT	24.72	0.33	−0.47	0.25	141.09	4.55	3.70	23.76
	IPW	2.69	−0.71	−0.80	−0.17	10.51	3.50	3.51	4.89
	WLS	−1.95	−2.01	−1.99	0.49	3.86	3.88	3.88	3.31
	DR	0.01	0.01	0.01	0.01	2.62	2.56	2.56	2.56
$n = 1000$	HT	69.13	−2.14	−1.55	−0.10	1329.31	3.12	2.63	10.36
	IPW	6.20	−0.87	−0.73	−0.04	13.74	1.87	1.80	2.23
	WLS	−2.67	−2.68	−2.69	0.18	3.08	3.13	3.14	1.48
	DR	0.05	0.02	0.02	0.02	4.86	1.16	1.16	1.15
(4) Both models incorrect									
$n = 200$	HT	25.88	0.39	−0.41	−0.14	186.53	4.64	3.69	23.65
	IPW	2.58	−0.71	−0.80	−0.24	10.32	3.49	3.50	4.92
	WLS	−1.96	−2.01	−2.00	0.47	3.86	3.88	3.88	3.31
	DR	−5.69	−2.20	−2.18	0.33	39.54	4.22	4.23	3.69
$n = 1000$	HT	60.60	−2.16	−1.56	0.05	1387.53	3.11	2.62	10.52
	IPW	6.18	−0.87	−0.72	−0.04	13.40	1.86	1.80	2.24
	WLS	−2.68	−2.69	−2.70	0.17	3.09	3.14	3.15	1.47
	DR	−20.20	−2.89	−2.94	0.07	615.05	3.47	3.53	1.75

CBPS Sacrifices Likelihood for Better Balance



Revisiting Smith and Todd (2005)

- Evaluation bias: "true" bias $= 0$
- CBPS improves propensity score matching across specifications and matching methods
- However, specification test rejects the null

	1-to-1 Nearest Neighbor			Optimal 1-to-N Nearest Neighbor		
Specification	GLM	Balance	CBPS	GLM	Balance	CBPS
Linear	-835	-559	-302	-885	-257	-38
	(886)	(898)	(873)	(435)	(492)	(488)
Quadratic	-1620	-967	-1040	-1270	-306	-140
	(1003)	(882)	(831)	(406)	(407)	(392)
Smith & Todd	-1910	-1040	-1313	-1029	-672	-32
	(1004)	(860)	(800)	(413)	(387)	(397)

Standardized Covariate Imbalance

- Covariate imbalance in the (Optimal 1–to–N) matched sample
- Standardized difference-in-means

	Linear			Quadratic			Smith & Todd		
	GLM	Balance	CBPS	GLM	Balance	CBPS	GLM	Balance	CBPS
Age	−0.060	−0.035	−0.063	−0.060	−0.035	−0.063	−0.031	0.035	−0.013
Education	−0.208	−0.142	−0.126	−0.208	−0.142	−0.126	−0.262	−0.168	−0.108
Black	−0.087	0.005	−0.022	−0.087	0.005	−0.022	−0.082	−0.032	−0.093
Married	0.145	0.028	0.037	0.145	0.028	0.037	0.171	0.031	0.029
High school	0.133	0.089	0.174	0.133	0.089	0.174	0.189	0.095	0.160
74 earnings	−0.090	0.025	0.039	−0.090	0.025	0.039	−0.079	0.011	0.019
75 earnings	−0.118	0.014	0.043	−0.118	0.014	0.043	−0.120	−0.010	0.041
Hispanic	0.104	−0.013	0.000	0.104	−0.013	0.000	0.061	0.034	0.102
74 employed	0.083	0.051	−0.017	0.083	0.051	−0.017	0.059	0.068	0.022
75 employed	0.073	−0.023	−0.036	0.073	−0.023	−0.036	0.099	−0.027	−0.098
Log-likelihood	−326	−342	−345	−293	−307	−297	−295	−231	−296
Imbalance	0.507	0.264	0.312	0.544	0.304	0.300	0.515	0.359	0.383

Extensions to Other Causal Inference Settings

- Propensity score methods are widely applicable
- This means that CBPS is also widely applicable
- Potential extensions:
  1. Non-binary treatment regimes
  2. Causal inference with longitudinal data
  3. Generalizing experimental estimates
  4. Generalizing instrumental variable estimates
- All of these are situations where balance checking is difficult

Non-binary Treatment Regimes

- Multi-valued treatment regime: $T_i \in \{0, 1, \ldots, K-1\}$
- Propensity scores: $\pi_\beta^k(X_i) = \Pr(T_i = k \mid X_i)$
- Score equation: multinomial likelihood
- Balancing moment conditions:

$$\mathbb{E}\left\{\frac{\mathbf{1}\{T_i = k\}\widetilde{X}_i}{\pi_\beta^k(X_i)} - \frac{\mathbf{1}\{T_i = k-1\}\widetilde{X}_i}{\pi_\beta^{k-1}(X_i)}\right\} = 0$$

for each $k = 1, \ldots, K-1$.

Generalizing Experimental Estimates

- Lack of external validity for experimental estimates
- Target population $\mathcal{P}$
- Experimental sample: $S_i = 1$ with $i = 1, 2, \ldots, N_e$
- Non-experimental sample: $S_i = 0$ with $i = N_e + 1, \ldots, N$
- Sampling on observables: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp S_i \mid X_i$
- Propensity score: $\pi_\beta(X_i) = \Pr(S_i \mid X_i)$
- Weighted regression with the weight $= 1/\pi_\beta(X_i)$
- Score equation: binomial likelihood
- Balancing between experimental and non-experimental sample:

$$\mathbb{E}\left\{\frac{S_i \widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1 - S_i)\widetilde{X}_i}{1 - \pi_\beta(X_i)}\right\} \quad = \quad 0$$

- You may also balance weighted treatment and control groups

Causal Inference with Longitudinal Data

- Time-dependent confounding and time-varying treatments
- Notation:
  - N units
  - J time periods
  - Outcome Y_{ij}
  - Treatment: T_{ij}
  - Treatment history: $\overline{T}_{ij} = \{T_{i0}, T_{i1}, \ldots, T_{ij}\}$
  - Covariates: X_{ij}
  - Covariate history: $\overline{X}_{ij} = \{X_{i0}, X_{i1}, \ldots, X_{ij}\}$
- Assumption: Sequential ignorability

$$\{Y_{ij}(1), Y_{ij}(0)\} \perp\!\!\!\perp T_{ij} \mid \overline{T}_{i,j-1}, \overline{X}_{ij}$$

- Propensity score:

$$\pi_\beta(\overline{T}_{i,j-1}, \overline{X}_{ij}) = \Pr(T_{ij} = 1 \mid \overline{T}_{i,j-1}, \overline{X}_{ij})$$

Marginal Structural Models

- Weighted regression of Y_{ij} given $\overline{T}_{ij}$ where the stabilized weight for unit i at time j is given by

$$w_{ij} = \frac{\prod_{j'=1}^{j} \Pr(T_j = T_{ij'} \mid \overline{T}_{j'-1} = \overline{T}_{i,j'-1})}{\prod_{j'=1}^{j} \pi_\beta(\overline{T}_{i,j-1}, \overline{X}_{ij})}$$

- Do not adjust for $\overline{X}_{ij}$ in outcome regression $\Longrightarrow$ posttreatment bias

- Challenge: balance covariates at each time period
- The score equation: logistic regression
- The balancing moment conditions (for each time period j):

$$\mathbb{E}\left\{\frac{T_{ij}\widetilde{Z}_{ij}}{\pi_\beta(\overline{T}_{i,j-1}\overline{X}_{ij})} - \frac{(1-T_{ij})\widetilde{Z}_{ij}}{1-\pi_\beta(\overline{T}_{i,j-1}, \overline{X}_{ij})}\right\} = 0$$

where $\overline{Z}_{ij} = f(\overline{T}_{i,j-1}, \overline{X}_{ij})$

Review of Instrumental Variables (Angrist et al. *JASA*)

- Encouragement design
- Randomized encouragement: $Z_i \in \{0, 1\}$
- Potential treatment variables: $T_i(z)$ for $z = 0, 1$
- Four principal strata (latent types):
  - compliers $(T_i(1), T_i(0)) = (1, 0)$,
  - non-compliers $\begin{cases} \textit{always} - \textit{takers} & (T_i(1), T_i(0)) = (1, 1), \\ \textit{never} - \textit{takers} & (T_i(1), T_i(0)) = (0, 0), \\ \textit{defiers} & (T_i(1), T_i(0)) = (0, 1) \end{cases}$
- Observed and principal strata:

	$Z_i = 1$	$Z_i = 0$
$T_i = 1$	Complier/Always-taker	Defier/Always-taker
$T_i = 0$	Defier/Never-taker	Complier/Never-taker

- Randomized encouragement as an instrument for the treatment
- Two additional assumptions
  1. Monotonicity: No defiers

$$T_i(1) \geq T_i(0) \quad \text{for all } i.$$

  2. Exclusion restriction: Instrument (encouragement) affects outcome only through treatment

$$Y_i(1, t) = Y_i(0, t) \quad \text{for } t = 0, 1$$

  Zero ITT effect for always-takers and never-takers

- ITT effect decomposition:

$$\begin{aligned} \text{ITT} &= \text{ITT}_c \times \Pr(\text{compliers}) + \text{ITT}_a \times \Pr(\text{always} - \text{takers}) \\ &\quad + \text{ITT}_n \times \Pr(\text{never} - \text{takers}) \\ &= \text{ITT}_c \Pr(\text{compliers}) \end{aligned}$$

- Complier average treatment effect or (LATE): $\text{ITT}_c = \text{ITT} / \Pr(\text{compliers})$

Generalizing Instrumental Variables Estimates

- Compliers may not be of interest
  1. They are a latent type
  2. They depend on the encouragement
- Generalize LATE to ATE
- No unmeasured confounding: ATE $=$ LATE given X_i
- Propensity score: $\pi_\beta(X_i) = \Pr(C_i = c \mid X_i)$
- Weighted two-stage least squares with the weight $= 1/\pi_\beta(X_i)$
- Score equation is based on the mixture likelihood:
- Balancing moment conditions: weight each of the four cells and balance moments across them

Concluding Remarks

- Covariate balancing propensity score:
  1. simultaneously optimizes prediction of treatment assignment and covariate balance under the GMM framework
  2. is robust to model misspecification
  3. improves propensity score weighting and matching methods
  4. can be extended to various situations
- Open questions:
  1. Empirical performance of proposed extensions
  2. How to choose model specifications and balancing conditions
- Open-source software in development