STA 640 — Causal Inference

Chapter 3.2: Observational Studies - Stratification and Matching

Fan Li

Department of Statistical Science
Duke University

Covariate Balance: standardized difference

- Under strong ignorability, valid causal inference can be obtained by comparing the observed distributions of Y under treatment and control if the covariates are balanced
- In a causal study, a good practice is always to first check covariate balance
- Many metrics of balance - the most common one is the absolute standardized difference (ASD)

$$\text{ASD}_1 = \left| \frac{\sum_{i=1}^{N} X_i Z_i}{\sum_{i=1}^{N} Z_i} - \frac{\sum_{i=1}^{N} X_i (1 - Z_i)}{\sum_{i=1}^{N} (1 - Z_i)} \right| \Bigg/ \sqrt{s_1^2 / N_1 + s_0^2 / N_0},$$

  where s_z^2 is the sample variance of the covariate in group z for $z = 0, 1$
- For a continuous covariate, ABD is the standard two-sample t-statistic, and the threshold is based on a t- or z- test (e.g. 1.96)

Covariate Balance: standardized difference

- Debate on whether N_0 and N_1 should be in the denominator: with large sample size, imbalance will always be declared based on ASD_1
- In comparative effectiveness research and some disciplines, the ASD is often defined as:

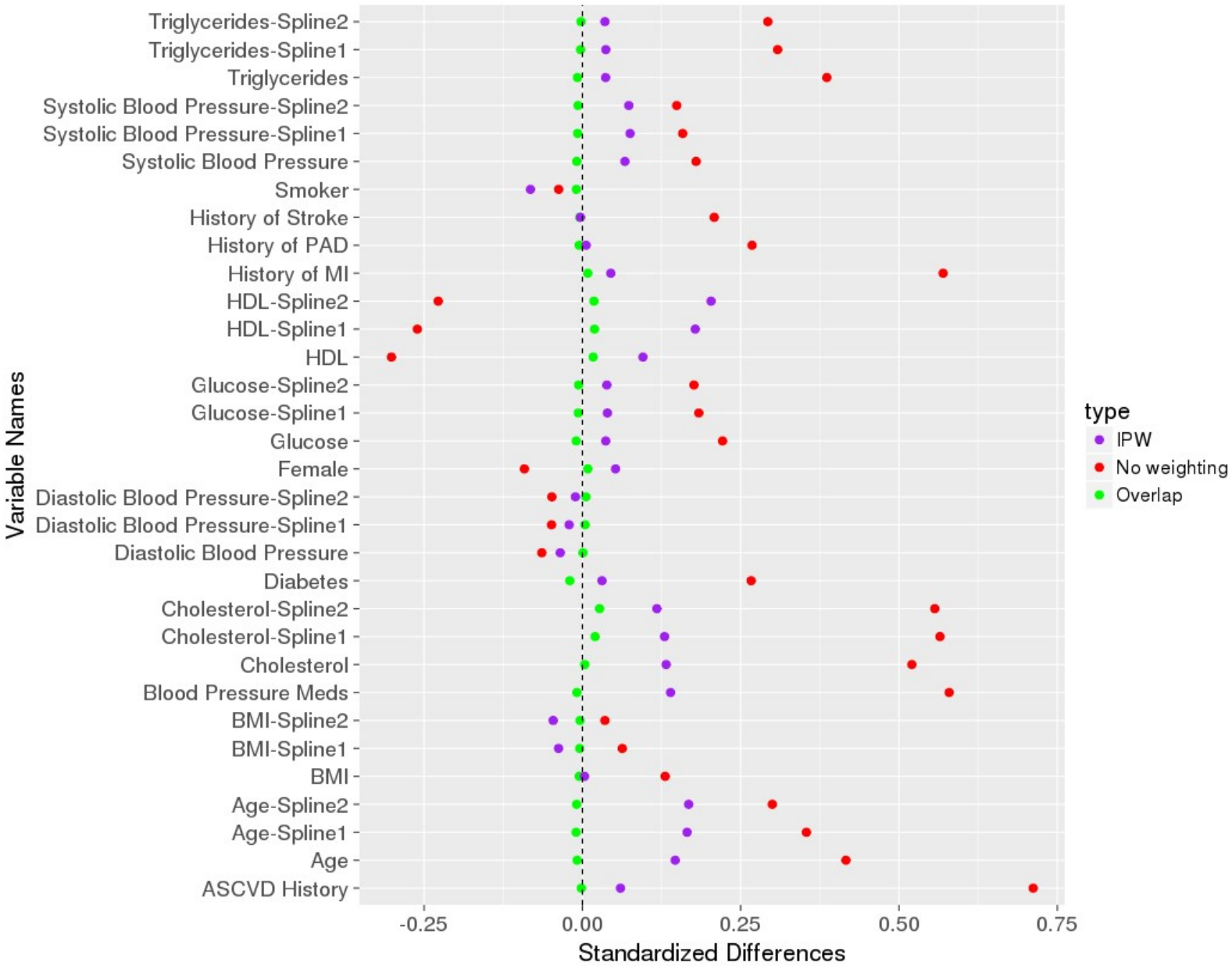
$$\text{ASD}_2 = \left| \frac{\sum_{i=1}^{N} X_i Z_i}{\sum_{i=1}^{N} Z_i} - \frac{\sum_{i=1}^{N} X_i (1 - Z_i)}{\sum_{i=1}^{N} (1 - Z_i)} \right| \Big/ \sqrt{s_1^2 + s_0^2},$$

- The common threshold is 0.1 (Austin, 2011)
- More general, multivariate, balance metrics are available: e.g. (1) variance ratios, (2) Kolmogorov-Smirnov (KS) statistic
- Always good to check higher order terms and interactions

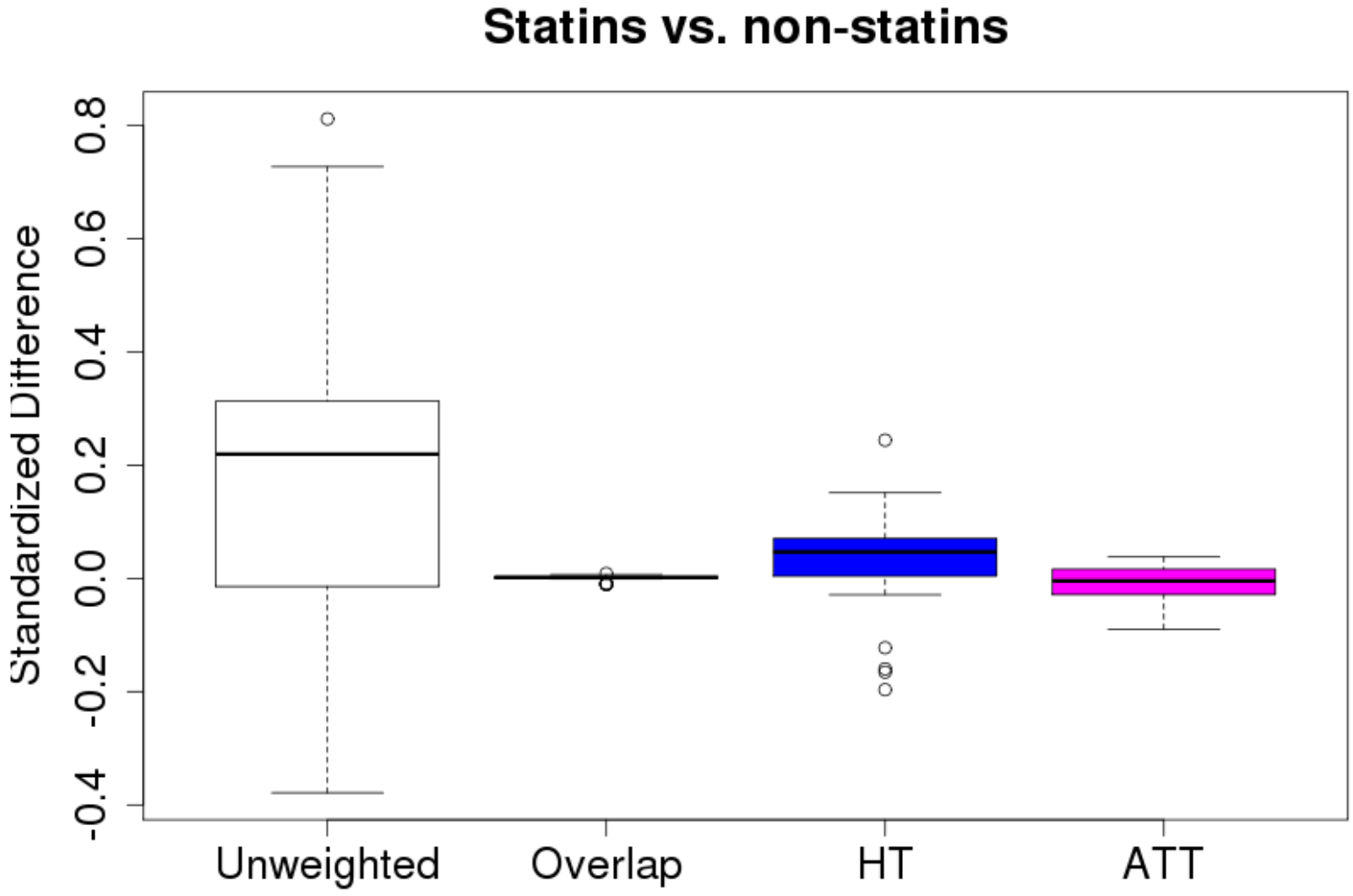
Visualize Balance

- Covariate-by-covariate before-after table (see Framingham example)
- Boxplot: of the ASDs or standardized diff of all covariates (can include higher order terms and splines)
- Love plot: of the ASDs or standardized diff of all covariates (can include higher order terms and splines)
- The Love plot is named after Thomas Love

Love Plot of Framingham Study



Boxplot of Standardized Difference in Framingham Study



Balancing covariates: small number of covariates

- When the number of covariates is small, the adjustment can be achieved by exact matching or stratification
- Exact matching: for each treated subject, get a control with exact same value of the covariate
- Exact matching ensures distributions of covariates in treatment and control groups are exactly the same, thus eliminate bias due to difference in X.
- Exact matching is usually infeasible, even with low dimensional covariates

Stratification

- Suppose we have a single covariate X with k levels (e.g. age in tens) and X makes assignment ignorable. Want to estimate the ATE τ.
- We have: $\mathbb{E}[Y(1)] = \sum_k \mathbb{E}(Y|X_i = k, Z = 1)Pr(X = k)$
- n_k: number of people in cell $X_i = k$; $\bar{Y}_{k,z}$: the sample average of Y among people in the cell $X_i = k$ and $Z_i = z$
- Estimate $\mathbb{E}[Y(1)]$ by a consistent estimator $\sum_k \bar{Y}_{k,1} \frac{n_k}{n}$.
  Therefore, τ can be estimated by:

$$\hat{\tau} = \sum_k (\bar{Y}_{k,1} - \bar{Y}_{k,0}) \frac{n_k}{n} \tag{1}$$

- Note: Equation (1) is not generally applicable for a non-linear contrast of the potential outcomes.

Stratification

- What if X is continuous?
- Stratification (subclassification): split X into k classes. Then for class k, define $n_k, \hat{Y}_{k,z}$ as before. An estimator of ATE τ is:

$$\hat{\tau}^k = \sum_k (\bar{Y}_{k,1} - \bar{Y}_{k,0}) \frac{n_k}{n}$$

- $\hat{\tau}^k$ is generally biased for τ
- Denote $R_k = 1 - \frac{\mathbb{E}(\hat{\tau}^k) - \tau}{\mathbb{E}(\hat{\tau}^1) - \tau}$. Cochran (1968) showed that $R^k \approx 90\%$ for $k \geq 5$ for a large class of underlying models
- Therefore, generally stratification of over 5 blocks can remove 90% of the bias!

Matching

- Regression estimators impute the missing potential outcomes using the estimated regression function
- Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of nearest neighbours of the opposite treatment group (similar to nonparametric kernel regression methods)
- They have often (but not exclusively) been applied in settings where
  - the interest is in the ATT
  - and there is a large reservoir of potential controls. This allows matching each treated unit to one or more distinct controls (matching without replacement)
- More general settings: both treated and control units are (potentially) matched and matching is done with replacement

Nearest-Neighbor (NN) Matching with Fixed Number of Matches

- let $\mathcal{M}_i$ be the set of the indices of the M closest matches of unit i in terms of the distance measure based on the norm $\|\cdot\|$

$$\sum_{j|W_j \neq Z_i} 1\{\|X_j - X_i\| \leq \|X_l - X_i\|\} = M$$

- Let

$$\hat{Y}_i(0) = \begin{cases} \sum_{j \in \mathcal{M}_i} Y_j / M, & Z_i = 1, \\ Y_i, & Z_i = 0, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} Y_i, & Z_i = 1, \\ \sum_{j \in \mathcal{M}_i} Y_j / M, & Z_i = 0. \end{cases}$$

Nearest-Neighbor (NN) Matching with Fixed Number of Matches

- The treatment effect within a pair is then estimated as the difference in outcomes, and then average these within-pair difference

$$\hat{\tau}^{\text{ATE}} = \sum_i \left(\hat{Y}_i(1) - \hat{Y}_i(0)\right) \Big/ N,$$

$$\hat{\tau}^{\text{ATT}} = \sum_i \left(Y_i - \hat{Y}_i(0)\right) Z_i \Big/ N_1.$$

- Pros: Matching estimators ensure good balance in covariates between groups are generally robust
- Cons: With fixed number of matches and matching with replacement, the NN matching estimators are generally biased, the asymptotic bias is of the order $O(N^{-1/p})$, where p is the number of continuous covariates (Abadie and Imbens, 2006)
- Intuition: matching is a non-smooth procedure

Nearest-Neighbor (NN) Matching: Variance Estimation

- With fixed number of matches and matching with replacement, bootstrap estimate of s.e. of simple NN matching estimators is generally biased (Abadie and Imbens, 2008)
- Intuition: bootstrap fails to reproduce the distribution of the number of times each unit is used as a match
- Abadie and Imbens (2006) use normal approximation to derive asymptotic variance of NN matching estimators
- Weighted bootstrap (Otsu et al., 2017), subsampling inference (Politis and Romano, 1994)
- Matching estimators are generally not efficient, estimators combining matching and regression adjustment are usually more efficient

Matching: Tuning

Matching involves lots of tuning

- distance metric
- fixed or varying no. matches
- for fixed M, number of matches
- with or without replacement

Tuning for matching is an art, with some theory and general guidelines available...

Matching: Tuning

- Distance metric (later): Mahalanobis distance, propensity score, tree-based
- Fixed M or varying M? For varying M:
  - Matching with caliper: define a caliper (say 0.1) and all units within that caliper are matches
  - M increases with sample size (e.g. kernel-based matching (Heckman et al. 1998))
- For fixed M, the choice of M (number of matches per unit) has a bias-variance tradeoff: smaller M, smaller bias but larger variance; larger M, larger bias but smaller variance
- Also depends on the proportion of treatment versus control: when there is a much larger control group, can use 1-to-many matching

Matching: Tuning

- Matching with replacement:
  Pros: (1) computationally easier, (2) both controls and treated can be matched, but with high variances, and (3) not order-depended;
  Cons: some units (especially ones with extreme ps) can be matched many times and thus heavily influence overall estimates, similar to extreme weights in weighting

- Deal with ties

- Matching is a vast topic: an excellent review up to 2010 is Stuart (2010), but many new matching methods have been developed since (e.g. Rosenbaum and students)

- Software in R: `Matching` (Sekhon) (this one is faster and easier to use by my experience); `Matchit` (Ho et al.), and many more

Matching: Target Population?

- Even with the same sample, different matching methods may lead to very different matched samples, which may corresponds to very different subpopulation
- So what is the target population (the one causal effect is defined on)? Can be ambiguous
- It is irrelevant to compare different matching methods corresponding to different target populations, when treatment effect is heterogenous
- Also applying the same matching method to different samples can lead to very different target populations and conclusions. E.g. samples with different proportions of treatment and control
- Always ask ahead what is your target population?

Distance Metric: Mahalanobis distance

- Mahalanobis distance (Mahalanobis, 1936).
- For two random vector $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$ and $\mathbf{y} = (y_1, y_2, ..., y_p)'$ from the the same distribution with the covariance matrix S, the Mahalanobis distance is:

$$d(\boldsymbol{x}, \mathbf{y}) = \sqrt{(\boldsymbol{x} - \mathbf{y})' S^{-1} (\boldsymbol{x} - \mathbf{y})}.$$

- If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the *Euclidean distance*
- If the covariance matrix is diagonal, reduces a *normalized Euclidean distance*

$$d(\boldsymbol{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2 / s_i^2},$$

where s_i is the standard deviation of x_i (and y_i)

- Widely applicable, but computational intensive for high dimensions

Dimensional Reduction

- What if there is a large number of covariates? With just 20 binary covariates, there are 2^{20} or about a million covariate patterns
- Direct matching or stratification is nearly impossible
- Need dimensional reduction: propensity score

Matching as a Pre-Processing Step

- More generally, matching can be used as a pre-processing step for causal inference with observational data
  - Use a generic matching method to obtain a matched sample (nothing optimal), to ensure good balance
  - Then use regression (e.g. bias-corrected, or simply use the regression coefficient as causal estimate) or weighting on the matched sample to correct for residual imbalance and improve efficiency
- Empirical results (e.g. on the famous Lalonde data) show that once a matched sample with reasonable balance is obtained, the point estimates from different additional methods are usually similar
- In high dimensional cases, combining matching (or stratification or weighting) with regression is the way to go (i.e. augmented or double learning)
- Matching is the starting point, not the end

Bias-corrected matching: mix matching with regression

- Residual imbalance in matching
- Rubin (1973): perform bias correction via regression on the matched sample
- Abadie and Imbens (2011) provided theoretical basis and software for bias-corrected matching estimator
- Let $\mu_z(\boldsymbol{x}) = \mathbb{E}[Y(z)|\mathbf{X} = \boldsymbol{x}]$, and $\hat{\mu}_z(\mathbf{X}_i)$ be a consistent estimator of $\mu_z(\mathbf{X}_i)$, for $z = 0, 1$. A regression estimator uses $\hat{\mu}_z(\mathbf{X}_i)$ to impute missing potential outcomes $Y_i(z)$.

Bias-corrected matching: mix matching with regression

- Let

$$\tilde{Y}_i(0) = \begin{cases} \sum_{j \in \mathcal{M}_i} \left[Y_j + \hat{\mu}_0(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_j) \right] / M, & Z_i = 1, \\ Y_i, & Z_i = 0, \end{cases}$$

$$\tilde{Y}_i(1) = \begin{cases} Y_i, & Z_i = 1, \\ \sum_{j \in \mathcal{M}_i} \left[Y_j + \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_1(\mathbf{X}_j) \right] / M, & Z_i = 0. \end{cases}$$

- The bias-corrected matching estimators:

$$\hat{\tau}^{\text{ATE}}_{\text{mix}} = \sum_i \left(\tilde{Y}_i(1) - \tilde{Y}_i(0) \right) \Big/ N$$

$$\hat{\tau}^{\text{ATT}}_{\text{mix}} = \sum_i \left(Y_i - \tilde{Y}_i(0) \right) Z_i \Big/ N_1 .$$

- Empirical evidence that the mixed method outperform its nonparametric matching counterpart.

References

Abadie, A., Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. Econometrica, 74(1), 235-267.

Abadie, A., Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. Econometrica, 76(6), 1537-1557.

Abadie, A., Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. Journal of Business and Economic Statistics, 29(1), 1-11.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3), 399-424.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, 295-313.

Heckman, J. J., Ichimura, H., Todd, P. (1998). Matching as an econometric evaluation estimator. The review of economic studies, 65(2), 261-294.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Politis, D. N., Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. The Annals of Statistics, 2031-2050.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics, 185-203.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, I: Some examples. Biometrics, 109-120.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science, 25(1), 1.