観察研究において選択biasを制御するために用いられるPropensity Score IPTWと層化調整法の、頑健性の観点からの使い分けについて

古川敏仁
株式会社バイオスタティスティカル　リサーチ　代表取締役

IPTW or Adjustments using the estimated propensity score, which, how and why can we use from a robustness and bias reduction points of view?

Toshihito Furukawa
President, Biostatistical Research Co,ltd.

要旨

観察研究においては選択 bias を制御するために Propensity Score を用いた IPTW (Inverse Probability of Treatment Weighted)や、Propensity Score を層化変数とした背景因子の調整が行われる。生存時間解析においては､SAS　Ver9.以降 WEIGHT ステートメントが利用できるようになり IPTW がより容易に解析できるようになった。しかしながら、Propensity Score は、被験者の情報は観察された n 個の変数ですべて説明されるという strongly ignorable な仮定の上に成り立つ手法であり、この仮説上の制約、あるいは、Propensity Score 計算上の線形性の仮定に基づく制約がある。

ゆえに、今回、

1. Propensity Score の理論的背景からこれらの手法の原理を説明し、
2. 具体的な SAS coding でそのイメージを具現化し、
3. 作成した Propensity Score の Bias reduction の評価方法を述べ、
4. 理論的な制約や頑健性の観点から両者の使用上のポイントを解説する。

キーワード：propensity score, adjustment, IPTW, bias reduction, observational study

はじめに

例えば観察研究のデータで特定の薬剤の効果や、医療機器や手術などの効果を対照群と比較する場合、無作為化比較臨床試験とは違い治療群は無作為化割付されていないため、比較する治療群の被験者背景は異なり、治療効果の比較は難しい、あるいは、推定結果には被験者の選択 bias が介在することになる。この選択 bias を軽減するために、治療効果に影響がある背景項目を共変量としてモデルに組み込み、目的とする治療群の効果を調整(adjust)することが広く利用されている。しかし、治療効果に影響がある共変量をすべて調整

に用いようとすると、モデルに組み込まれる共変量の数が増え、モデルの線形性の制約から調整には限界があり、また、治療効果のパラメータ推定値自体も不安定なものとなる。つまり、従来広く用いられてきた共変量の調整による被験者選択 biais の軽減には限界があった。そこで、その解決法として propensity score(PS)を用いた手法が数多く提案されてきた。その原理は以下である。

propensity score の基本的なコンセプト：PS の基本的なコンセプトは至って明快である。今、T1 群（研究対象の治療群）、T0 群（対照群）、2 つの治療効果の比較を考えると以下のようになる。

- もし、治療 T1 と T0 の治療効果の差が知りたければ、同じ被験者に同時に T1 と T0 の治療を行い、応答 Y1 と Y0 を観察し、その差の集団平均 E[Y1－Y0]（平均因果効果 average causal effects）を求めればよい。
- 上記は、最も明快な T1 と T0 の治療効果の推定ではあるが、しかし、残念ながら 1 人の被験者には 1 つの治療しか行うことができない(つまり、仮に T1 の治療を行えば Y1 しか実際のデータは観察されない)。
- そこで、すべての治療効果に影響を与える被験者背景が同じ患者は同じ治療効果を持つ（同一人と考えて良い）と考え、観察された被験者背景で被験者に与える治療効果はすべて説明できるという strongly ignorable な仮定が成り立つとすれば、以下のような Rubin Causal Model[1)]の展開から、観察された被験者背景と応答（Y1 もしくは Y0）から平均因果効果を推定することができる。

Rubin Causal Model

平均因果効果 E[Y1－Y0]は、Y1 と Y0 が独立ならば以下となる。

E[Y1－Y0]=E[Y1]—E[Y0]　(1)

しかし、実際に観察されるデータは、T1 と T0 に割り付けられた被験者は背景が違うため、

E[Y1]—E[Y0]≠E[Y1|T1]—E[Y0|T2]　となり、観察されたデータから平均因果効果を単順に推定することはできない。

治療群　T1 に割り付けられた治療効果　E[Y1|T1]

対照群　T0 に割り付けられた治療効果　E[Y0|T0]

（ただし、T1: 治療群　への割付状態、T0: 対照群への割付状態を示す。）

これが、無作為化比較試験であれば、無作為化割付により T1、T0 の割付は被験者背景とは独立に実施され、その結果 T1、T0 と Y1、Y0 は独立となる。ゆえに、(2)式により、無作為化比較試験では平均因果効果 E[Y1－Y0]は、観察されたデータから推定することができる。これが、臨床試験に無作為化が必要な理由である。

E[Y1|T1]＝E[Y1|T0]＝E[Y1]、E[Y0|T0]＝E[Y0|T1]＝E[Y0]　より、

E[Y1|T1]—E[Y0|T2]= E[Y1]—E[Y0]= E[Y1－Y0]　(2)

しかし、観察研究では Y0 と T0、Y1 と T0 は独立ではないため、このスキムは利用できない。そこで以下のように展開する[2) 3)]。

今、ある被験者 u が被験者背景 ***Z***（Z1：重症度、Z2：年齢、Z3：性別、･･･Zp、Zp+1、･･･）を持ち、被験者の情報は観察された p 個の情報ですべて説明されるという strongly ignorable な仮定をおくと、p 個の被験者背景が同じなら、同一人物と同等と見なすことができ、そのようなペア集団からは平均因果効果を推定することができる。

E[y1(T=1|Z=z） － y0(T=0|Z=z）)　　≒E[Y1]—E[Y0]

strongly ignorable な仮定のもとでは、Z が同一ならば同一被験者と考えることができるので、Z を与えた下で Y（応答）と T(割付)は条件付独立となる。

$$\begin{pmatrix} Y0 \\ Y1 \end{pmatrix} \perp T \mid Z = z \qquad (3)$$

今 T(0,1)を、T=1 治療群への割付、T=0 対照群へ割付を意味する変数とし、Propensity score e(z)を、e(z)=E[T=1|z] 、すなわち、被験者背景 z をもつ被験者が T=1、すなわち、治療群に割り付けられる確率と定義すると

$$E\left[\frac{Y1\,T}{e(z)}\middle|z\right] = E[Y1 \mid z] \times E\left[\frac{T}{e(z)}\middle|z\right] = E[Y1 \mid z] \qquad (4)$$

$\because Y1$とTは、z　givenのもとで条件付独立

$\because E[T \mid z] = e(z)$

すると、$E[Y1 \mid z]$は(5)式のように展開できる。

$$E[Y1 \mid z] = E\left[\frac{Y1\,T}{e(z)}\middle|z\right] \qquad (5)$$

任意の確率変数X、Yにおいて$E[Y] = E_x[E[Y \mid X]]$だから、(5)式をzに関して期待値を求めると$E[Y1]$を推定することができる

$$E[Y1] = E_z\{E[Y1 \mid z]\} = E_z\left\{E\left[\frac{Y1\,T}{e(z)}\middle|z\right]\right\} = E\left[\frac{Y1\,T}{e(z)}\right] \qquad (6)$$

同様に

$$E[Y0] = E\left[\frac{Y0(1-T)}{1\text{-}e(z)}\right] \qquad (7)$$

(6)、(7)式を具体的に現実のデータで推定すると、n 人の患者集団のうち、m 人が T=0 治療を行い、n-m 人が T=1 治療を行ったとすれば、E[y1]と E[y0]の推定値は以下となる。

$$E[Y1] = E\left[\frac{Y1\,T}{e(\mathbf{z})}\right]$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y1}i \approx \frac{1}{n}\sum_{i=1}^{n}\frac{Y1i\,t_i}{e(\mathbf{z}i)}$$

$$= \frac{1}{n}\sum_{i=(tm+1)}^{n}\frac{Y1i}{e(\mathbf{z}i)} \approx \frac{1}{\sum_{i=(m+1)}^{n}\frac{1}{e(\mathbf{z}i)}}\sum_{i=(m+1)}^{n}\frac{Y1i}{e(\mathbf{z}i)} + 0 \qquad (8)$$

$$\because E\left[\sum_{i=(m+1)}\frac{1}{e(\mathbf{z}i)}\right] = E\left[\sum_{i=1}^{n}\frac{ti}{e(\mathbf{z}i)}\right] = nE\left[\frac{T}{e(\mathbf{z}i)}\right] = n$$

同様に

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y0}i \approx \frac{1}{\sum_{i=1}^{m}\frac{1}{1-e(\mathbf{z}i)}}\sum_{i=1}^{m}\frac{Y0i}{1-e(\mathbf{z}i)} + 0 \qquad (9)$$

$$E[Y1-Y0\} \approx \frac{1}{\sum_{i=(m+1)}^{n}\frac{1}{e(\mathbf{z}i)}}\sum_{i=(m+1)}^{n}\frac{Y1i}{e(\mathbf{z}i)} - \frac{1}{\sum_{i=1}^{m}\frac{1}{1-e(\mathbf{z}i)}}\sum_{i=1}^{m}\frac{Y0i}{1-e(\mathbf{z}i)} \qquad (10)$$

IPTW 法と Propensity Score(PS)による調整解析法

IPTW(Inverse Probability of Treatment Weighted)とは、(10)式をそのまま利用した方法で、治療群の平均効果 Y1 と対照群の平均効果 Y0 は、被験者背景から計算した各群に属する確率の逆数を重みとした、各群の治療効果 yi の重み付け平均として計算される。ただし、治療群の確率 e(zi)、対照群の確率 1-e(zj)の逆数をそのまま用いると、小数例の群の重みが大きくなるので、各群の割合 p1 と p0 で群に帰属する確率を除した条件付確率の逆数の重み wki が用いられる。

$$\text{治療群の重み} w1i = \frac{p1}{e(zi)}$$

$$\text{対照群の重み} w0\,ji = \frac{1-p1}{1-e(zj)}$$

これら式の意味は、被験者背景から勘案して当該治療群に割り付けられにくい確率をもつ被験者の重みを相対的に重くしてやれば、治療群、対照群間の被験者背景の偏りは補正されるということである。わかりやすく、SAS coding で示せば以下となる。

```
/* IPTW Coding    */
Proc logistic data=被験者背景と治療群が存在するデータセット;
 Model TRT(EVENT='1')=A B C D E ‥‥;   /* TRT 1=治療群、0=対照群  */
                                        /* A,B,C‥‥観察された背景因子  */
 OUTPUT OUT=PS PRED=P;                  /* P: Propensity Score */
 RUN;
 DATA PS;
  SET PS ;
  IF TRT=1 THEN WGT=1/P*&n1./&N;     /* &n1  治療群の例数、N   全例数、&n0   対照群の例数  */
  IF TRT=0 THEN WGT=1/(1-P)*&n0/&N; /* WGT IPTW の重み  */;
 RUN;
```

これを例えば、Cox 回帰に用いれば以下となる。

```
 PROC PHREG DATA=解析用変数と WGT が存在するデータセット ;
   MODEL TIME*CENSOR(0)=TRT ;
  WEIGHT WGT;
 RUN;
```

一方、生存時間解析における PS による調整解析法では、被験者背景をもとに算出した治療群に属する確率=Propensity score を、イベントの分布に応じて適切に 5 区分した層化変数を調整因子として用いる。この意味は、複数の被験者背景の情報を 1 つの Propensity score に縮約するということである。なぜ、5 区分かというと、5 区分以上の層化は調整と調整の効率をあまり向上することはできず、実際的にイベント数から、5 区分以上の層化は難しいからである。また、Cochran[4) 5)]や Rosenbaum and Rubin[6)]によれば、傾向スコアが 5 区分化でき、充分解析に適する分布であれば、背景因子の偏りによるバイアスを 90%減少できると報告されてい

る。また、4 区分では 85%、3 区分でも 80%のバイアスが除去できることが報告されている。わかりやすく、SAS coding で示せば以下となる。

```
/* PS による調整解析法 Coding　*/
Proc logistic data=被験者背景と治療群が存在するデータセット;
    Model TRT(EVENT='1')=A B C D E ‥‥;　/* TRT 1=治療群、0=対照群 */
                                          /* A,B,C‥‥観察された背景因子 */
    OUTPUT OUT=PS PRED=P;                 /* P: Propensity Score */
 RUN;

DATA PS;
  SET PS ;
  IF P>0 AND P<=?? THEN PS1=1;ELSE PS1=0;　/*　PS を適切に 5 区分　*/
  IF P>?? AND P<=??? THEN PS2=1;ELSE PS2=0;
  ・・
  IF P>???? THEN PS5=1;ELSE PS5=0;
RUN;
```

これを例えば、Cox 回帰に用いれば以下となる。

```
  PROC PHREG DATA=解析用変数と PS1,PS2, ‥,,PS5 が存在するデータセット ;
   MODEL TIME*CENSOR(0)=TRT　PS1 PS2 PS3 PS4;
  /*あるいは　STRATA　PS を 5 区分した変数 */
 RUN;
```

IPTW 法と PS による調整解析法に共通する重要事項

IPTW 法も PS による調整解析法も、基本的には作成した propensity score が妥当であるかがすべてである。作成した PS のチェックポイントは以下である。以下を満たさない PS は解析に用いることはできない。

① 作成した propensity score は、目的とする被験者背景の bias をきちんと減少させるか。
② propensity score 作成に用いた変数（被験者背景）は、治療効果にあたえる bias をすべて説明できるものなのか=strongly ignorable な仮定を満たすといえるのか。

① 作成した propensity score は、目的とする被験者背景の bias をきちんと減少させるか

これが propensity score を用いた解析を行う場合、最初のチェックポイントであり、また、最も重要なチェックポイントである。作成した PS の妥当性を評価する場合、被験者背景ごとの治療群間の effect size を PS 調整前後で比較するのが最も簡単で重要なチェックポイントとなる。

$effect\ size(j) = (\overline{X}_{1J} - \overline{X}_{0j}) / se(\overline{X}_{1J} - \overline{X}_{0j})$

$\overline{X}_{1J}, \overline{X}_{0j}$: j番目の背景因子の治療群と対照群の平均値

$sese(\overline{X}_{1J} - \overline{X}_{0j})$: 標準誤差

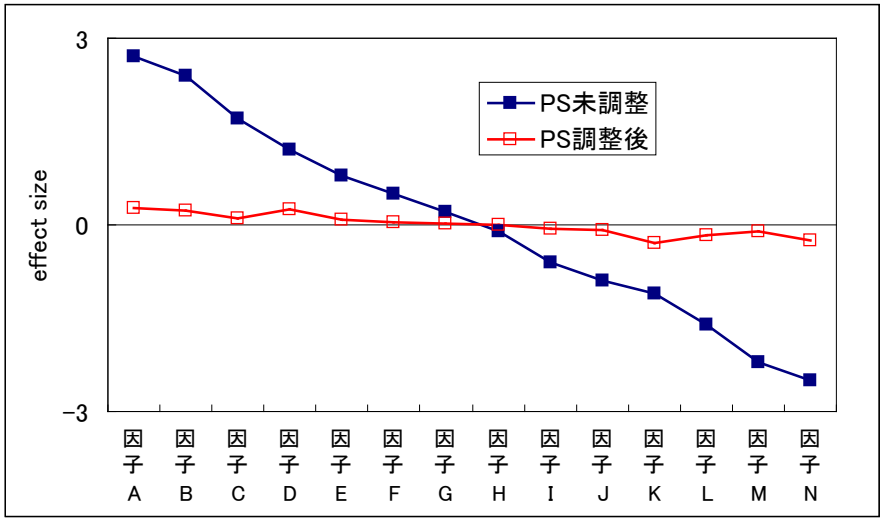
図 1 のように結果を示せばわかりやすい。PS が適切に作成されていれば、図 1 のように、PS で調整した被験者背景の治療群間差の effect size は 0 に近づく。effect size(調整後)/ effect size(未調整)の平均値が、0.1 以下すなわち 90%以上の bias 減少が得られることが望ましい。逆に言うと bias 減少が 70%を超えないような PS は解析に用いるべきではなく、そのような PS を用いた解析結果は意味を持たない。PS 作成に Logistic モデルを用いる場合、すべての被験者背景が正規分布をしていれば、bias 減少はすべての被験者背景で均一となることが理論上知られている。仮に、bias 減少が均一ではなかったり、bias 減少自体あまり起こらなかったりする場合は以下の点に注意する。

- PS 作成に使用した背景因子の分布は偏っていないか（例、一様分布や正規分布からずれていないか）
- カテゴリ背景因子などでは治療群、対照群で度数のないセル（欠測セル）などはないか
- 背景因子間で極端に相関の高い項目が存在しないか、かつ、それら項目の分布が偏っていないか。

上記の問題点をクリアするためには、以下のような対処方法がある。

- 被験者背景の分布を正規分布に近づける変数変換を行う。
- 3 以上の水準を持つカテゴリ変数は、連続量として取り扱えないか検討する（欠測セルをなくすため）
- PS 作成モデル中、例えば p 値が 0.8 以上の値の背景項目は PS モデルから除外する。
- その他

図 1　PS 調整前後の被験者背景の治療群間差の effect size

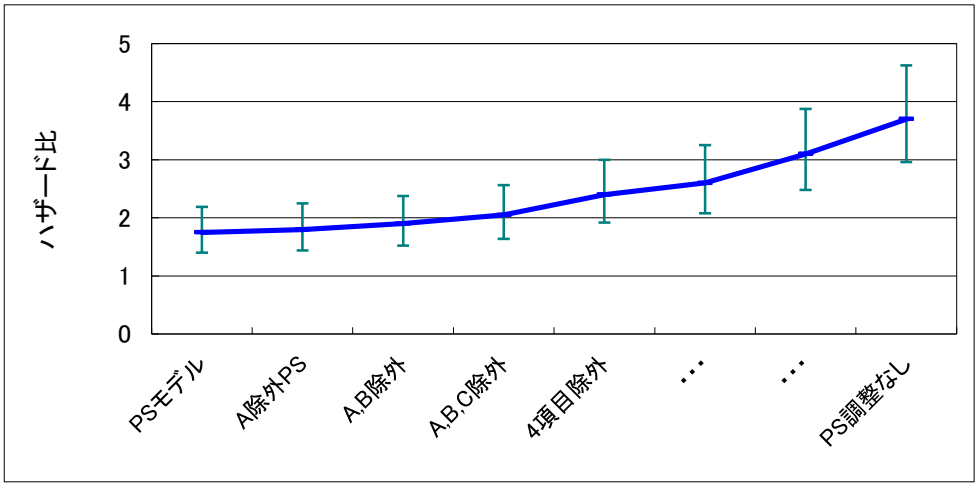


②　propensity score 作成に用いた変数（被験者背景）は、治療効果に与える bias をすべて説明できるものなのか

これは、strongly ignorable な仮定に関する最も重要な観点であるが、現実的には、未観察の背景項目の影響は調べようがない。この確認のためにいくつもの提案があり、それは感度分析から確認される。1 つは、例えば、未測定項目が一定の割合で、IPTW あるいは PS による調整解析の結果に影響を与えているとしたら、現在の結果がそれによってどのくらい説明できるかを確認する方法である [7]。また、IPTW あるいは PS による調整解析の結果と、未調整の結果が大きく違う場合は、筆者が良く行う方法は、PS 作成に使用した被験者背景の変数群から治療群と関連の強い変数から順次 PS 計算モデルから除外し、最終的に得られるハザード比

の変化を確認する方法である。もし作成した propensity score が安定的なものであれば、例え、治療群と関連の強い背景項目を PS 作成モデルから除外しても、PS 解析後のハザード比はあまり変化しない（図 2）。

図 2　PS 作成モデルを構成する背景変数を治療群に関して有意なものから順次除外した場合の、PS 調整解析の結果。



上記例では、PS 解析で調整しない場合、対照群の治療群に対するイベント発生ハザードは 3.7(3.0-4.6)であるが、PS 調整後は 1.8(1.4-2.2)である。PS 調整によりハザード比は 1.0 に近づいているが、PS モデル近辺ではハザードの変化はほとんどないので、背景因子の偏り（観察変数、未観察変数）を考慮しても治療群は対照群より治療効果が高いと予想される。

propensity score 作成に用いた変数（被験者背景）は、治療効果に与える bias をすべて説明できるものなのかは、未観察の説明変数の問題ばかりではなく、観察された変数でも問題となることが多い。例えば、疾患の重症度のような治療効果に明らかに影響を与える背景変数が、治療群では重症と軽症が 50%ぐらいであるが、対照群では重症がほとんどいないということはよくある。このような場合、PS 作成モデルには重症度変数を含めることができなくなる。その場合、PS 解析の結果に関しては、当然、被験者背景の bias 減少は不十分である。このような場合、PS 解析の結果から、重症度のイベントに対するハザード比が一定の割合で存在すると仮定し、重症度の影響を仮に除外した場合のハザード比を検討するぐらいしか PS 解析の結果の妥当性を確認する方法はない。このような場合、PS 解析自体、無意味である場合が多い。

IPTW 法と PS による調整解析法の使い分け

教科書によれば、IPTW は調整解析法と比較して以下のような利点があることが記載されている [8)]。

- Direct に統計計量が計算できる。
- 調整解析法のように PS の治療群、対照群の分布の重なりがあまりなくても計算できる。

一見すると、IPTW 法と PS による調整解析法を選択する場合、常に IPTW を選択すれば問題がないように思える。しかし、以下のような問題がある。

- 例えば、作成した PS の治療群、対照群の分布の重なりがあまりない場合、応答に影響を与える背景因子の分布が、治療群、対照群のどちらかで極端な場合がある。例えば、先に例にあげた重症度のような例

では、対照群の重症患者が極端に少ない場合、対照群の応答に対する重みは、この極端に少ない被験者が非常に大きなものを持つことになる。そのような重みを持った平均値が信頼できるだろうか？PS の治療群、対照群の分布の重なりがあったとしても、このような問題は散見する。

- そもそも、背景因子から計算した治療群、対照群への割付確率が、確率としてどのような意味を持つのであろうか。PS を作成する例えばロジスティックモデルにおいても線形性の制約から、確率として意味を持つのは説明変数（被験者背景）の重心周りであり、重心から遠く、重みとしては重くなる重心から遠い点の確率としての信頼性は低い。ゆえに、PS による調整解析法においてもロジット値（確率値）をそのまま用いるのではなく、層化変数として解析に用いている。

以上のことを勘案すると、IPTW 法と PS による調整解析法の使い分けは以下のようなことが推奨される。

1. 無作為化比較臨床試験のデータにおいて、無作為化割付群以外の要因のハザード比を検討するような場合、あるいは無作為化割付の事後的な偏りを補正する場合、被験者背景の分布は比較群間で極端な偏りが存在する場合は少ない。このように、調査する治療群間の被験者背景の分布があまり治療群間で偏りがない場合は、IPTW の方がスマートだと思われる。
2. しかし、多くの観察研究の場合、治療群間の被験者背景の分布が許容をこえる偏りが存在する場合が多い。このような場合、被験者背景の偏りをある程度調整できるのは、イベントを持つ被験者の被験者背景の重心周りの限られた範囲であり、PS による調整解析法の方が misleading な結論を引気起こすことが少ない。
3. ただ、いずれの方法を用いても、真の平均因果効果を推定しているわけではなく、あくまでも、得たデータの制約のもとで、1 つの推定値を得ているだけである。ゆえに、どのような手法も感度分析の 1 つであり、その妥当性の確認もまた感度分析に依存する。

参考文献

1) Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997; 127:757-63.
2) 「統計的因果推論」　宮川雅巳　朝倉書店　2004
3) [不完全データ解析の基礎と統計的因果推論」　狩野裕 2010：10-13　統計数理研究所　夏期講座
4) Cochran, W. The planning of observational studies of human populations. Journal of the Royal Statistical Society, Series A 1965; 128:234-255.
5) Cochran, W. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968; 24:205-213
6) Rosenbaum, P. and D. Rubin. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 1984; 79:516-524
7) Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome, P. R. Rosenbaum; D. B. Rubin Journal of the Royal Statistical Society. Series B (Methodological), Vol. 45, No. 2. (1983), pp.212-218
8) Analysis of Observational Health Care Data Using SAS, Douglas E. Faries et al. 2010, SAS