Propensity Score Matching in Accounting Research

Shipman, Jonathan E.　　Swanquist, Quinn T.　　Whited, Robert L.

The Accounting Review (2017), 92 (1), pp. 213–244

概要

会計研究において、傾向スコア・マッチング (propensity score matching, PSM) が平均処置効果 (ATEs) を推定するための一般的な手法として利用されるようになった。本論文では、PSM の有用性と限界について、伝統的な重回帰 (MR) 分析と比較して議論する。さまざまな PSM の設定 (design choice) について検討し、2008〜2014 年に主たる会計雑誌に掲載された、PSM を採用している論文 86 本をレビューする。2008 年には、PSM を採用している論文数は 0 本であったのに対し、2014 年では 26 本と、顕著な増加が確認できる。しかしながら、論文によっては PSM の特性を過大評価しすぎてしまい、重要な設定およびまたは理論的に一貫した手続きに則って PSM を実施することを明記していない。そこで、はじめに、会計研究における 3 つの例から、PSM の複雑性について実証的に明らかにする。まず、処置群を表す変数 (treatment) がバイナリー変数でない場合、PSM を利用することで、効果量が最小限になるようなサブサンプルを対象とした分析になってしまうという例を示す。また、一見問題ないように思われる設定が、サンプルの構成 (sample composition) および ATE の推定に深刻な影響を及ぼすという例も提示する。さいごに、マッチング手法の利用を検討している将来の研究に対して、いくつかの示唆を提供する。

1　Introduction

■研究の背景: 内生性の問題に対する従来の手法の問題点

- 実証的会計研究において、因果処置効果 (causal treatment effects) を推定することが、主たる目的とされることがある (Gow, Larcker, and Reiss, 2016)。
- 非実験的データを利用した研究では、処置群が無作為割り当てではない (non-random treatment assignment) ために、内生性の問題が生じてしまう。
- 内生性を緩和させるための従来の手法
  - アーカイバル研究では、伝統的には重回帰 (multiple regression, MR) モデルを利用
  - ただし、MR を利用してバイアスのかかっていない推定値を得るためには、結果変数 (Y) と説明変数 (X) の関係に適切な仕様 (前提) を満たす必要がある。
  - Y と X の関係が前提を満たしていない場合、MR は “回帰式の特定ミス (functional form misspecification, FFM)” の影響を受けることになり、バイアスのかかった推定値を得ることになり得る。

  - FFM による潜在的なバイアスは、処置群 (treatment groups) が似ていないほど、強くなってしまう。

■Propensity score matching (PSM)

- 変数間の特定の仮定を少なくすることで、上記のような問題を緩和
- 処置群から選択された観測値と処置を受ける傾向の推定値を利用して、機械的に、様々な側面から判断したコントロール・グループをマッチ
- 変数間の関係 (functional relation) に関する緩い (relaxed) 仮定のもとで、処置効果を直接的かつ直感的に推定可能
- ただし、PSM は伝統的な MR の手法が持つ理論的有用性を損なう。

■本論文の検討事項

- 内生性、MR の手法、FFM に関する問題、マッチング手法のメリット、PSM の設定のインプリケーションについて議論
- 以下の雑誌に掲載されている、PSM を利用した 86 本の論文に対する議論
  - *The Accounting Review*,*Contemporary Accounting Research*,*Journal of Accounting and Economics*, *Journal of Accounitng Reserach*,*Review of Accounting Studies*
  - PSM を利用した論文数は増加 (2008 年: 0 本、2014 年: 26 本)
  - 各論文について、PSM の利用および設定が正当なものであるのかを評価
- 財務報告に関する研究における 3 つの設定において、PSM を利用した場合に問題が生じる例
  1. 監査人の規模
  2. 内部統制の弱さ
  3. フォローしているアナリスト数
- マッチング手法の利用もしくは、FFM に基づく内生性問題に取り組む他の手法の利用を検討している研究に対するいくつかの提言

■本論文の貢献

- DeFond, Erkens and Zhang (2015) を補完
  - DeFond, Erkens and Zhang (2015)
    * マッチング手法を利用して、監査人の規模および監査の質の関係について分析
    * 設定 (design) をランダムに数千回実行し、総合的な結果を提示
    * 分析結果の多くは、4 大監査法人は、それ以外の監査法人よりも優れた監査の質を提供していることを示唆
    * Lawrence, Minutti-Meza and Zhang (2011) の結果と整合

  - 本論文と DeFond et al. (2015) の違い
    * DeFond et al. (2015) の主たる目的は、4 大監査法人の影響に関する実証的証拠を提供すること
    * 本研究も同様のテーマを共有
    * ただし、会計研究において PSM を利用している論文をレビューし、PSM の有用性および限界に関する議論を提供し、一般的な会計研究における設定 (setting) の、固有の設定 (design choices) に関する影響について明示する、という点で異なる。
    * 本論文におけるどの設定 (settings) においても、実証的証拠を提供するものではない。
- PSM の利用を検討している将来の研究に対する情報提供

■構成

1. Background on propensity score matching: PSM の有用性、誤解、適切なリサーチ・デザインの設計に関する議論
2. Propensity score matching in accounting research: 主な会計雑誌に掲載された PSM を利用した研究のサーベイ
3. Empirical examples of propensity score matching in accounitng settings: 3 つの会計的な例を設定した場合における、PSM が引き起こす問題に関する実例
4. Suggestion and consideration for future research: マッチング手法の利用を検討している研究に対する提言
5. Conclusion: 本論文の発見事項とインプリケーション

2 BACKGROUND ON PROPENSITY SCORE MATCHING

Endogeneity, Functional Form Misspecification, and Propensity Score Matching

- 大学の学位（D_i）が個人の所得（W_i）に与える影響を検討する。
- 実験ならば、大学の学位がある（$D_i = 1$）と、大学の学位がない（$D_i = 0$）を無作為に割り当て、このグループ間で W_i を比較し、平均処置効果（average treatment effect: ATE）を得ることになる。
- この場合、所得の決定要因（たとえば、能力や動機）は、大学の学位を得ることと独立であることが仮定され、内生性の懸念はない（つまり、$E[W_{0i}|D_i = 1] = E[W_{0i}|D_i = 0]$）。
- ここで、実験ではなく、観察的な研究の場合を考え、（1）式を提示する。

$$W_i = \beta_0 + \beta_1 D_i + \varepsilon_i \tag{1}$$

- （1）式で求められた $\hat{\beta}_1$ は、大学の学位が所得に与える影響を十分に説明していない。

- 大学在籍の決定要因（たとえば、家庭の事情、能力、動機、キャリアに対する興味）が、所得も決定している（この要因を X_i とする）
- これらの要因をコントロールしないと、X_i の影響が（1）式の誤差項（ε_i）に入り込んでしまい、$\hat{\beta_1}$ は、バイアスのある推定値となる（つまり、$E[W_{0i}|D_1 = 1] \neq E[W_{0i}|D_i = 0]$）。
- ここで、（1）式の独立変数に X_i を含めると

$$W_i = \beta_0 + \beta_1 D_i + \beta X_i + \varepsilon_i \tag{2}$$

- たとえば、知能（IQ_i）が大学在籍と所得の両方に関連するただひとつの要因だとする。
- IQ_i を（2）式に含めることで、$\hat{\beta_1}$ は、バイアスのない推定値となるはず。
- つまり、X_i が交絡因子（confounds）を十分に捉えているのなら、（2）式は、バイアスのない推定を行う。
- ここで考慮しないといけないのは、W_i と X_i の関係である。
- この関係が不正確（misspecified）に考えられているのなら、“zero conditional mean assumption”（$E[\varepsilon|X_i] = 0$）が守られておらず、（2）式は、バイアスのある推定を行うことになる。
- これは、functional form missepecification（FFM）と呼ばれる内生性の一形態の問題
- マッチングは、FFM の問題を緩和するのに有効である。
- つまり、大学の学位を有する人（$D_i = 1$）と同じ IQ_i（X_i）を有する、大学の学位を有していない人（$D_i = 0$）をマッチングし、処置群と対照群の差異を低下させる。
- こうすれば、W_i と X_i の関係の仮定を置く必要なく、IQ_i（X_i）の影響を調整できる。
- 会計研究のようなアーカイバルデータを用いた研究では、処置群と対照群への割り当ての決定要因は複数存在する。
- Rosendaum and Rubin（1983）は、X_i が複数想定される場合に、処置群に割り当てられる確率を X_i にもとづいて求め、その確率を用いてマッチングを行う手法を提案した。
- その確率のことを「傾向スコア（propensity score）」という。
- 傾向スコアは、以下の（3）式で推定される。

$$D_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3}$$

- 処置群（$D_i = 1$）に属する観測値は、（3）式で推定された傾向スコアが近似している対照群（$D_i = 0$）に属する観測値とマッチングされる。
- つまり、処置群と対照群は、X_i が類似している観測値同士の組み合わせとなる。
- これは、D_i と X_i の相関を最小にし、FFM の懸念を緩和する。

■マッチングの質と外的妥当性

- PSM は、“common supprt”（あるいは overlap）の範囲内に存在する観測値をサンプルとす

ることになる。

- もし、IQ_i が大学の学位取得に強く影響を与えるのなら、サンプルをマッチングしたのち、大学の学位を有して（有してなくて）、かつ IQ_i の高い（低い）観測値の多くを排除する可能性がある。
- つまり、IQ_i と D_i 間の関連性が強まるごとに、マッチングの質の程度が減少する。
- っまｔ、PSM の外的妥当性は、サンプルの ATE と母集団の ATE の近似性で判断されるが、サンプルサイズが小さくなることは、その妥当性に影響を与えるかもしれない（Cram, Karan, and Stuart 2009; Heckman, Ichimura, and Todd 1998）。

■以上の例の重要な仮定

1. モデルには、大学の在籍と所得のいずれにも関連する、すべての要因が含まれている。
2. 共変量（covariates）は、以上の要因から正確に計算できる。

- しかし、実証分析では、要因の特定や測定が困難であることから、以上の仮定が必ず遵守されるわけではない。
- このことは、MR と PSM のいずれにもバイアスがもたらされていることを意味する。
- 以下でこの限界について議論を行う。

Propensity Score Matching in Accounting Research ― Misconceptions and Limitations

■MR に対する PSM の優位性

- MR は、ATE を推定するのに柔軟なフレームワークを提供しており、観察的な研究のスタートポイントとなる。
- しかし近年、MR の頑健性を証明するためや、FFM の問題を緩和するために、PSM を用いる研究が存在している（Armstrong, Ittner, and Larcker 2012b; Armstrong, Jagolinzer, and Larcker 2010; Laerence et al. 2011; Miunutti-Meza 2013）。
- たとえば、クライアント企業の規模と Big 4 の監査法人を選択することには、強い関連性があると考えられる。
- この場合、MR のモデルにクライアント企業の規模を組み込みコントロールしようとするが、それと結果変数との関連が不適切に特定されている場合、Big 4 の監査法人を選択した効果についての推定にバイアスがもたらされる。
- PSM は、企業規模や他の要因を用いてマッチングを行うことで、クライアント企業の規模と監査法人の選択の効果を最小にする。
- この場合、変数同士の関係の仮定を置く必要がなく、PSM は FFM の問題を緩和することになる。
- ここに、MR に対する PSM の優位性がある。

■先行研究の誤解

- しかし、この優位さを認識している先行研究は少ない
- 先行研究は、「内生性」「(自己) 選択バイアス」「欠落変数バイアス」を広く是正する手法として、あるいは、操作変数法（instrumental variables approaches）の代替法として、PSM を用いている。
- MR と同様に、PSM は、自己選択や内生性の懸念に対処するような、処置と結果に関連するすべての要素を正確に特定したり測定したりする能力はない。
- つまり、PSM が Heckman（1979）のタイプの選択モデルの代用ではなく、「内生性」「欠落変数」あるいは「自己選択」に関連する広い懸念を緩和するわけではない。

■PSM は、実験研究を完全に再現するわけではない

1. 観察不能な要因の存在は、観察可能な要因のみでマッチングした場合に、処置への割り当てが完全にランダムになるわけではないことを示している。
    - 実験研究は、処置への割り当てが完全にランダムなので、観察可能な要因「と」観察不能な要因の両方ともコントロールできる。
2. 実験とは異なり、PSM は、観測値を処置に実際に割り当てているわけではない。
    - PSM は、あくまでサンプルの選択や重み付けを行なっている。

- PSM の追加的な議論は、外的妥当性と関連する。限定された重なりを有するセッテングにおいて、PSM は、ATE の推定値がサンプルの外側に一般化しうる程度を妥協する、欠落した反事実のある観測値をシステマティックに除外する。重なりの範囲内でさえも、PSM の発見は、デザインの選択に敏感となりうる。多くの“overlapping”な観測値は、適切な反事実の欠如以外の要因を原因として釣り合いが取れない（go unmatched）かもしれない。具体的には、我々のサンプルにおける観測値のたった 7% が、internal control weaknesses（ICW) を有する。もし我々が ICW の観測値と非 ICW の観測値を 1 対 1 で置き換えもなくマッチングするなら（これは会計研究で普通)、マッチングされたサンプルは、多くの非 ICW の観測値を除外するだろう。さらにいえば、多くの除外された非 ICW の観測値が適切なマッチングであるが、わずかにより良いマッチングが存在するためにそれは廃棄される。事実、たとえ、より多くの共通の支持があるにしても、1 対 1 でマッチされたサンプルは、すべての非 ICW 観測値のせいぜい 7.5% を含む。ゆえに、外見上問題なく見える特定の変化は、「再サンプル」を認め、テストの再現性を毀損する。したがって、代替的な特定は類似した結果を生むことを保証するそれぞれの研究に義務がある（このコンセプトを描いた DeFond et al.［2015］を見よ)。

■PSM を手法として用いる時の注意

- 以上の問題は、マッチングされたサンプルで分析する妥当性、PSM と MR の結果の相違時の解釈、および母集団の ATE とサンプルの ATE を比較することで判断される結果の一般化に影響を与える。
- 観察的な研究では、分析手法を事後的に（post hoc）選択することが可能なので、PSM を軽々に使ってしまいがちになる。
- そうではなく、基本的には先行研究のリサーチデザインにならい、PSM を用いる場合は、その妥当性を判断した上で利用するべきである。

Primary Design Choices in Propensity Score Matching

- PSM の手法は標準化されてないため、同じデータを用いたとしても、研究ごとにその結果が異なる可能性がある（Angrist and Pischke, 2009, 86）。
- 以下で、PSM の手法について議論する。

Primary Design Choices for Estimating the Propensity Score

1. 処置群と対照群の区別
    - 観測値を処置群と対照群に割り当てる場合、その割り当ての基準に問題が生じる。
    - たとえば、IFRS を適用している企業とそうでない企業のような、はっきりと 2 つに分けられるものについては、割り当ての際に問題が生じないだろう。
    - 一方、監査法人の規模、アナリストのカバレッジ、あるいは経営者予想のような場合、処置群と対照群の分かれ目であるカットオフの時点の選択の際に問題が生じる。
    - また、連続的な要因で割り当てを行う場合、マッチングは、カットオフの付近にある観測値同士で行われる可能性が高まる。
    - この場合、第二種の過誤の可能性が高まってしまう。
2. 交絡因子の特定
    - 交絡因子（confounding factors）の選択は、サンプルの構成や結果に影響を与える。
    - ここで、交絡因子は、理論的な裏付けによって選択されるべきである。
    - モデルの適合度（fit）や予測力（predicive power）にもとづくべきというのは誤解である（Peel and Makepeace,2012）。
    - また、PSM は、MR と同様の変数を用いるべきであり、当該変数が理論によって MR に導入するべきでないとするなら、PSM にも導入するべきではない。
    - MR と PSM で変数が異なることは、内的な一貫性がないことや事後的なリサーチデザインの選択という点から批判を招く恐れがある。

Primary Design Choices for Forming the Matched Sample

1. マッチングの置換を認めるか認めないか
    - マッチングの置換を認めないとは、マッチングを 1 回のみ行うことである。
    - この場合、ある対照群の観測値が、処置群の観測値いくつかとマッチングすることが最良としても、そのうちのひとつの処置群の観測値としかマッチングしないことになってしまう。
    - そうなると、マッチングの質が低くなったり、置換を認めている場合よりサンプルサイズが小さくなったりする。
    - マッチングの置換を認める場合、ある対照群の観測値は、複数の処置群の観測値とマッチングされる。
    - 1 回だけマッチングを行う場合よりも、バイアスが低減されたり、サンプルサイズが大きくなったりする可能性がある。
    - ATE の推定の際、反復のマッチングは、マッチングの回数を反映して適切に重み付けされなければならず、標準誤差は、反復のマッチングの程度によって調整されなければならない（Armstrong et al. 2010; Stuart 2010）。置き換えのあるマッチングの際の有意さの議論は、過度の傾向スコアを有する置き換えられた観測値が多くの回数マッチングされ、それゆえに大きく重み付けされる可能性がたいていより高いということである。もし、傾向スコアの外れ値を有する観測値が代替できないのなら、この問題は、欠陥のある結果をもたらしうる。最後に、置き換えのあるマッチングの際、研究者は、どのグループが対照群として設計され、置き換えられるかもしれないのかについて考慮しなければならない（つまり、より大きいグループか、より小さいグループか）。これは、サンプルの構成に対する有意な効果がありうるからである。
2. キャリパー距離（Caliper Distance）
    - 処置群と対照群がマッチングする傾向スコアの距離を制限する。
    - こうすることで、傾向スコアの距離が過度に大きいのにマッチングされてしまい、マッチングの質が低下する問題を防ぐことができる。
3.「1 対 1」か「1 対多」か
    - 会計研究では、処置群と対照群の観測値を 1 対 1 でマッチングするのが主流である。
    - ただし、“common support”の部分に対照群の観測値が処置群の観測値よりも多く存在する場合、1 対多のマッチングの方が有効となるかもしれない。
    - この部分に、処置群の観測値の反事実（counterfactual）となりうる、対照群の観測値が多く存在すると考えられるから。
    -「1 対多」のマッチングは、マッチングの質の低下を招く恐れがある。
    - しかし、本稿と DeFond et al.（2015）で指摘するように、サンプルの変動という問題を部分的に緩和するかもしれない。
    - また、置換を認めたマッチングと同様に、1 対多のマッチングを行なった場合、観測値

を重み付けするべきである。

Evaluating Matched Sample

1. • PSM は、処置群と対照群の共変量を「バランスさせる」。
   • しかし、PSM は、常に完璧なマッチングを行うとは限らない（特に、割り当てが連続変数による場合）。
   • したがって、マッチングに対する評価を行う必要がある。
   • 一般的には、処置群と対照群の共変量の平均値や中央値を検定することで、バランスしているかどうかを評価する。
   • しかし、この差異が非有意であろうとも、FFM によるバイアスを除去できているとは限らない。
   • 一方、この差異が有意であったとしても、マッチングしていない時よりは差異は小さいだろうし、FFM のバイアスも緩和されている可能性はある。
   • 共変量のバランスを評価は、その差異の検定「および」仮設検定のインパクトに依存することとなる。

Estimating the Treatment Effect

1. *t* 検定か MR か
   • マッチングののち、ATE は、単純に *t* 検定、もしくは MR により評価されることになる。
   • 共変量のバランスが達成されているのなら、*t* 検定で妥当かもしれない。
   • しかし、そうでないのなら、処置群と対照群に残っている共変量の相違をコントロールするために、MR を用いることが推奨される（Ho et al. 2007; Lawrence et al. 2011）。

3 PROPENSITY SCORE MATCHING IN ACCOUNTING RESEARCH

■会計研究における PSM の利用 (Table 1 Panel A)

- 2008 ~ 2014 年における、*The Accounting Review, Contemporary Accounting Research, Journal of Accounting and Economics, Journal of Accounting Research, and Review of Accounting Studies* に掲載された論文延べ 86 件が対象。
- 会計研究で PSM が用いられはじめたのは最近 (86 件中 70 件は 2012 ~ 2014 の期間に刊行)。

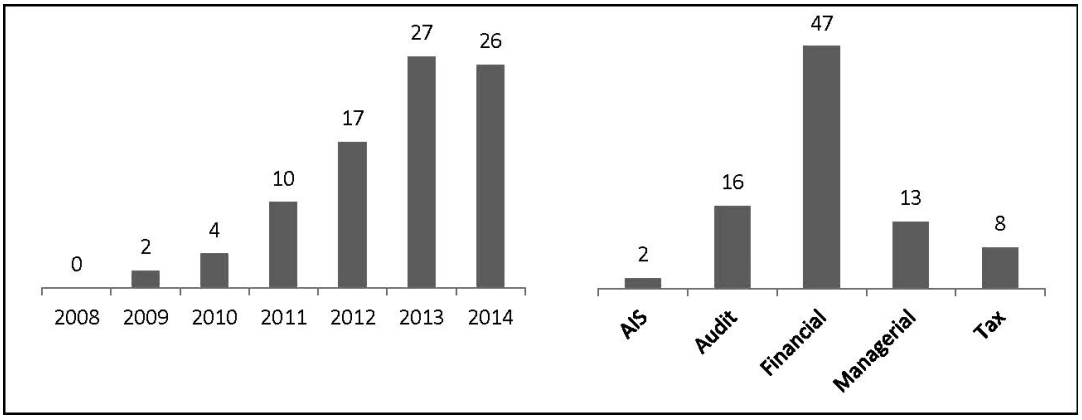
■各研究の PSM の位置付け (Table 1 Panel B)

- 主要な分析 (primary analyses) として用いている研究が 37 件であるのに対し、ロバスト・チェック (sensitivity or robustness tests) として用いている研究は 49 件。

TABLE 1

Descriptive Statistics for Accounting Studies Using PSM

Panel A: Number of Studies in Top Accounting Journals Using PSM Techniques by Year and Topic[a] (2008–2014)



[a] Topics are classified using the BYU Accounting Research taxonomy (Coyne et al. 2010). Twenty-three studies had more than one BYU topic classification (the majority of which include "financial"). Based on judgement, we placed each study into just one classification.

Panel B: Purpose and Reliance on PSM in Empirical Tests (2008–2014)

Is PSM used as a primary or sensitivity analysis?	Primary	Sensitivity
	37	49
If used as a primary analysis, is PSM the only method for at least one conclusion?	Yes	No
	22	15
Is PSM motivated by concerns about FFM or nonlinearities?	Yes	No
	20	66
Did the paper test for FFM or nonlinearities?	Yes	No
	2	84
Is PSM motivated by generic concerns about "self-selection," "endogeneity," or "omitted variable" bias?	Yes	No
	33	53

Panel C: Implementation of PSM (2008–2014)

Was the underlying treatment construct dichotomous?	Yes	No	
	52	34	
For the 64 studies that used other non-PSM tests, were the matching/control variables consistent with other tests?	Yes	No	Unknown
	13	43	8
Does the study match with replacement or without replacement?	With	Without	Unknown
	5	26	55
Did the paper impose a caliper distance?	Yes	No/Unknown	
	29	57	
Was the matching procedure 1:1 or 1:m?	1:1	1:m	Unknown
	68	11	7
Did the paper discuss covariate balance?	Yes	No	
	51	35	
Does the study use MR or a t-test for the second stage?	MR	t-test	Unknown
	58	22	6

Table 1 presents descriptive statistics on the use of PSM in the leading accounting journals from 2008–2014. Studies were identified by searching all publications in *The Accounting Review* (28 studies), *Contemporary Accounting Research* (20 studies), *Journal of Accounting and Economics* (13 studies), *Journal of Accounting Research* (16 studies), and *Review of Accounting Studies* (nine studies) for PSM-related key words (e.g., "propensity," "PSM") and manually determining whether a PSM technique was used. Panel A categorizes the studies by year and topic. Panels B and C classify the studies by motivation and methodology. All studies identified are listed in Appendix A.

- PSM を採用する理由として FFM や重回帰分析の線形性の仮定を挙げている研究はわずか 20 件。
- PSM が対処しうる内生性の問題を提示することなく、広く “自己選択 (self–selection),” “内生性 (endogeneity),” および “欠落変数バイアス (omitted variable bias)” への対応として PSM を用いている研究が 33 件ある。
- Heckman (1979) の代わりとして誤用してしまっている研究も存在する。

■処置群の選択の方法 (Table 1 Panel C)

- 問題の所在
  - 処置が 2 値変数 (dichotomous) であるならば、PSM の実施は単純である。
  - しかしながら、多様な状況でマッチングを実施するため、連続 (あるいは順序) 変数に閾値を設けて変換することがある[*1]。
- 問題点
  - このような場合、閾値付近の観測値が over–represent される傾向があり、それによって、効果の大きさ (および平均処置効果) が消失し、第 II 種の過誤が生じる可能性が増大する。
- 連続変数を用いている研究の数
  - 34 件。また、この影響により、効果の大きさのみならずサンプル・サイズも低下する。
  - 59 (12) 件の研究において、MR のサンプル・サイズの大きさは PSM の 3 (10) 倍である。
  - サンプルサイズが小さいほど、サブサンプルは母集団を代表しなくなる。

■コントロール変数の選択 (Table 1 Panel C)

- MR と PSM のいずれを用いるにせよ、同様のコントロール変数を用いるべきであるにもかかわらず、しばしば異なるコントロール変数が用いられていることがわかった。
- MR からマッチングに用いた変数を除外することは、その変数が処置変数 (treatment) にも結果変数 (outcome) にも影響を与えないこと、ひいては、その変数によるマッチングが不必要であることを意味するに他ならない。
- 分析においては、*post hoc* なモデルの特定 (model specification) をおこなっているという疑念 (appearance; 外観) を避けるため、PSM と他のテストとの説明変数の不一致を検討すべきである。

[*1] 2 値変数を用いた処置群の選択について、例えば修正再表示のアナウンスメントや IFRS のアドプションがあげられる。一方で、非 2 値変数を用いた処置群の選択について、企業の所有構造や監査人の産業特殊性 (auditor industry specialization) があげられる

■傾向スコア推定後のマッチング・プロシージャ (Table 1 Panel C)