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、PSM の特性を過大評価する、重要な設定の選択を開示していない、および/あるいは、PSM の理論的背景を誤って解釈している研究が散見される。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) がバイナリ変数でない場合、PSM を利用することで、効果量が最小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、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)

- 変数間の関係の特定に対する依存 (reliance) を減少させることで、上記のような問題を緩和
- 処置群から推定された傾向 (likelihood; 尤度?) を用いて、様々な側面から、処置群に属する観測値と対

照群をマッチング
- 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観的に推定可能
- ただし、PSM は伝統的な MR の手法と比較して、理論的有用性 (theoretical benefits) が少ない

■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

Introduction

本論文の目的と意義 (担当: 久多里)

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 (担当: 井上)

Propensity score matching in accounting research

主要な会計雑誌に掲載された PSM を利用した研究のサーベイ (担当: 大洲)

Empirical examples of propensity score matching in accounitng settings

3 つの会計的な設定を例として、PSM が引き起こす問題に関する実例 (担当: 井上)

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 (担当: 久多里)

Conclusion

本論文の発見事項とインプリケーション (担当: 大洲)

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、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ことになる。
  - つまり、$E[W_{0i}|D_i = 1] = E[W_{0i}|D_i = 0]$
- ここで、観察的な研究の場合を考え、（1）式を提示する。

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

- ここで、所得は、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。これを X_i）からも影響されることが考えられる。
- 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

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

- 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ことが想定される。
  - 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（$D_i = 1$）を、同程度の IQ_i（X_i）であ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減少させる。
  - 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、マッチングは、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。
- 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; Lawrence et al. 2011; Miunutti-Meza 2013）。
- 例を挙げると、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。
  - しかし、そのコントロール変数と結果変数との関連の特定が不適切である場合、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。

■先行研究の誤解

- 以上の優位さを認識している先行研究は数少ない。
- 先行研究は、“内生性（endogeneity）”“（自己）選択バイアス（(self）selection bias）”“欠落変数バイアス（omitted variable bias）” 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。
- しかし、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。

■マッチングによってサンプルサイズが小さくなる問題

- マッチングは、処置群と対照群が重なり合う（overlapping）する部分で主として生じ、その範囲外のサンプルはマッチングされない可能性が高い。
- その範囲においてさえも、リサーチ・デザインの設定によってマッチングされないサンプルが生じることもある。
  - マッチングでサンプルサイズが小さくなることは、外的妥当性についての問題を有することになる。
- 4 節の分析のサンプルを見ると、内部統制の弱さ（internal control weaknesses: ICW）を監査報告書で指摘されているのは、サンプルのわずか 7% にすぎない。
- この状況において、一般的な会計研究のように 1 対 1 でマッチングを行うと、内部統制の弱さを指摘されていない観測値をサンプルから除外することになる。
- 除外された観測値には、マッチングされても何らおかしくない観測値も含まれているはずである。
- 実際、内部統制の弱さについては、overlapping している部分が大きいにもかかわらず、内部統制の弱さの指摘のない観測値のうち、サンプルとして選択されるのは 7.5% しかない。
- この場合、マッチングごとに選択されるサンプルが変わってしまう可能性があるので、再現性に問題が生じてしまう。
- 研究ごとに同様の結果が確認されることが重要となってくる（DeFond et al.［2015］が指摘している）。

■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。
  - しかし、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、あるいは経営者報酬のような場合、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、このような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。
   - 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. マッチングの置き換えを認めるか認めないか
   - マッチングの置き換えを認めないとは、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。
   - マッチングの置き換えを認める場合、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。
   - 置き換えを認めたマッチングの際、生じたマッチングの数に応じて、ATE や標準誤差に重み付けや調整を行う必要がある（Armstrong et al. 2010; Stuart 2010）。
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。
     - 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。
   - しかし、“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. マッチングの質の評価
   - PSM は、処置群と対照群の共変量を “バランスさせる（balancing）”。
   - 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。
     - 共変量のバランスは、処置群と対照群の共変量の差の大きさ、“および” 検定によって確認される差のインパクトによって評価されるべきである。

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。
   - 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)。

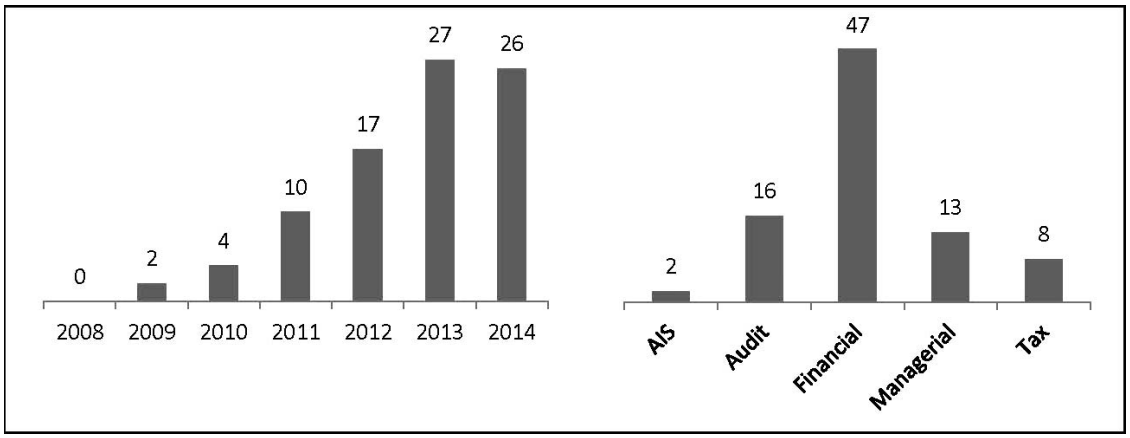
■各研究の PSM の位置付け (Table 1 Panel B)

- 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。
- 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

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.

(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、閾値の近傍の観測値が 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。

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

- 置き換え処理 (replace)
  - 55 件の研究において、マッチングに際して置き換え処理がおこなわれているか否か (matching is performed with or without replacement) 開示されていない。
  - 開示している 31 件の研究のうち、5 件が置き換えあり、26 件が置換なしであった。
- キャリパー距離 (caliper distance)
  - マッチング・プロシージャとしてキャリパー距離を開示している研究は 29 件のみ。
  - 開示されているキャリパー距離の分布は、0.00005 から 0.23 までで、よく用いられている距離は 0.01 (4 件)、0.03 (6 件)、および 0.10 (5 件) である。
- 1 対 1 と 1 対多のどちらのマッチングを用いるか

[*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

  - 1 対 1 (one–to–one) が 68 instances であるのに対し、1 対多 (one–to–many) が 11 instances である。

■covariate balance (Table 1 Panel C)

- マッチング変数の数 (number of matching variables)、キャリパー距離 (caliper distance)、グループ・サイズ (group size) などの要因は、PSM によってサンプルに covariate balance が生じる程度に影響する。
- しかしながら、covariate balance の決定はマッチング・クオリティについての主観的な判断を要求するため、マッチングが実際に適切な (sufficient) balance を達成しているか否か、しばしば不透明である。
- サーベイの結果、35 件はマッチされたサンプル (matched sample) の covariate balance について議論しておらず、4 件のみ傾向スコアの平均差について議論している。
- 研究においては、残存した covariate imbalance の効果を緩和するために、PSM のサブサンプルにおいて MR を使うことができる。
- サーベイの結果、58 件は (第 2 段階の) ATE を推定する際に MR を用いている、22 件は結果変数の距離を t 検定することによって、処置効果を推定している、そして、6 件は推定手法を開示していないということが明らかとなった。

■サーベイの結論

- 本節で指摘した問題点は現在の文献でも改善されていない。
- Heckman (1979) モデルについて含意を示した Lennox et al. (2012, 589) の結論と同様に、われわれは、多数の研究が “重要な計量経済学上の問題点および PSM の利用をとりまく問題点に対する理解 (appreciation) がほとんど無いままに、” PSM を実施していると結論付ける。

4 EMPIRICAL EXAMPLES OF PROPENSITY SCORE MATICHNG IN ACCOUNTING SETTINGS

監査法人の規模、内部統制、アナリストのフォロー数が財務報告の質に与える影響を検証する。

Sample Selection and Descriptive Statistics

■サンプルの選択

- 期間は、Sarbanes-Oxley (SOX) 法にかかる観測値が取得可能な 2004～2012 年である。
- 以下の要件に該当するサンプルを除外する。
  - 海外企業、および金融業（2 桁 SIC コードの 60–69）に属する観測値
  - 総資産が 500 ドル以下の観測値
  - 2 桁 SIC コードに基づく産業年が 10 観測値未満のもの（裁量的会計発生高を計算するための要件）
  - 欠損値のある観測値
- 最終サンプルは、監査法人の規模の分析とアナリストのフォローの分析で 29,227 観測値、内部統制の弱さの分析で 20,385 観測値である。
- 最終サンプルの内訳は、以下のとおりである。

  - Big 4 に監査されている（$BIG4_{it} = 1$）のが 19,988 企業年
  - 少なくとも 1 つの内部統制の弱さを監査報告書で指摘されている（$WEAK_{it} = 1$）のが 1,422 企業年
  - 少なくとも 1 人のアナリストがついている（$ANALYST_{it} = 1$）21,144 企業年

■データソース

- 監査法人、内部統制の弱さ、財務諸表の再報告にかかるデータ：Audit Analytics
- アナリストにかかるデータ：the Institutional Brokers' Estimate System (I/B/E/S)
- 財務データ：Compustat
  - Table 2：記述統計量

Research Design

$BIG4_{it}$、$WEAK_{it}$、$ANALYST_{it}$ のそれぞれを割り当て変数とし、PSM を行い、ATE を推定する。

■予測モデル（第 1 段階）

$$D_{it} = \alpha_0 + \alpha_1 X_{it} + \varepsilon_{it} \tag{4}$$

- D_{it} は、$BIG4_{it}$、$WEAK_{it}$、$ANALYST_{it}$ で、処置群と対照群に割り当てるダミー変数
- PSM のキャリパー距離は、0.03 とする。

■結果モデル（第 2 段階）

$$QUALITY_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 X_{it} + \varepsilon_{it} \tag{5}$$

- $QUALITY_{it}$ は、財務報告の質を示し、本分析では、裁量的会計発生高（$ABSACC_{it}$）、および財務諸表の再報告（$RESTATE_{it}$）のそれぞれで評価している。
- X_{it} は、企業規模（$LNASSETS_{it}$）、パフォーマンス（ROA_{it}、$ATURN_{it}$）、財政状態（$CURR_{it}$、LEV_{it}、$DISTRESS_{it}$）、設立年数（AGE_{it}）、成長性（$GROWTH_{it}$）、企業価値（BTM_{it}）、年度固定効果を用いる。
  - Appendix B：変数の定義
- 本分析で観察したい ATE は、（5）式の β_1 で示される。

Diagnosing Function Form Misspecification

- MR において、FFM が懸念されるのかについて判定する。
- Ramsey（1969）の RESET テストが FFM を見る手法として用いられることがある（Lawrence et al. 2011 を見よ）。
- ただし、これは、変数の非線形性が追加的な説明を与えているのかどうかについて検証するものであり、結果変数と非線形な関係にある変数が ATE の推定にバイアスを与えてるのかどうかを見るものではない。

■変数を追加して FFM を判定する方法（Table 3）

- ここで、MR における FFM を判定する手法として、コントロール変数（たとえば、変数を 2 乗したもの、3 乗したもの）を追加する手法を提唱する。
- もし、元のモデルと変数を追加したモデルの結果が異なるのなら、FFM に対する懸念がある。
- （5）式を推定した結果を見ると、変数を追加したとしても（（2）列と（5）列）、基本的な結果は元のモデル（（1）列と（4）列）大きくは変わらない。
- しかし、監査法人の規模とアナリストのフォローで割り当てた場合、Chow（1960）の検定をすると、元のモデルの ATE と変数を追加したモデルの ATE の間で有意な差が認められる。
  - これは、X と結果変数が非線形な関係にあることを示す。

First-Stage Prediction Model

■第 1 段階（（4）式）の推定結果（Table 4）

- 先行研究では、pseudo-R^2 が高ければ高いほど、PSM は良い状態であることが示唆されている。
- しかし、この説明力は、割り当てが大きく影響を与えている。
- つまり、処置群の X が対照群の X とより大きく相違しているのなら、回帰式の説明力は大きくなる。
- overlap が大きければ大きいほど、pseudo-R^2 は小さくなるとも言える。
  - ただし、説明力が PSM の有効性を必ずしも示しているわけではない。

Demonstration of Propensity Score Overlap

処置群と対照群の傾向スコアの overlap について確認するため、Shaikh, Simonsen, Vydacil, and Yildiz（2009）と同様に、傾向スコアの密度をプロットする。

■監査法人の規模で割り当てた場合の傾向スコアのプロット（Figure 1, Panel A）

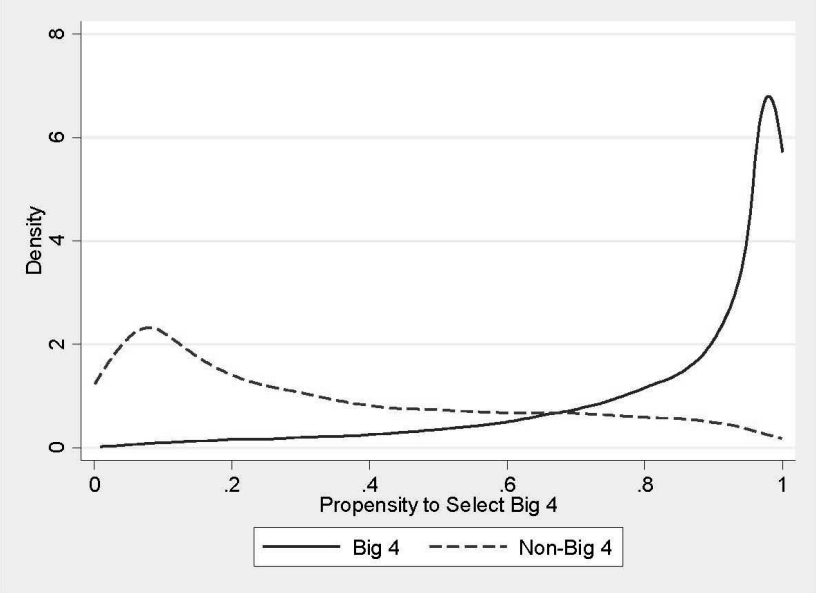
- Big 4 の監査法人のクライアントの傾向スコアは、0.9〜1 の範囲内に集中し、非 Big 4 の監査法人のクライアントの傾向スコアは、0〜0.2 に集中している。
- これは、共変量が割り当ての決定に大きく影響していることを示している。
- 結果的に、マッチングは、主として極端でない傾向スコアの範囲（ここでは、0.2〜0.9）の範囲内でなされることになる。

■Second Tier を考慮して割り当てた場合の傾向スコアのプロット（Figure 1, Panel B）

- 非 Big 4 のクライアントから、Second Tier の監査法人のクライアントを識別する。
- この識別をしたプロットを見ると、Second Tier のクライアントは、0〜0.2 の範囲内で、小規模監査法人のクライアントよりも傾向スコアが小さいが、0.2 以上では小規模監査法人よりも大きな傾向スコアであることがわかる。
- これは、Second Tier のクライアントが Big 4 のクライアントとより多くマッチングが成立することを示す。

FIGURE 1
Auditor Classification Density Plots

Panel A: Distribution of Propensity Scores by Treatment Status: Big 4 and Non-Big 4



Panel B: Distribution of Propensity Scores by Treatment Status: Big 4, Second Tier, and Small Auditors

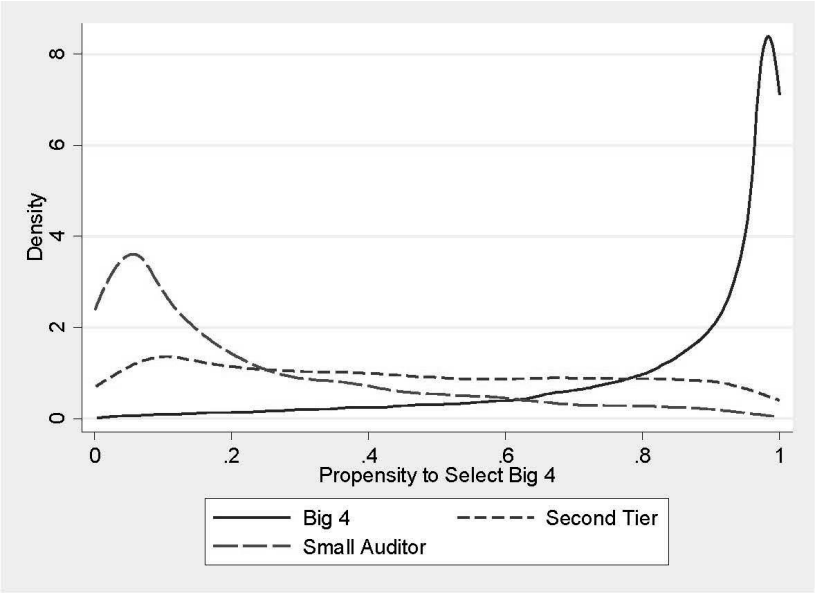
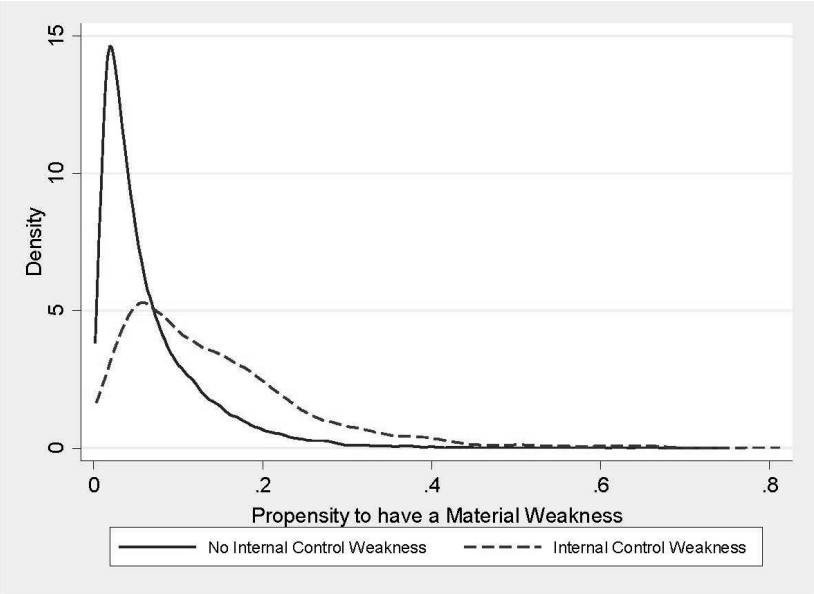
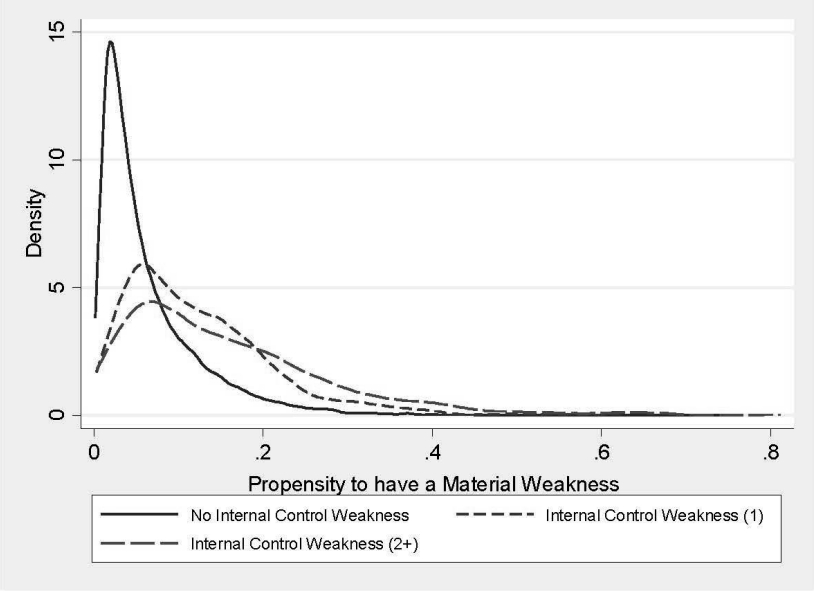


FIGURE 2
Internal Control Weakness Density Plots

Panel A: Distribution of Propensity Scores by Treatment Status: Internal Control Weakness and No Internal Control Weakness



Panel B: Distribution of Propensity Scores by Treatment Status: Internal Control Weakness Count Categories



■内部統制の弱さで割り当てた場合の傾向スコアの密度のプロット（Figure 2, Panel A）

- 監査法人の規模のときよりも、overlap の範囲が大きいことがわかる。
- Table 4 の pseudo-R^2 も、低い値であることもこのことを示唆している。

■内部統制の弱さの判定を詳細にした場合の傾向スコアの密度のプロット（Figure 2, Panel B）

- 内部統制の弱さの判定をより詳細にする（No ICW、1 ICW、2+ICW）。
- 監査法人の規模の場合と同様に、overlap は、No ICW と 2+ICW の間よりも、No ICW と 1 ICW の間の方がより大きいことがわかる。

■アナリストのフォローで割り当てた場合の傾向スコアの密度のプロット（Figure 3, Panel A）

- 監査法人の場合と同様に、処置群と対照群との間で傾向スコアの密度の範囲は異なっており、overlap が限定的なものであることがわかる。

■アナリストのフォロー数を詳細にした場合の傾向スコアの密度のプロット（Figure 3, Panel B）

- アナリストのフォロー数を詳細にみる（1〜5 人、6〜10 人、11 人以上）。
- プロットした結果、アナリストの数が少なければ少ないほど、アナリストに全くフォローされてない観測値との overlap が大きくなることがわかる。
- これは、マッチングがフォローしているアナリストの数の少ない企業となされる可能性が高いことを示している。

Matching without Replacement

置き換えを認めず、1 対 1 でマッチングを行う。

■サンプルの構成（Table 5, Panel A）

- 置き換えをせず 1 対 1 でマッチングすると、かなりの数の観測値を除去することになる。
- 事実、“full sample”を比較して、監査法人の規模で 30%、内部統制の弱さで 14%、アナリストのフォローで 28% のサンプルしか確保できなかった。
- また、マッチングしたサンプルは、Second Tier のクライアントや、アナリストのフォロー数が少ない観測値で構成されていることもわかる。
  - サンプルが小さいことは、外的な妥当性を減少させる、もしくは欠落させてしまう可能性を有している。

■MR と PSM 間における共変量と ATE の比較（Table 5, Panel B）

- 共変量は、PSM を行うことにより、バランスが向上していることがわかる。
  - これは、MR において、処置群と対照群の相違が残ったまま、ATE を推定していることを示す。
- 裁量的会計発生高を従属変数とした場合、MR の ATE はすべて有意であるが、PSM はすべて非有意で

あり、Chow（1960）の検定も、すべての係数間で有意である。
- 一方、財務諸表の再報告を従属変数とした場合は、MR と PSM の ATE ともにすべて類似した結果を示すが、係数を比較すると、有意な差が検出されるものが存在している。

■まとめ

- 以上の分析から、MR と PSM では推定される ATE が有意に異なっていることが明らかとなった。
- もし、PSM のみで分析すると、監査法人の規模、内部統制の弱さ、およびアナリストのフォローは、財務報告の質に影響を与えていないという結果のみを得ることになる。
- しかし、この “帰無仮説を採択する” (accept the null hypothesis) のは結論を急ぎ過ぎ (premature) であるかもしれない。
- 以下で、PSM の手法を変更することで、結果が変化することを示す。

Matching with Replacement

置き換えを認めて、1 対 1 でマッチングを行う。

■サンプルの構成（Table 6, Panel B）

- （2）列目は、サンプルサイズが小さい方の群を処置群とし、サンプルサイズの大きな方の群にマッチングの置き換えを認める場合の結果である（Panel A）。
- サンプルサイズが大きくなっており、キャリパー距離（0.03）の間でより多くの対照群の観測値がマッチングしていることがわかる。
- （3）列目は、サンプルサイズの大きな方の群を処置群とし、サンプルサイズの小さい方の群にマッチングの置き換えを認める場合の結果である。
- サンプルサイズがかなり大きくなっており、ウェイトがいずれも 5 以上であることから、対照群の観測値がより多くの回数マッチングしていることがわかる。
- マッチングの手法ごとにサンプルの構成を比較した結果、9 個の比較中 8 個で有意な差が認められる。
- 特に、Second Tier のクライアントの割合と、アナリストのフォロー数の平均値は、マッチングの手法間で顕著な差がある。
  - マッチング手法の違いが、サンプルサイズとサンプルの構成に影響を与えてしまうことが示唆されている。

■ATE の比較（Table 6, Panel B, C）

- 監査法人の規模で割り当てた場合の結果は、いずれも非有意である。
- しかし、ATE の差が有意に検出されているものが存在する（6 個の比較中 3 個）。
- 内部統制の弱さ、およびアナリストのフォローで割り当てた場合の結果は、マッチングの手法ごとに異なっており、ATE の差も有意に検出されているものがある（12 個の比較中 7 個）。
  - マッチングの手法が統計的な結果に影響を与えてしまう。

The Influence of Matching Variables on Estimates of the ATE

■変数を置き換えてマッチングを行なった場合のサンプルの構成（Table 7, Panel A）

- 企業規模を示す変数を、総資産の自然対数（$LNASSETS_{it}$）から時価総額の自然対数（$LNMARKET_{it}$）に入れ替える。
- マッチングの手法間でサンプルを比較すると、共通の観測値は、58.9〜71.4% の割合である。
- さらに、内部統制の弱さで割り当てた場合の対照群（内部統制の弱さが指摘されていない観測値）は、マッチングの手法間で、わずか 18.2% しかサンプルが共通してない。

■変数を置き換えてマッチングを行なった場合の ATE（Table 7, Panel B）

- $LNASSETS_{it}$ の ATE よりも、$LNMARKET_{it}$ の ATE の方が、有意であるものが多いことがわかる。
- 特に、アナリストのフォローで割り当てとし、財務諸表の再報告を従属変数とした時の ATE は、Table 5 の全サンプルとマッチングサンプルで非有意だったにもかかわらず、変数を置き換えると有意な結果が確認される。
- Chow（1960）の検定によれば、6 個の比較中、3 個で ATE が有意な差であることが確認される。

■変数を追加してマッチングを行った場合の結果（Table 8）

- 以下のとおり、変数を追加してマッチングを行う。
  - 監査法人の規模：営業活動にかかるキャッシュフロー（CFO_{it}）、海外売上高（$FOREIGN_{it}$）
  - 内部統制の弱さ：損失ダミー（$LOSS_{it}$）
  - アナリストのフォロー：$LNMARKET_{it}$
- このマッチングから ATE を求めた結果は、（2）列目と（5）列目に示されているが、いずれも有意であることがわかる。
- さらに以下のとおり変数を追加する。
  - 監査法人の規模：$LNASSETS_{it-1}$、$LOSS_{it}$、棚卸資産（$INVENTORY_{it}$）
  - 内部統制の弱さ：$FOREIGN_{it}$
  - アナリストのフォロー：BTM_{it-1}
- 以上でマッチングし、ATE を求めた結果は、（3）列目と（6）列目に示されているが、これまでの検証で一貫して有意である内部統制の弱さで割り当て、財務諸表の再報告を従属変数とした ATE 以外、すべて非有意である。
  - 以上から、ATE の推定は、PSM のモデルの設定から影響を受けやすいことが明らかである。
- 以上の変数の選択は、結果の頑健性を示すために、“事後的” に選択されたことに留意するべきである。
- すべての X は、結果に同じような影響を与えるとは限らない。
  - ここから、PSM の変数の選択において、十分な考慮をするこの重要さが示唆される。

5 Suggestions and consideration for future research

■本節の検討内容

- 内生性問題に対する “特効薬” は存在しないため、あらゆる方法で慎重に検討する必要がある。
- PSM は、観測不可能な要因を選択する上での内生性を特定するものではないが、FFM に関する内生性については、有用であると考えられる。
- 本節では、PSM を利用する上でより説得力を高めるための提言と、PSM を利用した分析結果を評価する際に検討すべき事項について述べる。

5.1 Suggestions for improved application of propensity score matching

■PSM の利用を検討している研究に対する提言

1. FFM に関する識別問題に取り組むための手段として PSM を利用すべきあり、PSM が “内生性”“自己選択”“欠落変数バイアス” に関する一般的な懸念事項を緩和すると示唆するのは避けるべきである。
2. PSM を利用した単一の (もしくは小数の) の結果だけをもとに、推論を行うべきではない。
   - “マッチング手法は回帰の修正 (regression adjustment) と対立するものではないと認識されるべきであり、実際、その 2 つの手法は補完的であり、組み合わせて使用することが望ましい。” (Stuart, 2010, 2)
3. PSM と MR の両方を利用する場合、PSM のマッチング段階において、MR から除かれる変数を含めるべきではない。
   - MR で除かれる変数についてマッチングを行うことで、MR と PSM が不均一 (unequal) になり、一方もしくは両者の特性 (specifications) に疑問が生じることになる。
   - PSM の第 2 段階において、全てのコントロール変数を含めて MR を推定した場合の処置効果を推定すべき (“二重にロバストな推定” (“doubly robust”estimation))
4. リサーチ・デザインの再現可能性と明瞭性を高めるために、PSM の設定を開示すべきである。具体的な内容は以下の通りである。
   （a）傾向スコアを推定するためのモデル (第 1 段階)
   （b）ATE を推定するためのモデル (第 2 段階)
   （c）マッチング手法によって観測値を置き替えるか否か
   （d）処置済みの各観測値 (treated observation) に対して、マッチさせたコントロールの観測値数
   （e）(実施する場合、) キャリパー距離 (caliper distance)
   （f）マッチングの質 (共変量のバランス (covariate balance))

5.2 Considerations when using propensity score matching

■PSM を利用している場合の検討事項

1. 処置特性 (treatment specification) に関する検討

- 適切なマッチングの数が十分に存在している場合、効果量が最大となるような、母集団から抽出した部分集合に焦点をあてるために、“最大限の” 処置を施された観測値を “最小限の” 処置を施された観測値にマッチさせることを、代替的手法として推奨する。

2. マッチング変数と処置効果の関係に関する検討
   - treatment の選択を決定する性質も、因果効果 (the effect of treatment) に関連している可能性がある。
   - 機械的に、“最大の” 因果効果を持つ観測値を選択し、また、“最小の” 効果を維持させることで、PSM はバイアスのかかった推定 (inference) を実施することになる (Heckman et al., 1998)。
3. 代替的なマッチング設定 (design choiced) が同様の結果を生みだすか否かの検討
   - DeFond et al. (2015) は、PSM の設定を無作為に実施し、証拠が誤りである可能性の評価についても明記している。

5.3 Alternatives to propensity score matching for alleviating functional form misspecification

- 以下の方法を実施することで、common support の範囲外の観測値に関する FFM を緩和することができる一方で、PSM における裁量性と標本分散の一部を排除することができる。
  1. 重複を制限するうえで有効な要因 (例. 企業規模、収益性) を特定し、common support に該当するサンプルを制限する。
  2. 傾向スコアを推定し、そのスコアが、$[\alpha, 1-\alpha]$ (ただし、α は、客観的に決定される cutoff point (例えば、0.10)) の範囲外となるような極端な値をとるサンプルを除く (Crump, Hotz, Imbens and Mitnik 2009)。
- その他にも様々な対処法があるが、どちらか一方を代替するのではなく、FFM を緩和するために補完しあうものである。
- PSM で得た結果の頑健性の確認 (stress testing) こそ重要であり、証拠の妥当性に対する信頼性を向上させる。

6 CONCLUSION

- Shipman, Swanquist, and Whited の分析対象
  - PSM の理論的基礎を議論し、昨今の会計研究における PSM の利用を調査し、そして、デザインの選択 (design choice) にかんする実践的なインプリケーションを例示した。
- 会計研究における PSM の利用 (第 3 節) の要約
  - 観察不能なデータによって生じる内生性を緩和するための Heckman (1979) モデルの代わりとして、誤って利用しているケースがしばしば確認される。
  - デザインの選択を開示していない、あるいは、PSM と MR とでコントロール変数が不一致である研究が散見される。
- デザインの選択 (第 4 節) の要約
  - カットオフ・ポイント (cutoff point) が処置群を決定する場合、カットオフの近傍の観測値は over–represent し、第 II 種の過誤が生じる可能性が増大する。
  - PSM による推定は変化しやすく (fickle) リプリケートが難しいため、マッチド・サンプルに対して

“ストレス・テスト” (stress testing) を実施し、また、代替的なリサーチ・デザインで PSM の追加検証を実施することが要求される。

- 将来研究への提言
  - 重要なデザインの選択を開示すること。
  - PSM の利用目的を適切に理解すること。
  - マッチド・サンプル以外の specification concerns も調査すること。
  - PSM に対して他のリサーチ・デザインをもとに追加分析を実施すること、および、FFM に対処する新たな手法を模索すること。

付録 B　変数の定義

表 1: Variable Descriptions

Variable Name	Variable Definition
$ABSACC_{it}$	全ての産業–年度において最低 10 個の観測値をもとに以下の回帰式を推定し、得られた誤差を業績とマッチさせて算定した裁量的会計発生高 $\frac{TA}{A} = \alpha + \lambda_0 \frac{1}{A} + \lambda_1 \frac{\Delta REV - \Delta REC}{A} + \lambda_2 \frac{PPE}{A}$ A 総資産の平均値 TA 総会計発生高 (= 経常利益 - 営業活動によるキャッシュ・フロー) ΔREV 売上高の変化額 ΔREC 売上債権の変化額 PPE 償却性固定資産 上記の回帰式から得られた各観測値の誤差を、同じ SIC (two-digit) コードで ROA が最も近似している観測値から引いた絶対値を $ABSACC$ としている。
$ANALYST_{it}$	1 人以上のアナリストがフォローしている企業であれば 1、そうでなければ 0 を与えるダミー変数
AGE_{it}	t 期時点で Compustat に収録されている企業年数
$ATURN_{it}$	売上高 ÷ 総資産の平均値
$BIG4_{it}$	監査人が 4 大監査法人であれば 1、そうでなければ 0 を与えるダミー変数
BTM_{it}	簿価時価比率
CFO_{it}	営業活動によるキャッシュ・フロー ÷ 総資産額の平均値
$CURR_{it}$	流動比率
$DISTRESS_{it}$	Altman (1983) に基づいて算定した Z スコア $0.717 \times \frac{\text{運転資本}}{\text{総資産}} + 0.847 \times \frac{\text{留保利益}}{\text{総資産}} + 3.107 \times \frac{\text{利息・税控除前利益}}{\text{総資産}} + 0.42 \times \frac{\text{自己資本簿価}}{\text{総負債}} + 0.998 \times \frac{\text{売上高}}{\text{総資産}}$
$FOREIGN_{it}$	外国為替損益が 0 でなければ 1、そうでなければ 0 を与えるダミー変数

表 1: Variable Descriptions

Variable Name	Variable Definition
$GROWTH_{it}$	$t-1$ 期から t 期への売上高の変化額 ÷ $t-1$ 期の売上高
$INVENTORY_{it}$	棚卸資産 ÷ 総資産
LEV_{it}	長期借入金 (long-term debt) と短期借入金の合計額 ÷ 総資産
$LNASSETS_{it}$	総資産額 (単位: 100 万) の自然対数
$LNMARKET_{it}$	時価総額 (単位: 100 万) の自然対数
$LOSS_{it}$	operating income after depreciation がマイナスであれば 1、そうでなければ 0 を与えるダミー変数
$RESTATE_{it}$	t 期の財務諸表について、後に訂正財務諸表を提出していれば 1、そうでなければ 0 を与えるダミー変数
ROA_{it}	当期純利益 ÷ 総資産の平均値
$WEAK_{it}$	監査報告書において、内部統制に関する脆弱性が 1 点以上指摘されていれば 1、そうでなければ 0 を与えるダミー変数