シングルソースデータにおける
調査・行動データの統合分析モデル

Multiple Overimputation による欠測・測定誤差の補正

本多 将大

慶應義塾大学大学院経済学研究科

統計学会 発表用

本報告の構成

本節の内容：研究背景

シングルソースデータ

単一の対象（顧客）について収集した複数データを統合したデータセット。行動データと調査データを総称して，以下では調査・行動データと呼ぶ。

行動データ

- POS 購買履歴，Web アクセスログなど
- 消費者の実際の行動を自動的・客観的に記録
- 全顧客対象に蓄積されやすい

調査データ

- アンケート，インタビュー，パネル調査
- 満足度，ブランド評価，価値観を把握 (里村卓也 2018)
- 取得コストと回答協力の制約がある

- 統合により，観測された購買行動の背後にある心理的要因を解釈できる (里村卓也 2018; Qian and Xie 2014)。
- 広告接触，顧客マインドセット，心理的経験を統合することで，個人行動や利益予測の精度向上が示されている (Danaher et al. 2020; Venkatesan et al. 2019; Gao et al. 2023)。

1. 欠測

- 調査は一部顧客からしか得られない。
- 回答者だけで分析すると選択バイアスが生じうる (Heckman 1979; 星野崇宏 2007)。
- 非回答者にも態度指標を補完する必要がある。

2. 測定誤差

- 自己申告や心理尺度は，真の値に対するノイズを含む代理変数（noisy proxy）である。
- 線形モデルでは，ランダム（非系統的）誤差でも推定効果量を減衰させ，因果効果の過小推定を招きうる (Imai and Yamamoto 2010)。
- 補完だけでは誤差由来のバイアスが残る。

表 1：本研究で想定するシングルソースデータの構造

	回答者	非回答者
調査データ	観測（測定誤差有）	欠測（MAR）
行動データ	共通の観測変数	

- Multiple Overimputation (MO) は，欠測と測定誤差を同時に扱う方法として提案された (Blackwell, Honaker, and King 2017a)。
- Venkatesan et al. (2019) は，製薬業界で顧客マインドセット指標と処方行動データを MO で統合し，将来利益予測の改善を示した。
- ただし同研究は，製薬業界に限られ，一般消費財の購買データへの適用例ではなく，顧客属性情報も分析に含まれていない。
- 本研究では，小売業の購買ログ，アンケートデータ，デモグラフィック情報を含むシングルソースデータに MO を応用する。
- さらに，シミュレーションにより提案モデルの統計的性質を検証する。

本節の内容：従来手法

定義

Rubin (1976) 以来，欠測は発生メカニズムに基づいて扱う。Little and Rubin (2002) では，Missing Completely At Random (MCAR)，Missing At Random (MAR)，Non-ignorable (NI) / Missing Not At Random (MNAR) に分類される。M を欠測の有無，D_{obs} を観測データ，D_{mis} を未観測データとする。

メカニズム	仮定	主な対応法
MCAR	欠測が観測・未観測データのいずれにも依存しない。$P(M \mid D) = P(M)$	listwise deletion，mean imputation，pairwise deletion
MAR	欠測が観測済みデータ D_{obs} にのみ依存する。$P(M \mid D) = P(M \mid D_{\mathrm{obs}})$	Multiple Imputation (MI) (Rubin 1987)，Inverse Probability Weighting (IPW)，Multiple Overimputation (MO; 本手法)
NI/MNAR	欠測が未観測データ D_{mis} にも依存する。$P(M \mid D) = P(M \mid D_{\mathrm{obs}}, D_{\mathrm{mis}})$	selection model (Heckman 1979)，Pattern Mixture Model (PMM)

本研究での前提

以降では，調査回答の欠測が観測済みの属性・購買履歴に条件づければランダムとみなせる，観測された MAR な欠測を仮定する。

Multiple Imputation (MI) フレームワーク (Rubin, 1987)

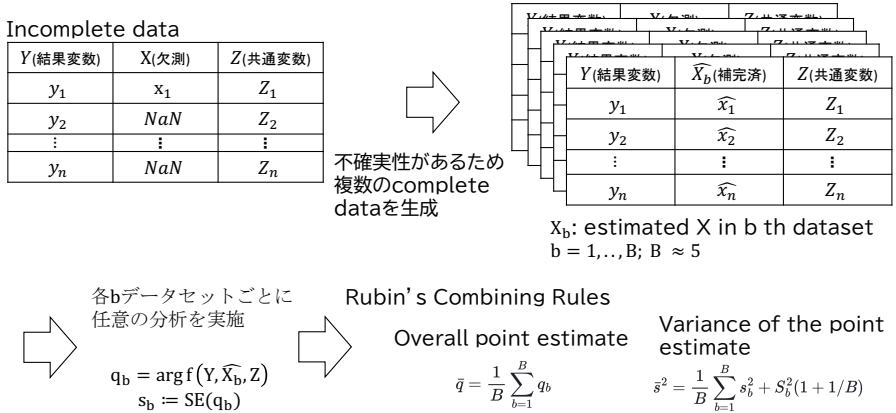

図 1：MI フレームワークの概要

MI は欠測部分を複数回補完し，完全データごとの推定結果を Rubin's Combining Rules で統合する。
ただし，観測済みの値は誤差のない真値として扱われ，測定誤差は補正されない。

従来の測定誤差への対応方法

- 以下では，測定誤差について真の値や結果変数 Y と独立な誤差（non-differential error）を仮定する。
- 独立でない測定誤差（differential error）では，別途誤差のモデリングが必要で，過大推定につながる可能性がある (Imai and Yamamoto 2010)。
- 従来は，操作変数法 (Instrumental Variables; IV)，構造方程式モデリング (Structural Equation Modeling; SEM)，regression calibration，simulation extrapolation などが提案されてきた。

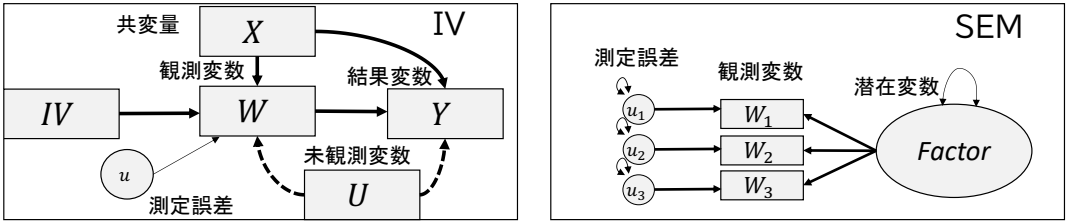

図 2：従来の測定誤差への対応手法

- IV では，測定誤差を含む変数と相関し，かつ測定誤差とは独立である操作変数の発見が必要であり，実務上困難である。
- SEM は測定誤差に柔軟に対応できるが，欠測の問題を同時に解決する枠組みとしては制約が残る。

本手法 (MO) では MI に SEM のアイデアを取り入れることで，測定誤差と欠測の両方の対応を目指す

本節の内容：Multiple Overimputation（MO）の概要と理論的構造

MO の定義

Multiple Overimputation (MO) は，欠測を測定誤差分散が極限的に大きい場合として位置づけ，Multiple Imputation (MI) を測定誤差を含む観測値へ拡張する補完枠組みである (Blackwell, Honaker, and King 2017a)。

MI との違い：観測値も上書き（“over”write）して補完（impute）する = “Over”imputation

観測値・潜在変数・誤差

$$w_i = x_i^* + u_i, \qquad u_i \mid x_i^* \sim \mathcal{N}(0, \sigma_u^2)$$

- w_i: 観測された調査データ
- x_i^*: 真の潜在変数
- u_i: 測定誤差

統計的役割

- 観測値 w_i は，潜在変数 x_i^* の測定誤差を含む観測 proxy として扱う。
- w_i は，x_i^* に対する観測単位別の informative prior を規定する。
- 観測 proxy の情報量は，分析者が指定する誤差割合パラメータ ρ により調整する。

MO | 基本仮定

以下の仮定のもとで，欠測と測定誤差を統一的に扱う (Blackwell, Honaker, and King 2017a; Blackwell, Honaker, and King 2017b)。

1. w_i の測定誤差メカニズムはランダム（Ignorable Measurement Mechanism Assignment; IMMA）である。
2. w_i は潜在変数 $x_i^* \in X^*$ と測定誤差 u_i の線形結合により表される。
3. x_i^* の欠測メカニズムは MAR に基づく。
4. x_i^* に条件づけた u_i は正規分布に従う。

特に仮定 2 と 4 が成り立つとき，w_i は以下の条件付き分布を用いて表される。

$$
\begin{aligned}
& w_i = x_i^* + u_i, \quad u_i \mid x_i^* \sim \mathcal{N}(0, \sigma_u^2) \\
\Leftrightarrow \quad & w_i \mid x_i^* \sim \mathcal{N}(x_i^*, \sigma_u^2) \\
\Leftrightarrow \quad & x_i^* \mid w_i \sim \mathcal{N}(w_i, \sigma_u^2).
\end{aligned} \tag{3}
$$

統計的双対性

ある二変数関数 $f(a, b)$ が，a を b に条件づけた密度 $p(a \mid b)$ としても，b を a に条件づけた密度 $p(b \mid a)$ としても表現できるとき，両者は統計的に双対である (Blackwell, Honaker, and King 2017b)。

$$f(a, b) = p(a \mid b) = p(b \mid a).$$

本研究では，正規分布の双対性を測定誤差モデルに適用する。

$$w_i \mid x_i^* \sim \mathcal{N}(x_i^*, \sigma_u^2) \quad \Longleftrightarrow \quad x_i^* \mid w_i \sim \mathcal{N}(w_i, \sigma_u^2).$$

MO | 経験的事前分布としての解釈

この操作は，観測変数 w_i に基づくデータ生成過程を，未観測な潜在変数 x_i^* に対する事前分布へ読み替えることを意味する。

- 外生的な専門知識や主観的信念ではなく，観測されたデータそのものに基づく経験的事前分布である (Blackwell, Honaker, and King 2017a)。
- w_i は測定誤差を含むが，x_i^* の中心位置について情報を持つ。
- σ_u^2 が小さいほど事前分布は狭く，w_i の情報量は大きい。
- σ_u^2 が大きいほど事前分布は広く，欠測に近い扱いとなる。

したがって，$\mathcal{N}(x_i^* \mid w_i, \sigma_u^2)$ は，測定誤差の大きさに応じて観測値 w_i の情報量を反映した observation-level prior として機能する。

観測値は潜在変数に関する情報を持つ

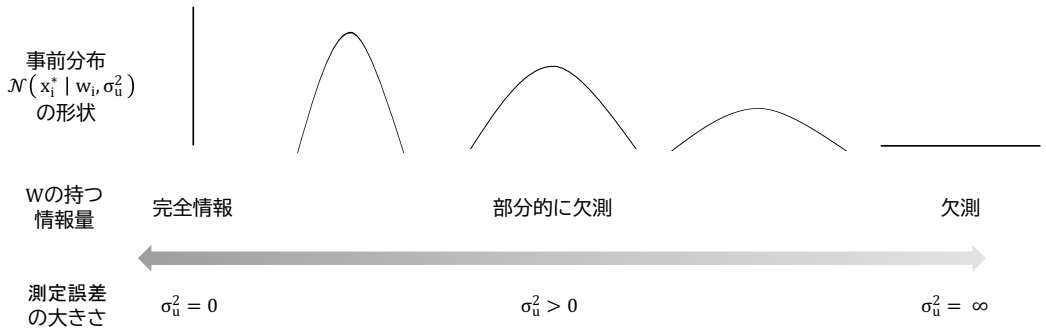

図 3：Multiple Overimputation の概念

観測 proxy w_i の情報量は測定誤差分散 σ_u^2 によって連続的に変化する (Blackwell, Honaker, and King 2017a)。

MO の補正プロセス

Incomplete with measurement error

Y(結果変数)	W(観測変数)	Z(共通変数)
y_1	w_1	Z_1
y_2	NaN	Z_2
⋮	⋮	⋮
y_n	NaN	Z_n

MO

$\rho = \sigma_u^2/\sigma_w^2, 0 < \rho < 1$ として定義し、ρ を用いて測定誤差の強さを調整して補完

事前分布：

$$\mathcal{N}(x_i^* \mid w_i, \sigma_u^2) \Leftrightarrow \mathcal{N}(x_i^* \mid w_i, \rho \cdot \sigma_w^2)$$

σ_w^2：w の母分散（回答のばらつき）
$\rho \in (0,1)$：測定誤差の強さを表す係数

複数のcomplete dataを生成する
$b = 1,..,B;$ B ≈ 10~25

Y(結果変数)	$\widehat{X_b}$(潜在変数)	Z(共通変数)
y_1	x_1^*	Z_1
y_2	x_2^*	Z_2
⋮	⋮	⋮
y_{n-1}	x_{n-1}^*	Z_{n-1}
y_n	x_n^*	Z_n

図 4：MO の補正プロセス

MI との違い

1. 測定誤差を考慮した欠測補完　2. 観測値 w_i も潜在変数 x_i^* で補正

- ρ は，観測 proxy のノイズの強さを調整する分析者指定パラメータである。
- 分析者には観察された w の分散はわかるが，真の測定誤差 σ_u^2 の程度は不明である。
- 実装上は，測定誤差の分散 σ_u^2 が proxy の分散 σ_w^2 に占める割合として ρ を指定する。

$$\rho = \frac{\sigma_u^2}{\sigma_w^2} = \frac{\sigma_u^2}{\sigma_\epsilon^2 + \sigma_u^2}, \qquad 0 < \rho < 1. \tag{4}$$

- $\sigma_u^2 = 0$ なら完全情報として扱う。
- $\sigma_u^2 \to \infty$ なら欠測に近い状態として扱う。
- ρ は，その中間で proxy をどの程度 noisy とみなすかを調整する。

MO | 事前分布と事後予測分布

ρ を用いて，潜在変数 x_i^* の事前分布を以下のように置く。

$$\mathcal{N}(x_i^* \mid w_i, \sigma_u^2) \Leftrightarrow \mathcal{N}(x_i^* \mid w_i, \rho \cdot \sigma_w^2). \tag{5}$$

観測データに基づく x^* の事後分布は，事前分布と尤度の積として表される (Blackwell, Honaker, and King 2017b)。

$$p(x^* \mid Z, w, \theta) \propto p(x^* \mid Z, w, \theta, \gamma) p(\theta, \gamma \mid Z, w). \tag{6}$$

さらに θ, γ を積分消去した事後予測分布からサンプリングする。

$$x^* \sim p(x^* \mid Z, w) = \int p(x^* \mid Z, w, \theta, \gamma) p(\theta, \gamma \mid Z, w)\, d\theta d\gamma. \tag{7}$$

注：実装では事後分布の直接抽出ではなく，EM with Bootstrapping (EMB) により近似する。

記号	定義・解釈
$w_i \in W$	観測された調査変数。潜在変数 x_i^* の測定誤差を含む観測 proxy。
$x_i^* \in X^*$	未観測の潜在的真値。MO では補完・上書きの対象となる。
u_i	測定誤差。$w_i = x_i^* + u_i$ として表す。
σ_u^2	測定誤差分散。w_i の情報量を規定する。
$\mathcal{N}(x_i^* \mid w_i, \sigma_u^2)$	w_i を中心とする観測単位別の事前分布。
ρ	観測 proxy のノイズの強さを調整する分析者指定パラメータ。

- $\sigma_u^2 = 0$：w_i は完全情報として扱われる。
- $\sigma_u^2 \to \infty$：w_i は欠測に近い情報として扱われる。
- $0 < \sigma_u^2 < \infty$：w_i は部分的な情報を持つ観測値として扱われる。

本節の内容：シミュレーション

1. 研究背景
2. 従来手法
3. Multiple Overimputation（MO）の概要と理論的構造
4. シミュレーション
5. 実証分析
6. 結論

DGP

- 真の潜在変数 x_i^* が結果 y_i に影響する。
- 観測 proxy w_i は測定誤差を含む。
- w_i は共通変数 Z_3 に依存して MAR で欠測する。
- 真の誤差割合は $\rho = 0.7$ とする。

手法

手法	役割
True X	真の x_i^* を使う理想ケース
Z only	共通変数 Z のみで推定
MI	欠測補完のみを実施
MO	欠測補完と測定誤差補正を実施

評価指標

- 予測 MSE
- 係数バイアス

潜在変数と観測 proxy

$$x_i^* = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{8}$$

$$w_i = x_i^* + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_u^2). \tag{9}$$

- 真の誤差割合は $\rho = 0.7$。
- $\sigma_\epsilon^2 = 0.3,\ \sigma_u^2 = 0.7$。

欠測と結果変数

$$\Pr(w_i \text{ missing} \mid Z3) = \begin{cases} 0.30 & Z3_i > \overline{Z3}, \\ 0.20 & \text{otherwise}, \end{cases} \tag{10}$$

$$y_i \sim \mathrm{Beta}(\alpha_i, \beta_i), \tag{11}$$

$$\mathrm{logit}(\mu_i) = \gamma_0 + \gamma_1 x_i^*. \tag{12}$$

評価モデルと指標

$$y_i \sim \mathrm{Beta}(\alpha_i, \beta_i), \quad \mathrm{logit}(\mu_i) = \gamma_0 + \gamma X_i \tag{13}$$

$$\mathrm{MSE}(\widehat{y}) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i^{real})^2, \quad \mathrm{Bias}(\widehat{\gamma}) = \widehat{\gamma} - \gamma^{real}.$$

比較手法

True X	$X_i = x_i^*$
Z only	$X_i = Z_i$
MI	$\widehat{w}_{i,b} = \beta_0 + \beta Z_i + \epsilon_i$
MO	$x_{b,i}^* \mid w_i, Z_i \sim \mathcal{N}(\mu_{w\mid Z_i}, \sigma_u^2)$

Rubin's Combining Rules

$$\overline{y} = \frac{1}{B}\sum_{b=1}^{B}\widehat{y}_b,$$

$$\overline{s}^2 = \frac{1}{B}\sum_{b=1}^{B}s_b^2 + S_b^2(1 + 1/B). \tag{15}$$

シミュレーション | 実験設定

- 各手法でのシミュレーション回数は $S = 100$，サンプルサイズは各回 $N = 1{,}000$ 件。
- データはランダムに train:test $= 8 : 2$ に分割した。
- MO は $\rho = 0.7$ とし，ρ の感度分析については後述する。
- MI と MO での補完データ数は $B = 20$ とした。
- 係数バイアスは表 2 で比較する。
- MI は欠測を補完するが，測定誤差を補正しない。
- MO は True X の方向にバイアスを改善する。

MO は係数バイアスを改善する

- True X は理想的な上限性能を表す。
- MI は欠測を補完するが，測定誤差を残す。
- MO は MI より係数バイアスが True X に近い。

	MSE		Bias	
Method	**Mean**	**SD**	**Mean**	**SD**
True x	0.009	0.003	-0.001	0.051
Z only	0.020	0.007	NaN	NaN
MI	0.031	0.009	-0.360	0.058
MO	0.021	0.007	-0.111	0.083

表 2：シミュレーションにおける各手法の評価指標

MO は MI 比で約 32%MSE を改善

- 推定結果を予測 MSE で比較。
- 方法 1 True X：真の x^* で推定。
- 方法 2 Z only：共通変数 Z のみで推定。
- 方法 3 MI：欠測補完後に推定。
- 方法 4 MO：欠測補完と測定誤差補正後に推定。
- MI は測定誤差を補正しないため MSE が大きい。
- MO は MI 比で約 32%改善。

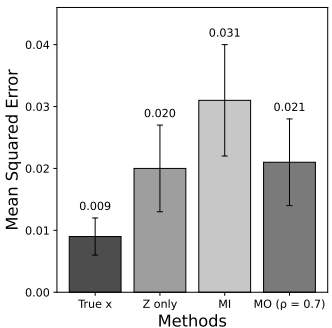

図 5：シミュレーションにおける各手法の MSE

ρ の設定が MO の精度を左右する

- 正確性パラメータ ρ を変化させる。
- $\rho = 0.1$–0.9 を 0.2 刻みで比較。
- 真の誤差構造に近い ρ で MSE が低い。
- ρ の設定は MO の性能を左右する。
- 事前知識がない場合は感度分析が必要。

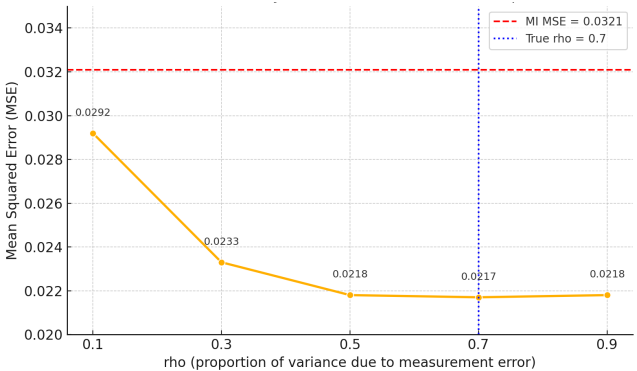

図 6：ρ ごとの MO の MSE

本節の内容：実証分析

実証分析 | データ概要

- データ：Nielsen IQ Household Panel
- 期間：2010/01/01–2012/12/31
- 対象：砂糖（granulated sugar）カテゴリにおける PB 購買
- 有効サンプル：24,771 世帯
- 態度変数回答：13,941 世帯（約 56%）
- 目的変数：2012 年の PB 購買数量
- 2010–2011 年の購買履歴，2011 年の属性，PB に関する調査回答を統合する。
- PB は低価格を特徴とする小売企業の自主企画ブランドである (Olbrich, Jansen, and Hundt 2017; Hara and Matsubayashi 2017)。
- 知覚品質や知覚リスクは購入意図に影響する (Gómez-Suárez, Quiñones, and Yagüe-Guillén 2017)。

実証分析 | 変数定義とデータ統合

記号	定義
N	本研究のサンプルサイズ。
n_{obs}	アンケート調査の回答者数。
n_{mis}	非回答者数。$n_{\mathrm{mis}} = N - n_{\mathrm{resp}}$。
$i = 1, \ldots, N$	世帯の添字。
$W_i^{(m)} \subset W_i$	世帯 i の m 番目の態度指標（$m = 1, \ldots, 4$）。
D_i	デモグラフィック変数。
$Y_{i,t}$	$t = 2010, 2011, 2012$ 時点のカテゴリ購買数量。
$Y_{i,t}^{PB}$	$t = 2010, 2011, 2012$ 時点の PB 購入数量。
$Y_{i,2012}^{PB}$	結果変数。2012 年の PB 購入数量。
$Z_i = (Y_{i,2010}, Y_{i,2010}^{PB}, D_i)$	共通変数。アンケート取得前年の購買履歴とデモグラフィック変数から構成する。

- 作成したシングルソースデータに MO を適用する。
- 全ての共通変数 Z_i が x_i^* の共変量であるとする。
- 観測 proxy のノイズ調整パラメータは $\rho = 0.5$ とおく。

$$\widehat{x}_{i,b}^{*(m)} \mid w_i^{(m)}, Z_i \sim \mathcal{N}\left(\mu_{w^{(m)}|Z_i}, \rho \cdot \sigma^2_{w^{(m)}}\right), \qquad m = 1, \ldots, 4. \tag{17}$$

- $\sigma^2_{w^{(m)}}$ は $w^{(m)}$ の分散を表す。
- 態度指標ごとに測定誤差の異質性を考慮する。
- $B = 20$ 個の補完データを生成する。

- 各補完データで 2012 年の PB 購買数を推定する。
- 説明変数は，前年までの購買数と MO で補正した潜在態度指標。
- 購買数データのため，負の二項分布モデルを用いる。

$$
\begin{aligned}
Y_{i,2012}^{PB} &\sim \mathrm{NegBin}(\mu_i, \kappa), \\
\ln \mu_i &= \gamma_0 + \sum_{t=2010,2011} \left(\gamma_{1,t} Y_{i,t}^{PB} + \gamma_{2,t} Y_{i,t} \right) + \gamma' \widehat{X}_i^*.
\end{aligned} \tag{18}
$$

従来手法との予測精度比較のため，$Y_{i,2012}^{PB}$ の予測値と真の観測値との MSE を算出する。

手法	使う情報	欠測・測定誤差の扱い
Baseline	行動データ＋属性	調査データを使わない。
MCAR	行動データ＋属性＋調査データ	欠測に平均値を代入し，MCAR とみなす。
MI	行動データ＋属性＋調査データ	Z_i で欠測補完するが，観測値の測定誤差は補正しない。
MO	行動データ＋属性＋調査データ	欠測補完に加えて，観測済み回答も潜在変数で上書きする。

Baseline：行動データ＋属性のみ

$$Y_{i,2012}^{PB} \sim \mathrm{NegBin}(\mu_i, \kappa), \quad \ln \mu_i = \gamma_0 + \sum_{t=2010,2011} \left(\gamma_{1,t} Y_{i,t}^{PB} + \gamma_{2,t} Y_{i,t} \right) + \gamma' D_i.$$

MCAR：平均代入で欠測補完

$$Y_{i,2012}^{PB} \sim \mathrm{NegBin}(\mu_i, \kappa), \quad \ln \mu_i = \gamma_0 + \sum_{t=2010,2011} \left(\gamma_{1,t} Y_{i,t}^{PB} + \gamma_{2,t} Y_{i,t} \right) + \gamma' W_i.$$

MI：MI で欠測補完

$w_{i,b} = \beta_0 + \beta Z_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ で補完し，各補完データで

$$Y_{i,2012}^{PB} \sim \mathrm{NegBin}(\mu_i, \kappa), \quad \ln \mu_i = \gamma_0 + \sum_{t=2010,2011} \left(\gamma_{1,t} Y_{i,t}^{PB} + \gamma_{2,t} Y_{i,t} \right) + \gamma' w_{i,b}$$

を推定する。

MO は最も低い予測 MSE を示す

- Baseline：行動データと属性のみ。
- MCAR：調査データを平均代入。
- MI：調査データの欠測を多重代入。
- MO：欠測補完に加えて測定誤差を補正。

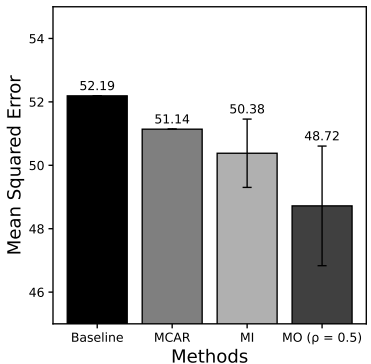

図 7：実証分析における各手法の MSE

PB 購買数量 (2011) と態度指標が PB 購買を説明する

変数	Baseline	MCAR	MI	MO
PB 購買数量 (2011) $Y^{PB}_{i,2011}$	0.042***	0.038***	0.038***	0.039***
	(0.008)	(0.008)	(0.008)	(0.008)
PB への知覚品質	–	−0.081**	−0.054***	−0.460***
		(0.025)	(0.016)	(0.128)
PB への知覚リスク	–	−0.026	−0.036*	0.451***
		(0.026)	(0.015)	(0.138)
PB への知覚価格	–	−0.013	−0.004	0.016
		(0.011)	(0.010)	(0.030)
品質への支払い意思	–	0.020·	0.025*	−0.003
		(0.012)	(0.010)	(0.038)
切片	✓	✓	✓	✓
デモグラフィック変数 D_i	✓	✓	✓	✓
購買履歴変数 Y	✓	✓	✓	✓

- 結果変数は $Y^{PB}_{i,2012}$。アウトカムモデルでは $t = 2010, 2011$ の購買履歴を投入する。
- 共通変数は $Z_i = (Y_{i,2010}, Y^{PB}_{i,2010}, D_i)$ とし，調査回答の補完に用いる。

注：*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.10$。完全な係数表は Appendix に掲載。

本節の内容：結論

結論 | 本研究の結論と意義

- 調査・行動データの統合では，調査データの欠測と測定誤差を同時に考える必要がある。
- MO は，観測された調査回答を潜在変数の事前情報として扱い，欠測補完と測定誤差補正を統一的に実行できる。
- シミュレーションでは，MI より予測精度と係数バイアスが改善した。
- 実証分析では，PB 購買数予測において Baseline，MCAR，MI より MO の MSE が低かった。
- 一般消費財における購買データへの理論的適用と実データでの応用可能性を示した。
- マーケティングの調査・行動データの文脈で MO を整理した。
- デモグラフィック情報を含む統合分析の有用性を示した。

理論的限界と今後の対応は以下である。

限界	今後の対応
測定誤差分散の同定	“gold standard” データと最適な ρ に関する研究の拡充 (Blackwell, Honaker, and King 2017a)。
欠測メカニズム NI への適用	選択モデルや Pattern Mixture Model との組み合わせ (Rubin 1987)。
IMMA ではない測定誤差への適用	階層ベイズモデル，群別 ρ，誤差分散の個人差・系統性のモデリング。
高次元データでの不安定性	変数選択，正則化，多変量正規仮定の妥当性確認。
heteroskedastic な測定誤差分布	仮定 4 の正規性を拡張し，誤差分散の異質性を明示的に扱う。
時系列形式の調査データ	脱落，系列相関，測定誤差を同時に扱う拡張。
マルチソースデータへの応用	シングルソース統合からデータ融合モデルへの発展。

補足：態度指標の欠測は共通変数に条件づけて扱う

- 行動データと属性は完全データとして扱う。
- 調査回答は回答者のみで観測される。
- 非回答世帯では態度指標 W が欠測する。
- 世帯サイズ・世帯年収・合計購買数で欠測率に差が見られる。
- 以降では観測済み共通変数に条件づけた MAR を仮定する。

表 5：実証分析におけるシングルソースデータの構造

	Y_{2012}^{PB}	$Y_t, Y_t^{PB} \subset Z$	$D \subset Z$	W
データ分類	行動データ		調査データ1	調査データ2
3節との対応	結果変数Y	共通変数 Z	共通変数 Z	観測変数 W
		完全データ	完全データ	不完全データ（欠測・測定誤差）
n_{obs}	-	$Y_{t,1}$ ⋮ $Y_{t,n_{\text{obs}}}$	D_1 ⋮ $D_{n_{\text{obs}}}$	W_1 ⋮ $W_{n_{\text{resp}}}$
n_{mis}		$Y_{t,n_{\text{obs}}+1}$ ⋮ $Y_{t,N}$	$D_{n_{\text{obs}}+1}$ ⋮ D_N	-

表 6：共通変数に条件づけた欠測率

	欠測率	
	平均未満	平均以上
年齢	44%	44%
世帯サイズ	41%	49%
世帯年収	45%	43%
合計購買数	43%	45%

補足：本文スライドの注記

- 本報告では unit nonresponse に焦点を当てる。item nonresponse にも MO は理論上適用可能。
- M は欠測インディケータ行列。MCAR でも平均代入は分散・共分散を過小評価しうる。
- IMMA は non-differential error を指す。仮定 4 は heteroskedastic error へ拡張可能だが，本報告では正規・等分散を仮定する。
- Nielsen IQ は約 6 万世帯規模。本分析では購買ログ不完全世帯と非購入世帯を除外し，購買・属性を完全データ化した。
- 推定時には母集団代表性に基づく projection factor で重みづけした。

補足：実証データの概要

表 3：実証分析の変数概要

Symbol	Variable	Mean	SD	Min	Max
D_i	世帯年収	1663.0	2390.9	129	17611
	世帯サイズ	2.4	1.2	1	9
	年齢	57.0	10.5	20	70
$Y_{i,t}$	カテゴリ購買数量(2010)	5.1	7.0	0	207
	カテゴリ購買数量(2011)	5.0	7.0	0	145
	カテゴリ購買数量(2012)	4.5	6.5	0	215
$Y_{i,t}^{PB}$	PB 購買数量(2010)	7.9	8.6	0	215
	PB 購買数量(2011)	8.1	8.7	0	193
	PB 購買数量(2012)	7.7	8.3	0	244
W_i	PBへの知覚品質	1.9	1.4	1	7
	PBへの知覚リスク	1.7	1.2	1	7
	PBへの知覚価格	4.9	1.9	1	7
	品質への支払い意思	4.6	2.0	1	7

- 購買履歴：2010–2012 年の砂糖カテゴリ年間購買数と PB 購買数。
- 属性：年齢，世帯人数，世帯年収。
- 態度変数：PB に関する 4 項目の調査回答。

行動・属性・態度指標を世帯単位で統合し，PB 購買数予測に用いる。

補足：態度指標の測定

表 4：アンケート用紙での質問内容

	Please tell us how much you agree or disagree with each of the following statements about Granulated sugar products. Please use a 7-point scale where 1=Agree and 7=Disagree.
Q1	Store brand products for granulated sugar are just as good as the brand name products.
Q2	Store brand products for granulated sugar are just as safe as the brand name products.
Q3	Store brand products for granulated sugar are just as expensive as the brand name products.
Q4	I am willing to pay extra to make sure I get top-quality granulated sugar products.

- 7 段階リッカート尺度で PB への態度を取得。
- 構成項目：知覚品質，知覚リスク，知覚価格，品質への支払い意思。
- 回答には心理尺度特有の測定誤差が含まれると考える。

態度指標は測定誤差を含む代理変数として扱う必要がある。

注：Tell Us More Survey 回収は 38,878 世帯。本分析の態度変数有効回答は 13,941 世帯で，item nonresponse はなし。

補足：実証分析における完全な推定結果

表 7：実証分析における推定結果

	Baseline	MCAR	MI	MO
(Intercept)	1.728 (0.100) ***	1.866 (0.142) ***	1.755 (0.130) ***	1.712 (0.186) ***
世帯年収	-0.027 (0.003) ***	-0.027 (0.003) ***	-0.027 (0.003) ***	-0.027 (0.003) ***
世帯サイズ	0.087 (0.010) ***	0.086 (0.010) ***	0.084 (0.010) ***	0.077 (0.011) ***
年齢	-0.003 (0.001) *	-0.003 (0.001) *	-0.002 (0.001) *	-0.002 (0.001) ·
カテゴリ購買数量(2010)	-0.004 (0.004)	-0.004 (0.003)	-0.003 (0.003)	-0.001 (0.003)
PB 購買数量(2010)	0.010 (0.006) ·	0.010 (0.006) ·	0.008 (0.006)	0.006 (0.006)
カテゴリ購買数量(2011)	-0.008 (0.005) ·	-0.005 (0.004)	-0.005 (0.004)	-0.006 (0.004)
PB 購買数量(2011)	0.042 (0.008) ***	0.038 (0.008) ***	0.038 (0.008) ***	0.039 (0.008) ***
PBへの知覚品質	—	-0.081 (0.025) **	-0.054 (0.016) ***	-0.460 (0.128) ***
PBへの知覚リスク	—	-0.026 (0.026)	-0.036 (0.015) *	0.451 (0.138) ***
PBへの知覚価格	—	-0.013 (0.011)	-0.004 (0.010)	0.016 (0.030)
品質への支払い意思	—	0.020 (0.012) ·	0.025 (0.010) *	-0.003 (0.038)

注：*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.10$。Baseline は complete-case analysis, MCAR は mean imputation, MI は multiple imputation, MO は multiple overimputation。Baseline には態度変数が含まれないため “–”。

補足：Amelia による MO 実装イメージ

```
prior_sd <- sqrt(rho * var(w_obs, na.rm = TRUE))
row_idx  <- which(!is.na(w_obs))

prior_mat  <- cbind(row_idx, 1, w_obs[row_idx], rep(prior_sd, length(row_idx)))
overimp_mat <- cbind(row_idx, 1)
mo_data <- data.frame(W = w_obs, Z)

mo_imp <- amelia(mo_data, m = m,
                 priors = prior_mat,
                 overimp = overimp_mat)
```

観測済みの W にも事前分布を与え，overimputation 対象として指定する点が通常の MI と異なる (Honaker, King, and Blackwell 2011)。

ご清聴ありがとうございました

参考文献 I

里村卓也 (2018). “トピックモデルによる顧客データの統合的分析”. In: オペレーションズ・リサーチ 63.2, pp. 67–74.

Qian, Y. and H. Xie (2014). “Which Brand Purchasers Are Lost to Counterfeiters? An Application of New Data Fusion Approaches”. In: *Marketing Science* 33.3, pp. 437–448.

Danaher, P. J., T. S. Danaher, M. S. Smith, and R. Loaiza-Maya (2020). “Advertising Effectiveness for Multiple Retailer-Brands in a Multimedia and Multichannel Environment”. In: *Journal of Marketing Research* 57.3, pp. 445–467.

Venkatesan, R., A. Bleier, W. Reinartz, and N. Ravishanker (2019). “Improving Customer Profit Predictions with Customer Mindset Metrics through Multiple Overimputation”. In: *Journal of the Academy of Marketing Science* 47, pp. 771–794.

Gao, L., E. de Haan, I. Melero-Polo, and F. J. Sese (2023). “Winning Your Customers’ Minds and Hearts: Disentangling the Effects of Lock-In and Affective Customer Experience on Retention”. In: *Journal of the Academy of Marketing Science* 51.2, pp. 334–371.

Heckman, J. J. (1979). “Sample Selection Bias as a Specification Error”. In: *Econometrica* 47.1, pp. 153–161.

星野崇宏 (2007). “学習科学研究の妥当性向上へ向けた統計解析法と複数データの統合手法について – 傾向スコアによる共変量調整とデータフュージョン –”. In: 教育システム情報学会誌 24.3, pp. 216–224.

Imai, K. and T. Yamamoto (2010). “Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis”. In: *American Journal of Political Science* 54, pp. 543–560.

参考文献 II

Blackwell, M., J. Honaker, and G. King (2017a). "A Unified Approach to Measurement Error and Missing Data: Overview and Applications". In: *Sociological Methods and Research* 46.3, pp. 303–341.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Blackwell, M., J. Honaker, and G. King (2017b). "A Unified Approach to Measurement Error and Missing Data: Details and Extensions". In: *Sociological Methods and Research* 46.3, pp. 342–369.

Olbrich, R., H. C. Jansen, and M. Hundt (2017). "Effects of Pricing Strategies and Product Quality on Private Label and National Brand Performance". In: *Journal of Retailing and Consumer Services* 34, pp. 294–301.

Hara, R. and N. Matsubayashi (2017). "Premium Store Brand: Product Development Collaboration between Retailers and National Brand Manufacturers". In: *International Journal of Production Economics* 185, pp. 128–138.

Gómez-Suárez, M., M. Quiñones, and M. J. Yagüe-Guillén (2017). "Private Label Research: A Review of Consumer Purchase Decision Models". In: *Advances in National Brand and Private Label Marketing*, pp. 165–172.

Honaker, J., G. King, and M. Blackwell (2011). "Amelia II: A Program for Missing Data". In: *Journal of Statistical Software* 46.7, pp. 1–47.

Gilula, Z., R. E. McCulloch, and P. E. Rossi (2006). "A Direct Approach to Data Fusion". In: *Journal of Marketing Research* 43.1, pp. 73–83.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Wiley.

参考文献 III

Mitsuhiro, M. and T. Hoshino (2020). "Kernel Canonical Correlation Analysis for Data Combination of Multiple-Source Datasets". In: *Japanese Journal of Statistics and Data Science* 3.2, pp. 651–668.

Petersen, J. A., V. Kumar, Y. Polo, and F. J. Sese (2018). "Unlocking the Power of Marketing: Understanding the Links between Customer Mindset Metrics, Behavior, and Profitability". In: *Journal of the Academy of Marketing Science* 46.5, pp. 813–836.

Rubin, D. B. (1976). "Inference and Missing Data". In: *Biometrika* 63, pp. 581–592.

星野崇宏 (2009). 調査観察データの統計科学. 岩波書店.