共同研究集会「官民オープンデータ利活用の動向及び人材育成の取組」
令和３(2021)年11月18日（木）

公的統計ミクロデータの利活用推進に資するSynthetic Dataの作成方法について

高部　勲
立正大学データサイエンス学部

Contents

☞**1．公的統計ミクロデータの概要**

2．疑似データの重要性と課題

3．Synthetic Dataに基づく方法

4．実データに基づく試作

5．まとめと今後の課題

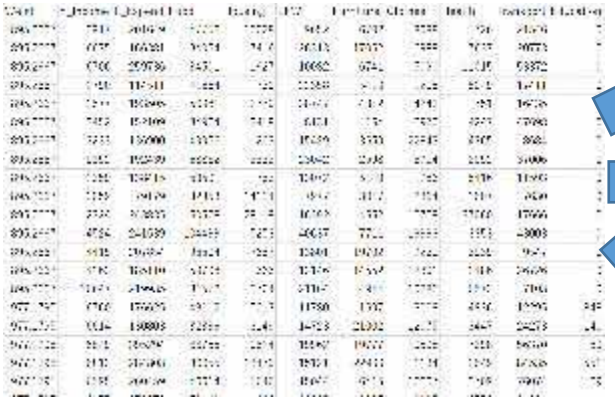
公的統計ミクロデータ（調査票情報）

- 国の行政機関が実施した統計調査の結果について、**調査対象の秘密の保護を図った上**で、**世帯単位や事業所単位といった集計する前の個票形式のデータ（ミクロデータ，調査票情報 ）**を提供
- 　ミクロデータ（調査票情報）を用いることで、研究者の方々は、より自由で多様な分析を行うことが可能となるため、新たな発見につながることが期待

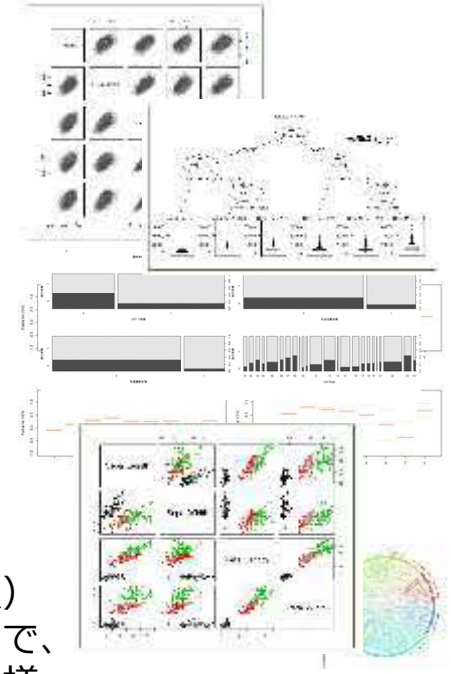
「政府統計の総合窓口（e-Stat）」
https://www.e-stat.go.jp/

行政機関による
集計・公表

ミクロデータ（調査票情報）



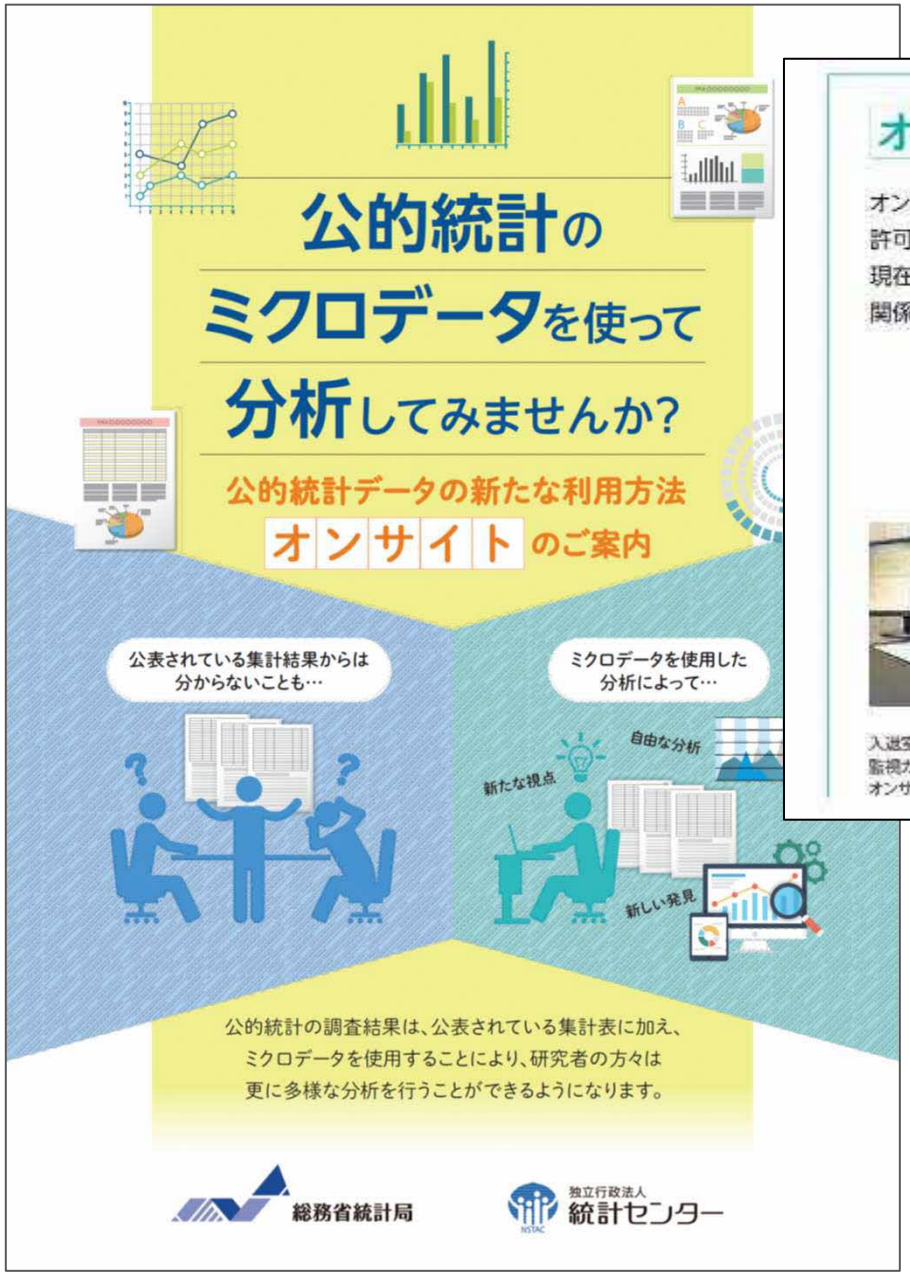
（※上図は擬似データ）



ミクロデータ
（調査票情報）
を用いることで、
より自由・多様
な分析が可能

（ミクロデータ利用ポータルサイト”miripo”から引用
https://www.e-stat.go.jp/microdata/）

公的統計ミクロデータのオンサイト利用



（ミクロデータ利用ポータルサイト”miripo”から引用
https://www.e-stat.go.jp/microdata/）

公的統計ミクロデータ研究コンソーシアム

公的統計ミクロデータ研究コンソーシアム：
ウェブサイト

http://jmodc.org/

１．公的統計ミクロデータの概要

☞**２．疑似データの重要性**

３．Synthetic Dataに基づく方法

４．実データに基づく試作

５．まとめと今後の課題

疑似データの重要性

①プログラムテスト用データとしての活用

- 公的統計ミクロデータの利用申請を行う前にデータの利活用のイメージを把握（※オンサイト利用）
- 各種のアプリケーションの開発・テスト
- 匿名加工情報等の作成におけるニーズの事前把握　など

②データを利用した教育への活用

- 経済・社会の計量モデルは、個人・企業等を単位としているものが多い（重回帰分析など）
  ⇒ 集計データではないレコード単位のデータの重要性
  ⇒ 元データと結果がそれほど変わらないのが望ましい
- 複雑な手続きを必要とせず自由に利用できるデータの重要性

疑似データの重要性（教育への活用）

数理・データサイエンス・AI　リテラシーレベルの教育方法

➢「導入」「基礎」「心得」「選択」のそれぞれの分類ごとに、推奨される具体的な教育方法を以下のとおりまとめた。

導入	**1. 社会におけるデータ・AI利活用** ・データ・AI利活用事例を紹介した動画（MOOC等）を使った**反転学習**を取り入れ、講義ではデータ・AI活用領域の広がりや、技術概要の解説を行うことが望ましい ・学生がデータ・AI利活用事例を調査し発表する**グループワーク**等を行い、一方通行で事例を話すだけの講義にしないことが望ましい	教育方法（例）※ 1，2，3，4 1，4
基礎	**2. データリテラシー** ・各大学・高専の特徴に応じて**適切なテーマ**を設定し、**実データ**（あるいは模擬データ）を用いた講義を行うことが望ましい ・実際に手を動かしてデータを可視化する等、学生自身がデータ利活用プロセスの一部を**体験**できることが望ましい ・必要に応じてデータハンドリングスキルを埋めるためのフォローアップ講義（**補講**等）を準備することが望ましい	教育方法（例）※ 1，2，3，4 1，4
心得	**3. データ・AI利活用における留意事項** ・身近で起こったデータ・AI活用における負の事例を通して、データ駆動型社会のリスクを**自分ごと**として考えさせることが望ましい（必要に応じてMOOC等の活用も検討する） ・データ・AIが引き起こす課題について**グループディスカッション**等を行い、一方通行で事例を話すだけの講義にしないことが望ましい	教育方法（例）※ 1，2，3，4 1，4
選択	**4. オプション** ・本内容は**オプション**扱いとし、大学・高専の特徴に応じて学修内容を選択する ・各大学・高専の特徴に応じて**適切なテーマ**を設定し、**実データ**（あるいは模擬データ）を用いた講義を行うことが望ましい ・学生が希望すれば本内容を受講できるようにしておくことが望ましい（**大学間連携**等）	教育方法（例）※ 1，2，3，4

※上記の「教育方法」欄の手法・形式 1～4 については次頁以降を参照

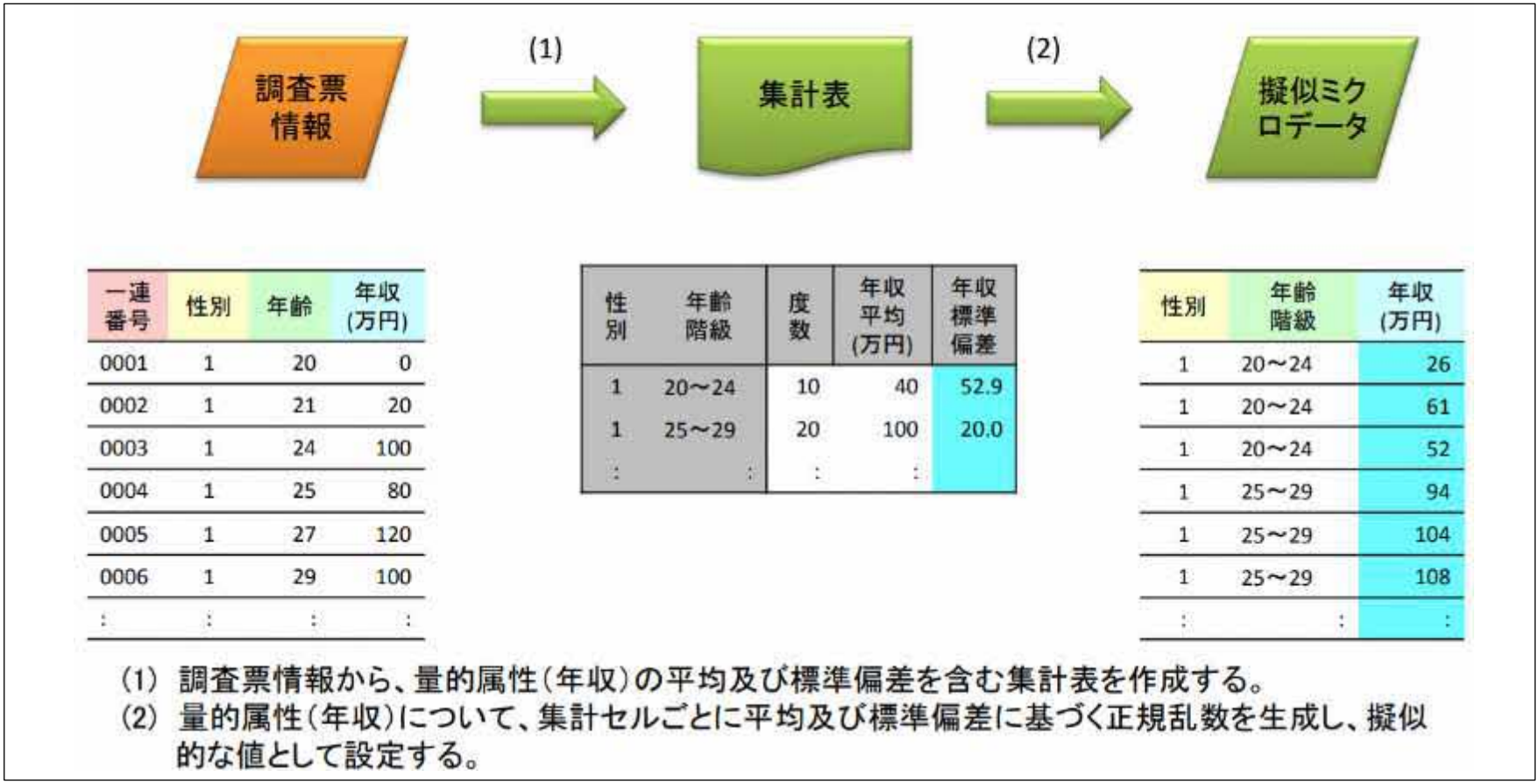
19

「数理・データサイエンス・AI（リテラシーレベル）モデルカリキュラム」
（2020年４月、数理・データサイエンス教育強化拠点コンソーシアム）　から引用

一般用ミクロデータ

（「一般用ミクロデータ（仮称）の作成及び利活用について」（2015年度統計関連学会連合大会）から引用）

- 現行の制度上、ミクロデータから**直接的にレコード単位の疑似データ**を**作成・提供**することは**できない**
- 中間的な集計表を公表し、そこから疑似データを作成



一連番号	性別	年齢	年収(万円)
0001	1	20	0
0002	1	21	20
0003	1	24	100
0004	1	25	80
0005	1	27	120
0006	1	29	100
:	:	:	:

性別	年齢階級	度数	年収平均(万円)	年収標準偏差
1	20～24	10	40	52.9
1	25～29	20	100	20.0
:	:	:	:	

性別	年齢階級	年収(万円)
1	20～24	26
1	20～24	61
1	20～24	52
1	25～29	94
1	25～29	104
1	25～29	108
:	:	:

(1) 調査票情報から、量的属性（年収）の平均及び標準偏差を含む集計表を作成する。
(2) 量的属性（年収）について、集計セルごとに平均及び標準偏差に基づく正規乱数を生成し、擬似的な値として設定する。

⇒ **変数間（離散・連続）の関係（回帰モデル）**を考慮した疑似データの作成が課題

１．公的統計ミクロデータの概要

２．疑似データの重要性と課題

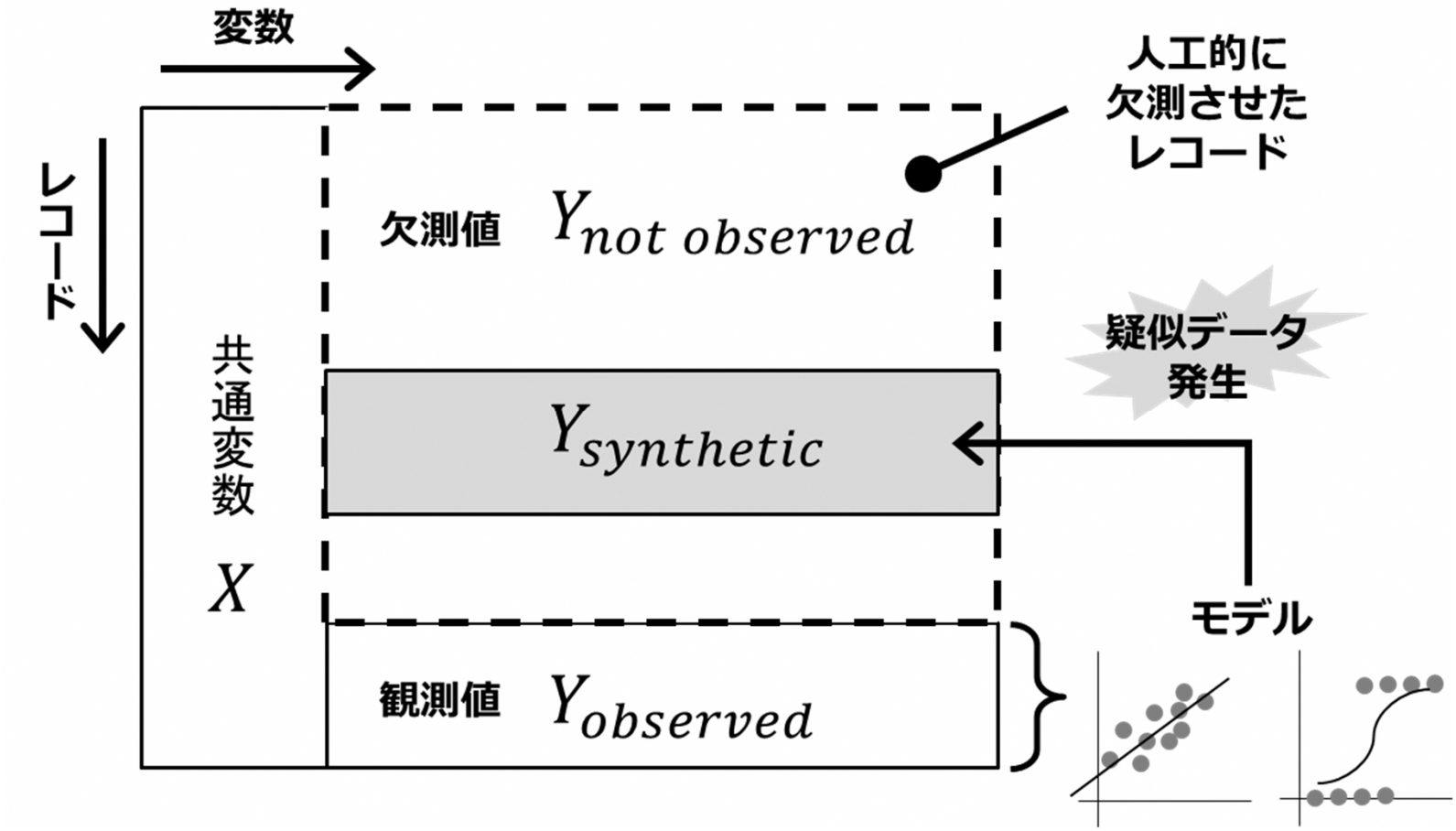
☞**３．Synthetic Dataに基づく方法**

４．実データに基づく試作

５．まとめと今後の課題

Synthetic Data

- 一部のレコードを人工的に欠測させ回帰モデル、ロジットモデル等により逐次的に疑似データを作成
  ⇒変数間の関係を保持したデータの作成が可能
  （Rのパッケージも提供されている（“synthpop”））



疑似データの作成方法

- ただし現行制度上、ミクロデータから直接的に**レコード単位の疑似データ**を**作成・提供できない**

  ⇒ **現行の制度に合わせた形**でSynthetic Data を作成するための工夫が必要

- 本研究では、**事前に集計・推定**した**統計表**と**モデルの結果**を基に、**Synthetic Dataの方法**に基づいて疑似データを作成する方法について検討

  ⇒ どのような集計結果、モデルの結果を公表すれば、秘匿性を確保した上で、元データの構造を保持した疑似的なデータが作成できるかを検討

１．公的統計ミクロデータの概要

２．疑似データの重要性と課題

３．Synthetic Dataに基づく方法

４．実データに基づく試作

５．まとめと今後の課題

分析に用いたデータ

- 商用データ(平成24年２月分)
  - ある県のデータを使用
  - サイズ：7,558レコード
  - 連続変数：
    - 従業員数（対数変換）
    - 売上高（対数変換）
  - 離散変数：
    - 産業（建設業、製造業、小売業、それ以外）
    - 地域（３地域）
    - 経営組織（株式会社、有限会社）
    - 開設年（～1984年, 1985年～1994年, 1995年～）
    - 資本金（300万円以上500万円未満, 500万円以上1000万円未満, 1000万円以上2000万円未満, 2000万円以上）

データ作成の流れ

- 周辺分布（地域(3)×産業(4)）の集計

	建設業	製造業	卸・小売業	左記以外	合計
地域１	1065	748	645	512	2970
地域２	935	508	831	827	3101
地域３	389	357	466	275	1487
合計	2389	1613	1942	1614	7558

- 上記の周辺分布に合うような7,558レコードを生成
- 元データから事前に推定したモデルで変数を逐次的に推測

(1) 地域・産業から経営組織を推測【２項ロジット】
(2) 地域・産業・経営組織から開設年を推測【順序ロジット】
(3) 地域・産業・経営組織・開設年から資本金を推測【順序ロジット】
(4) 地域・産業・経営組織・開設年・資本金から従業者数を推測【重回帰】
(5) 地域・産業・経営組織・開設年・資本金・従業者数から売上高を推測【重回帰】

※(4)・(5)：重回帰の残差から中央値(MD)、中央絶対偏差(MAD)を算出
⇒ $N(MD, MAD^2)$ に従う正規乱数を付与（0で切断）

モデルの推定結果

➢ 売上高を推測するモデルの推定結果

		Estimate	Std. Error	z value	Pr(>|z|)
定数項		2.948865	0.029566	99.737	< 2e-16
地域：	地域１【ベースライン】				
	地域２	0.105774	0.019574	5.404	6.73E-08
	地域３	0.007828	0.024043	0.326	0.745
産業：	建設業【ベースライン】				
	製造業	-0.103277	0.024815	-4.162	3.19E-05
	小売業	0.683146	0.023473	29.104	< 2e-16
	その他	-0.147661	0.024546	-6.016	1.87E-09
経営組織：	株式会社【ベースライン】				
	有限会社	-0.152575	0.027627	-5.523	3.45E-08
開設年：	～1984年【ベースライン】				
	1985年～1994年	0.085169	0.017506	4.865	1.17E-06
	1995年～	0.01336	0.015361	0.87	0.384
資本金：	300万円以上500万円未満【ベースライン】				
	500万円以上1000万円未満	0.350539	0.029154	12.024	< 2e-16
	1000万円以上2000万円未満	0.14328	0.018607	7.7	1.53E-14
	2000万円以上	0.083222	0.01993	4.176	3.00E-05
従業員数（対数）		0.912558	0.009206	99.129	< 2e-16

中央値(MD)：-0.02701386
中央絶対偏差(MAD)：0.8777895

試行データの作成結果

➢ 経営組織

【レコード数】	株式会社	有限会社	合計
元データ	4381	3177	7558
疑似データ	4416	3142	7558

【構成比】	株式会社	有限会社	合計
元データ	0.58	0.42	1.00
疑似データ	0.58	0.42	1.00

➢ 開設年

【レコード数】	～1984年	1985年～1994年	1995年～	合計
元データ	3018	2656	1884	7558
疑似データ	3092	2563	1903	7558

【構成比】	～1984年	1985年～1994年	1995年～	合計
元データ	0.40	0.35	0.25	1.00
疑似データ	0.41	0.34	0.25	1.00

試行データの作成結果

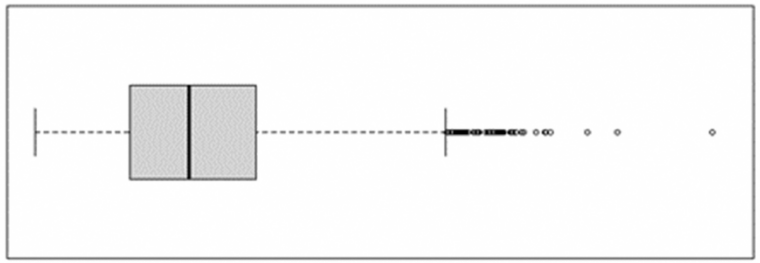
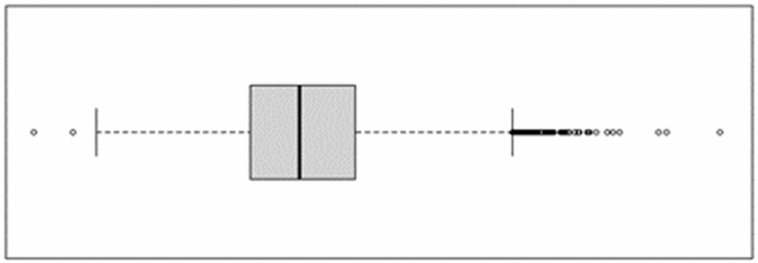
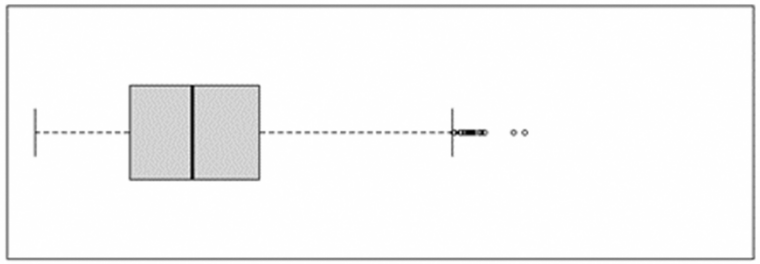
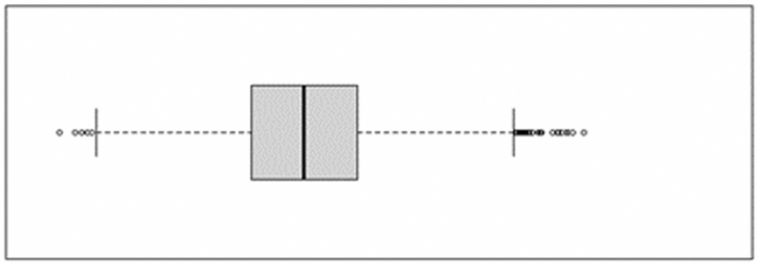
➢ 資本金

【レコード数】	300万円～	500万円～1000万円	1000万円～2000万円	2000万円～	合計
元データ	1896	1170	2864	1628	7558
疑似データ	1873	1169	2920	1596	7558

【構成比】	300万円～	500万円～1000万円	1000万円～2000万円	2000万円～	合計
元データ	0.25	0.15	0.38	0.22	1.00
疑似データ	0.25	0.15	0.39	0.21	1.00

試行データの作成結果

- 連続変数の分布

	従業員数	売上高
元データ		
疑似データ		

試行データの作成結果

➢ 相関係数

元データ	(1)地域	(2)産業	(3)経営組織	(4)開設年	(5)資本金	(6)従業員数	(7)売上高
(1)地域		0.09	-0.03	-0.04	0.01	-0.01	0.02
(2)産業			-0.02	-0.01	-0.04	-0.11	-0.04
(3)経営組織				0.27	-0.71	-0.40	-0.42
(4)開設年					-0.42	-0.25	-0.23
(5)資本金						0.48	0.49
(6)従業員数							0.79
(7)売上高							

疑似データ	(1)地域	(2)産業	(3)経営組織	(4)開設年	(5)資本金	(6)従業員数	(7)売上高
(1)地域		0.09	-0.03	-0.04	0.02	0.01	0.03
(2)産業			-0.03	0.01	-0.05	-0.12	-0.05
(3)経営組織				0.27	-0.72	-0.43	-0.44
(4)開設年					-0.41	-0.28	-0.25
(5)資本金						0.51	0.53
(6)従業員数							0.80
(7)売上高							

試行データの作成結果

重回帰モデル：売上高 ～ 地域 ＋ 産業 ＋ 経営組織 ＋ 開設年 ＋ 従業者数

		元データ				疑似データ			
		Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
定数項		2.948865	0.029566	99.737	< 2e-16	2.945073	0.027631	106.584	< 2e-16
地域：	地域１【ベースライン】								
	地域２	0.105774	0.019574	5.404	6.73E-08	0.124272	0.017619	7.053	1.90E-12
	地域３	0.007828	0.024043	0.326	0.745	-0.014865	0.021664	-0.686	0.493
産業：	建設業【ベースライン】								
	製造業	-0.103277	0.024815	-4.162	3.19E-05	-0.126383	0.022369	-5.65	1.66E-08
	小売業	0.683146	0.023473	29.104	< 2e-16	0.673735	0.02099	32.098	< 2e-16
	その他	-0.147661	0.024546	-6.016	1.87E-09	-0.166643	0.022165	-7.518	6.19E-14
経営組織：	株式会社【ベースライン】								
	有限会社	-0.152575	0.027627	-5.523	3.45E-08	-0.140898	0.023638	-5.961	2.63E-09
開設年：	～1984年【ベースライン】								
	1985年～1994年	0.085169	0.017506	4.865	1.17E-06	0.092696	0.015629	5.931	3.14E-09
	1995年～	0.01336	0.015361	0.87	0.384	0.008486	0.013555	0.626	0.531
資本金：	300万円以上500万円未満【ベースライン】								
	500万円以上1000万円未満	0.350539	0.029154	12.024	< 2e-16	0.372735	0.02591	14.386	< 2e-16
	1000万円以上2000万円未満	0.14328	0.018607	7.7	1.53E-14	0.161312	0.016809	9.597	< 2e-16
	2000万円以上	0.083222	0.01993	4.176	3.00E-05	0.079515	0.016976	4.684	2.86E-06
従業員数（対数）		0.912558	0.009206	99.129	< 2e-16	0.908968	0.0092	98.798	< 2e-16

変数の順序を変更した場合の影響

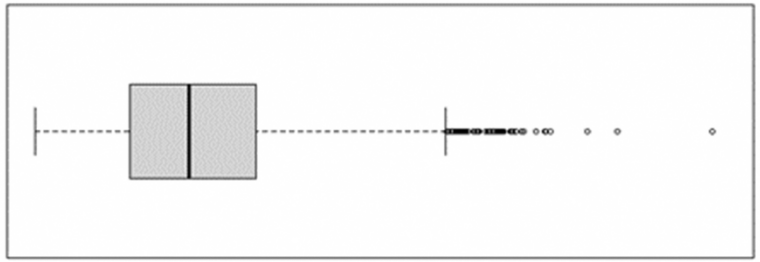
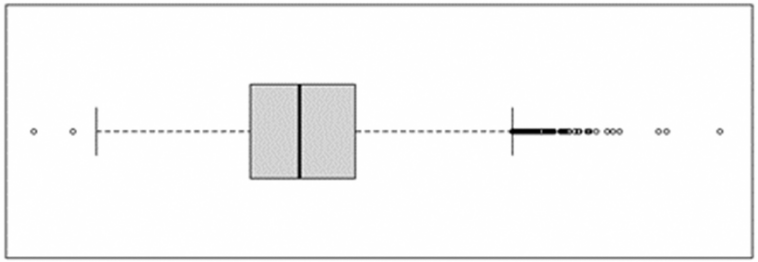
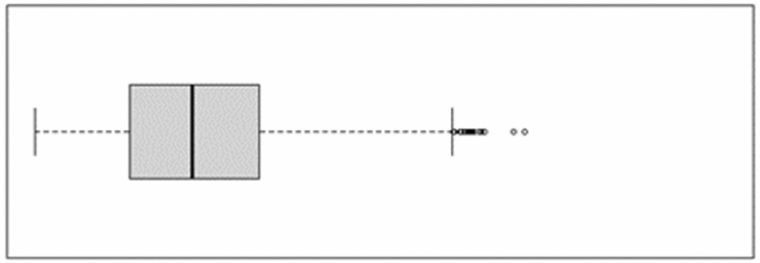
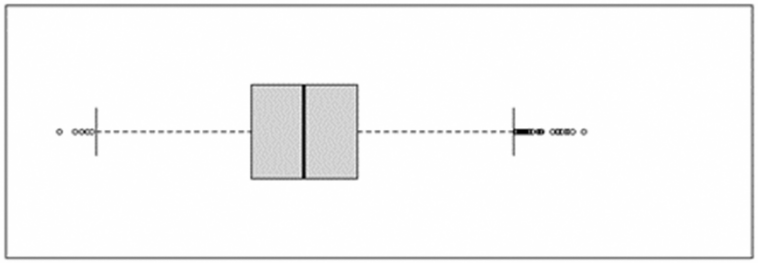
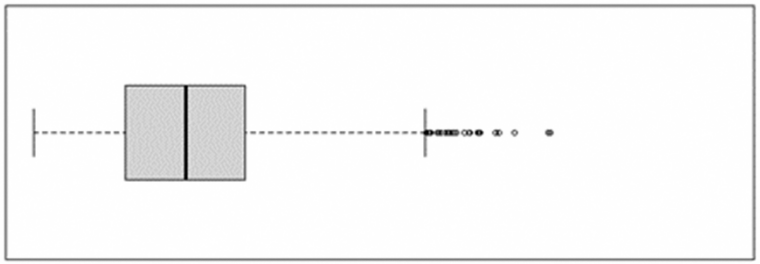
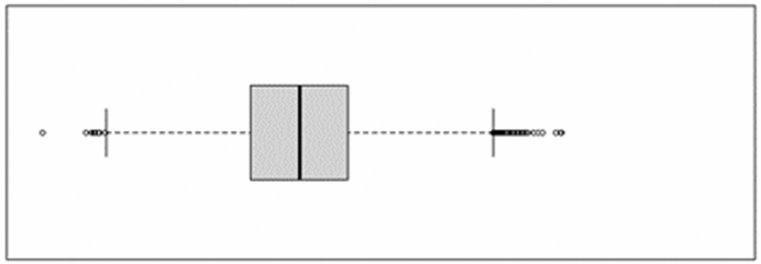
- 先に連続変数を推測するよう、変数の順序を変更

(1) 地域・産業から従業者数を推測【重回帰】
(2) 地域・産業・従業者数から売上高を推測【重回帰】
(3) 地域・産業・従業者数・売上高から経営組織を推測【２項ロジット】
(4) 地域・産業・従業者数・売上高・経営組織から開設年を推測
【順序ロジット】
(5) 地域・産業・従業者数・売上高・経営組織・開設年から
資本金を推測【順序ロジット】

※(1)・(2)：重回帰の残差から中央値(MD)、中央絶対偏差(MAD)を算出
⇒ $N(MD, MAD^2)$ に従う正規乱数を付与（0で切断）

- 離散変数の分布、相関係数、売上高を予測する重回帰モデルについては、変数の順序変更の影響は、それほど大きくないことを確認

変数の順序を変更した場合の影響

	従業員数	売上高
元データ		
疑似データ		
疑似データ（順序変更）		

１．公的統計ミクロデータの概要
２．疑似データの重要性と課題
３．Synthetic Dataに基づく方法
４．実データに基づく試作

☞５．まとめと今後の課題

まとめと今後の課題

- 商用データを用いた試作では、周辺分布の統計表、回帰モデルの結果、残差の分布に関する情報を事前に作成（公開）することにより、回帰モデルの構造をある程度保持したデータの作成ができた
- ただし、一部の変数では相関係数の係数の符号が逆になった
- 極端に大きな値の発生方法（発生させる必要があるか？、トップコーディング）
- 資本金など、特定の値に集中している変数では、ロジットモデル・回帰モデルの２段階推定が必要

まとめと今後の課題

- 周辺分布に合わせるための操作（比推定、繰返し比例補正など）の検討⇒必要な集計表の公開
- 変数の順序の影響を、より精緻に分析
- モデルの推測能力の向上、モデル推定の繰返し
- ロジットモデル・回帰モデルの２段階推定
  （資本金など、特定の値に集中している変数）
- 他の変数変換(Box-Coxなど)の検討⇒情報公開
- 元データに一致あるいは非常に近いデータが生じた場合の対応（データの削除、ノイズ付与）
- 差分プライバシーの考慮（適切なノイズの付与）

⇒世帯データを含む公的統計ミクロデータに対象を拡大して、さらに分析・検討を行っていく予定