CFPS を用いた家計のポートフォリオ分析

林田 実（北九州市立大学経済学部）

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」
2019年11月15日
於統計数理研究所

目次

1　はじめに

・研究の足跡

個票データを用いた、家計のポートフォリオ分析

総務省『家計調査　貯蓄負債編』
日本証券業協会『個人投資家の証券に関する意識調査』

プロビットモデル、オーダードプロビットモデル
ロジットモデル、サンプルセレクションモデル

計量経済学的手法の限界

機械学習→学習データで学習しテストデータでモデル評価
（検定に対する批判）

深層学習

SAS、stata→python、pandas、chainer

2　CFPS(China Family Panel Studies)

- 中国初のアカデミックな大規模縦断調査プロジェクト
- The Institute of Social Science Survey at Peking University
- 2010年調査がベースモデル
- その後、2012、2014、2016とフォローアップ調査
- 2010年調査：人口の95%が母集団。14,960の家計。42,590の個人
- 調査票5種類：①コミュニティ調査票、②家族構成調査票、③家族調査票、④子供調査票、⑤大人調査票

- 家族の新規分離、消滅なども考慮した調査設計

- ③家族調査票
  家族の日常生活、社会的なつながり、経済活動を記録する

  特徴的なものとして、社会的関係および主観的な自己観察（価値観、社会的態度、達成度、生活満足度）が含まれている→このような性格がポートフォリオ選択に影響を与えていることが分析できないか？

pandasの重要性

- 重要な前処理

③家族調査票には、家族がもつ株式についての情報あり
③家族調査票には、家族のメンバのIDが複数ある
CFPSは「世帯主」という概念は使わない

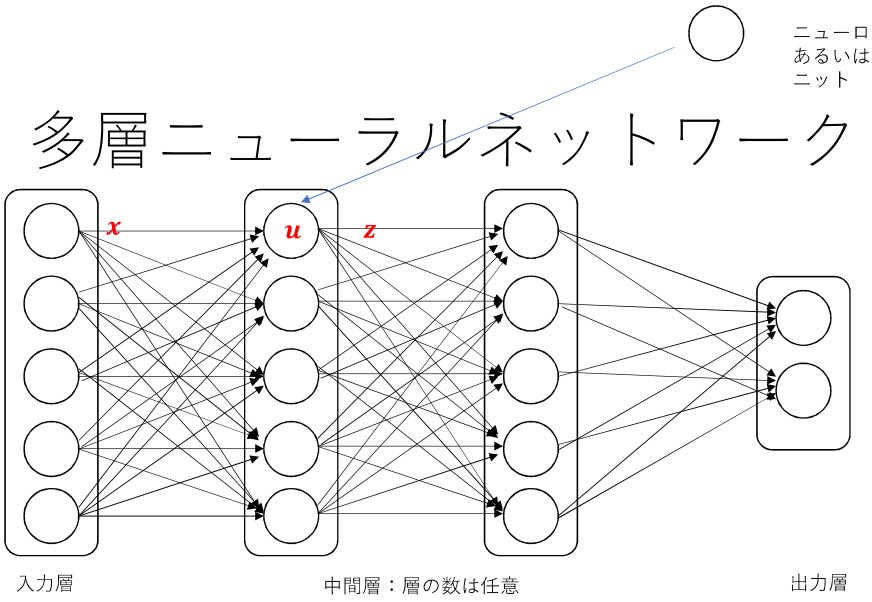
↓

そこで、⑤大人調査票から家族の稼ぎ頭を求め、それと③家族調査票とをマージする必要がある→pandasの利用（第3者の再利用・チェック可能）

3　深層学習

- 原語はDeep Learning
- 深い（多層の）ニューラルネットワーク（NN）による学習
- NNとは、脳の神経細胞（neuron）を数理モデルで表現したものをネットワーク上につなげたもの
- 「事前学習」を行うことにより多層のNNの学習が効率的となることが示された。
- ニューロンとは・・・

多層ニューラルネットワーク



- 他のニューロンからの電気刺激を集約し、ある値（閾値）を超えた場合に活性化する脳細胞
- これを以下のようにモデル化する。

- $u_j = \sum_{i=1}^{N} w_{ji}\, x_i + b_j,$
- $z_j = f(u_j).$

- u_j:第jニューロンの内部状態、　w_{ji}:第iから第jニューロンへの結合の強さ、
- x_i:第iニューロンの出力、　b_j:バイアス、
- $f()$：出力(活性化)関数、　z_j:第jニューロンの出力値。

活性化関数（出力関数）には次のようなものが用いられる。

- シグモイド関数　$f(u) = \frac{1}{1+e^{-u}}$.
- 正規化線形関数　$f(u) = \max(u, 0)$.

深層学習で重要なパラメータ等

- 中間層の数
- 確率的勾配降下法のバッチサイズ
- ネットワークデザイン

4　暫定的な分析結果

- 分析対象データ：2014年CFPS
- pandasを使って、前処理を行い、最終的に11,807家計を抽出
- 目的変数：株式の保有
- 説明変数（小）：所得、総金融資産
- 説明変数（大）：所得、総金融資産＋性格に関する変数（13）
- 学習データ：ランダムに選んだ9,807
- テストデータ：ランダムに選んだ2,000

ベースモデル：probit モデル
説明変数小（所得、総金融資産）

説明変数は対数ではない				説明変数は対数	
		予測値		予測値	
		非保有	保有	非保有	保有
実際値	非保有	1908	13	1921	0
	保有	76	3	79	0

89件　　79件

深層学習：中間層の数を変えて実験(unit=60, batch=5000)
説明変数小（所得、総金融資産の対数）

中間層の数			予測値		中間層の数			予測値	
			非保有	保有				非保有	保有
3	実際値	非保有	1921	0	7	実際値	非保有	1918	3
		保有	79	0			保有	77	2
4			非保有	保有	8			非保有	保有
	実際値	非保有	1914	7		実際値	非保有	1919	2
		保有	78	1			保有	77	2
5			非保有	保有	15			非保有	保有
	実際値	非保有	1921	0		実際値	非保有	1921	0
		保有	78	1			保有	77	2
6			非保有	保有					
	実際値	非保有	1913	8					
		保有	77	2					

深層学習：(続)説明変数小（所得、総金融資産の対数）

batchsize	中間層	ユニット数	train			test		AE回数	Epoch回数
5000	15	60	9436	9	0.963801	1921	0	100	500
			346	16		77	2		
5000	8	60	9435	10	0.9626	1919	2	100	500
			338	24		77	2		
5000	7	60	9435	10	0.964719	1919	3	100	500
			336	26		77	2		
5000	6	60	9434	11	0.963929	1919	3	100	500
			343	26		77	2		
5000	5	60	9441	4	0.963597	1921	0	100	500
			353	9		78	1		
5000	4	60	9429	16	0.963699	1914	7	100	500
			340	22		78	1		
5000	3		9445	0	0.96319	1921	0	100	500
		60	361	1		79	0		
5000	2		9445	0	0.963088	1921	0	100	500
			362	0		79	0		

5　おわりに

- 深層学習は従来手法よりも若干予測力が高い
- 深層学習を経済分野で利用していくには、経済データにあったネットワークデザインが必要
- 深層学習ができる大規模データの必要性→プライバシーよりも公的利益の方が重要ではないか
- 現存する日本人一人一人を、深層学習の入力にしてシミュレーションするということも可能なのではないか
- データの前処理の道具としてpandasは重要（エクセル不要）