情報・システム研究機構　機構間連携・文理融合プロジェクト
「言語における系統・変異・多様性とその数理」シンポジウム
(2018.02.02; 於TKP東京駅大手町カンファレンスセンター)

方言音声共通語化プロセスの確率モデル

前川 喜久雄（国立国語研究所）

自己紹介

- 音声学

  ⇒1979年来の一貫した興味の対象。特に自発音声とその変異

- 言語資源

  ⇒1999年からの副業（エフォート的には近年までこちらが本業）

- でも、今日の主題からすればふたつは結局つながっている

  ⇒どちらも統計的な分析が必須

- 第三回鶴岡調査（1991-92)に調査員として参加

  ⇒今日はこのデータについて30年近く前に考えたことを実証します

研究の目的

- 目的：言語変異・変化研究への（比較的に）先端的な統計モデリングを導入することを試みる

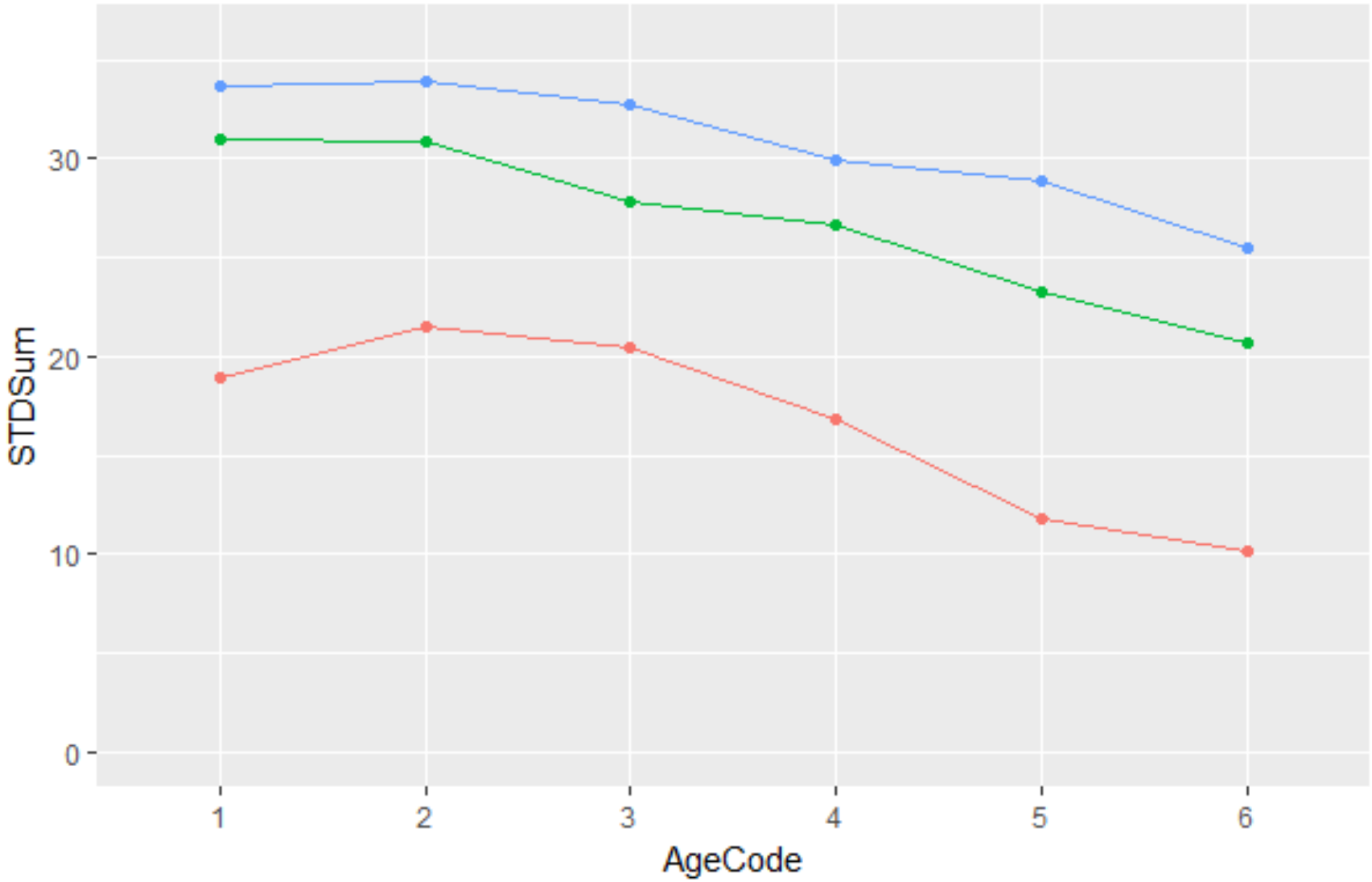
- 言語変異・言語変化の研究では従来から様々な統計手法が利用されてきている。しかし、必ずしも、適切に利用されてきていない

- 鶴岡調査データベースに記録された方言音韻の共通語化データを対象として、従来の分析の問題点を明らかにし、新しいモデリングの成果を紹介する

『鶴岡調査データベース』

- 鶴岡調査：国立国語研究所が1950 年以来、20 年間隔で４回実施してきた山形県鶴岡市における社会言語学的調査
- 方言の共通語化過程をリアルタイムで追跡
- 言語変化過程の実測データとして世界的に貴重
- 第１次~第３次調査の音韻項目（36 項目）のデータを一般公開（2017年4月, 同年10月に修正版公開）
- 本研究：公開データ（修正版）を用いて鶴岡における共通語化を統計的にモデル化
- 従来ほとんど検討されてきていない共通語化における個体差の問題の重要性を明らかにする

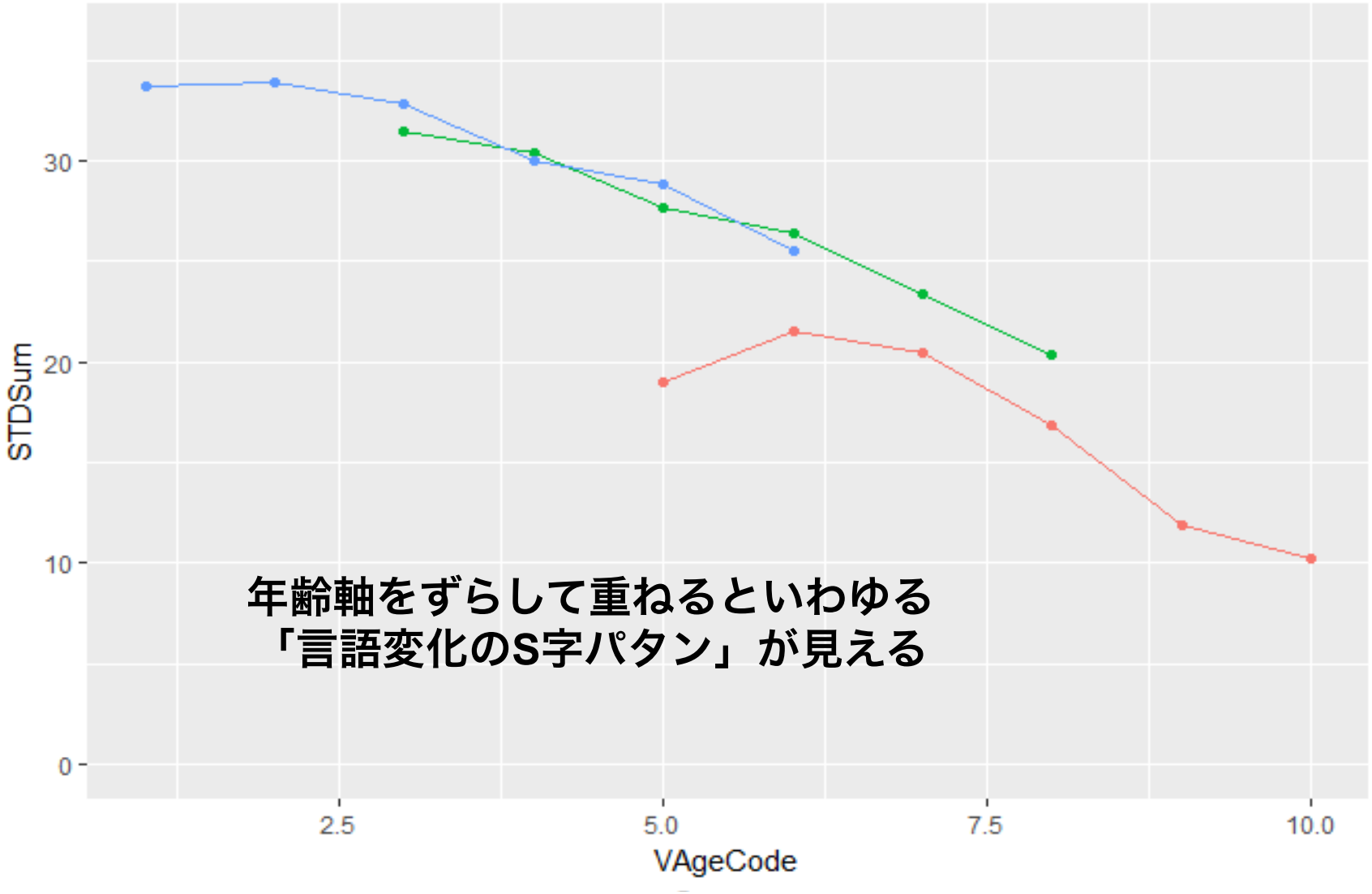
従来の成果：有名なグラフ



第1回～第3回調査の年代（10代～60代）毎の音韻項目の共通語化スコア平均値を比較

共通語化得点は0～36の区間に分布

従来の成果：有名なグラフ



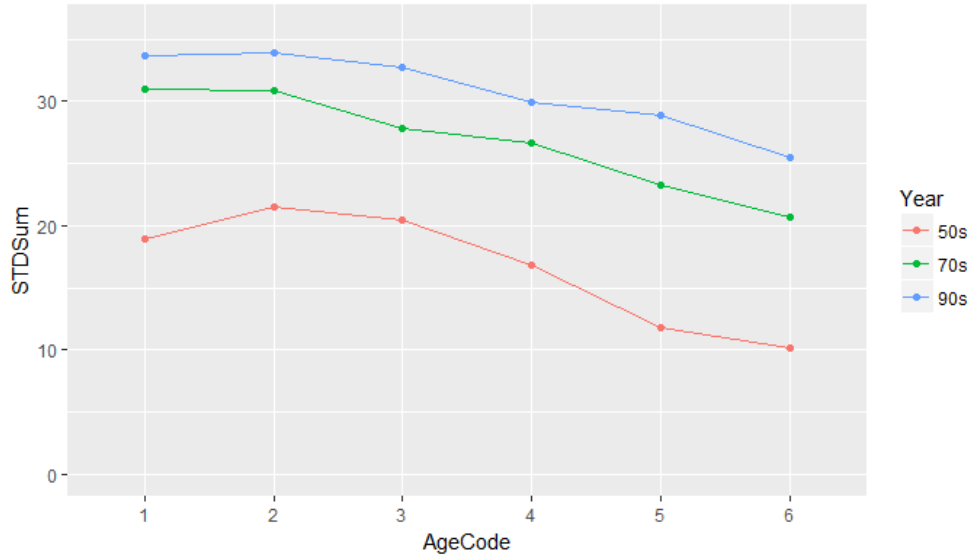
第1回～第3回調査の年代（10代～60代）毎の音韻項目の共通語化スコア平均値を比較

共通語化得点は0～36の区間に分布

このグラフは何を語っているか

1. 共通語化の諸要因のうち時間が最も重要である
2. 言語変化は時間とともに滑らかに進行する
3. 共通語化スコアには上限(36)と下限(0)がある

年齢（年代）から共通語化スコアを
予測できるとするモデル



実際にどの程度予測できるのか

- 第1次調査の対象者493名の共通語化得点を**年代**からロジスティック回帰分析（GLM)で予測

  ⇒平均予測誤差 7.06（36点満点に対して）

- 年代ではなく被調査者個人の**年齢**から予測

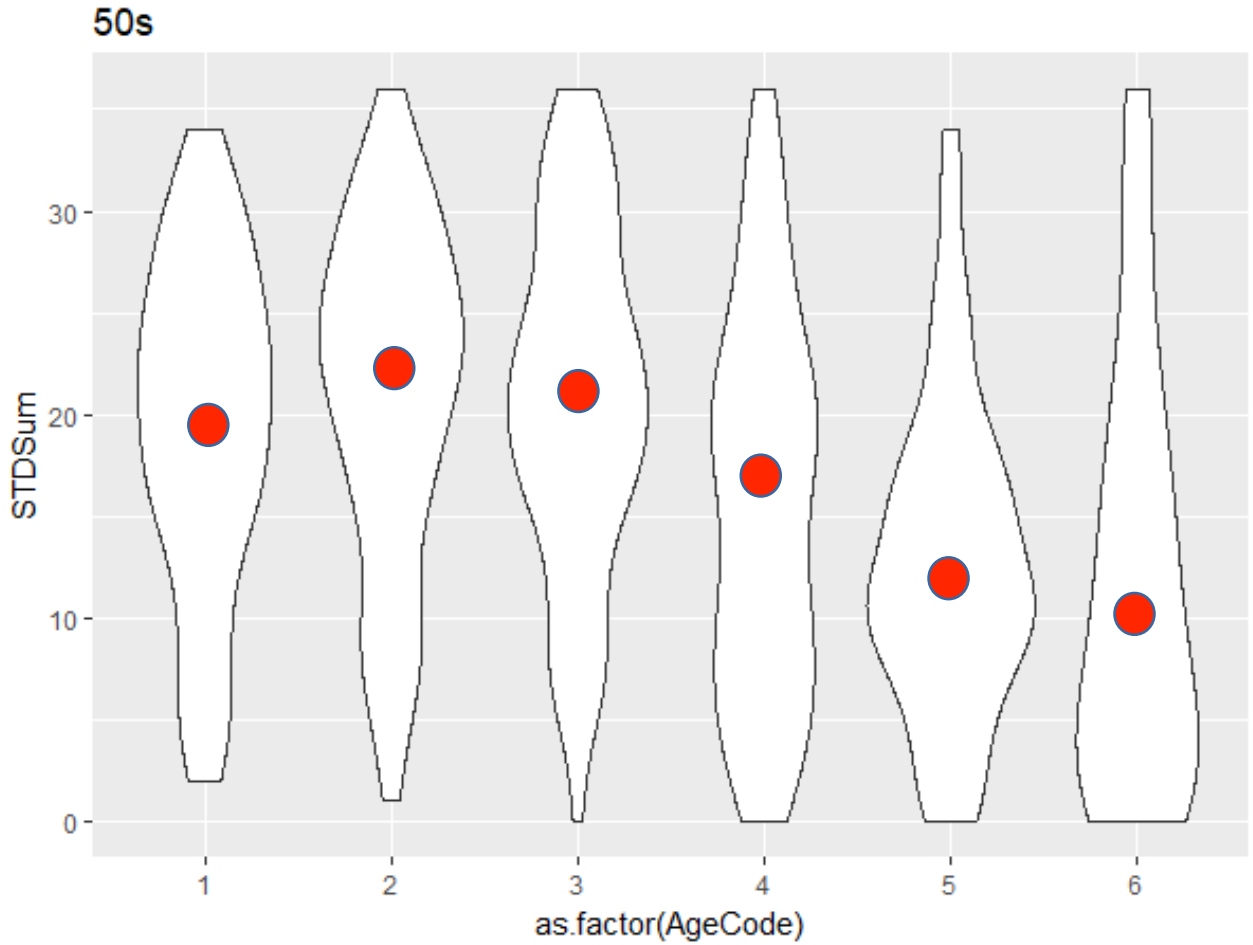
  ⇒平均予測誤差 6.95

- いずれも精度の良い予測とはとても言えない

- **性別**や**言語形成地域**の情報を追加してもたいして変わらない

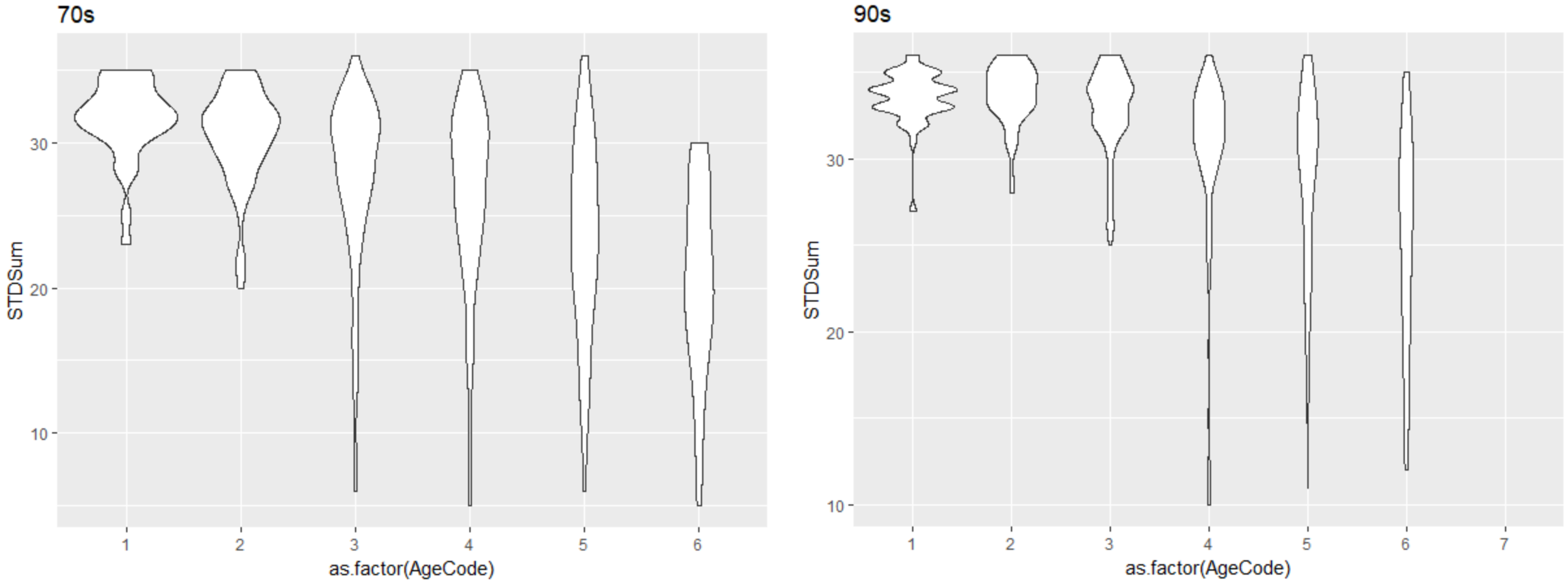
  ⇒平均予測誤差 6.53

根本的な問題：データと手法の不適合



第1回調査データの年代別バイオリンプロット：ヒストグラムを
平滑化して左右対称に描いたグラフ

根本的な問題：データと手法の不適合



第2回調査(70s)と第3回調査(90s)の年代別バイオリンプロット
やはり多くの年代で分布の裾野が広い

鶴岡データの数学的特性

1. 共通語化得点には上限（36）と下限（0）がある
   ⇒ いわゆる「カウントデータ」
2. 正規分布ではなく、二項分布にしたがうと考えられる

二項分布：成功確率 q のベルヌーイ試行（成功か失敗かの二値分布～コイン投げ）をN回繰り返したときにX回成功する確率を与える分布。

鶴岡データ音韻項目では、N＝36。共通語化の確率 q は年齢（年代）によって変化。

分散の理論値と実測値の比較

第1回調査(50s)

年代	被験者数	平均実測値	分散実測値	分散理論値	過分散指数
10	57	0.539	67.427	8.944	7.54
20	95	0.578	69.757	8.782	7.94
30	131	0.572	71.887	8.812	8.16
40	97	0.435	89.084	8.847	10.07
50	77	0.356	60.519	8.254	7.33
60	39	0.324	111.228	7.886	14.11

分散の理論値と実測値の比較

第2回調査(70s)

年代	被験者数	平均実測値	分散実測値	分散理論値	過分散指数
10	31	0.875	7.658	3.925	1.95
20	68	0.845	10.903	4.721	2.31
30	99	0.767	34.300	6.431	5.33
40	86	0.735	38.793	7.009	5.53
50	70	0.653	45.645	8.155	5.60
60	47	0.566	42.459	8.842	4.80

分散の理論値と実測値の比較

第3回調査(90s)

年代	被験者数	平均実測値	分散実測値	分散理論値	過分散指数
10	45	0.932	2.253	2.278	0.99
20	52	0.938	3.054	2.109	1.45
30	86	0.910	6.563	2.952	2.22
40	74	0.838	28.576	4.891	5.84
50	74	0.801	27.508	5.737	4.80
60	66	0.705	37.589	7.494	5.02

過分散が生じる原因

- データの個体差
  ⇒ 二項分布の成功確率 q が一定していない

- 個体差の発生要因
  1. 音韻クラス
  2. 個々の語彙項目
  3. 話者

音韻クラスと各クラスの語彙項目

- アクセント（団扇、烏、背中、旗、猫）
- 中舌化（地図、知事、辛子、島、烏、団扇、狐、炭）
- イとエ（息、駅、煙突）
- 唇音化（ひげ、百、蛇、西瓜、火曜日）
- 口蓋化（税務署、背中、汗）
- 前鼻音化（鈴、帯、窓）
- 有声化（糸、鳩、蜂、柿、猫、旗、口、靴、松）

統計的モデリング

- 鶴岡調査音韻項目の共通語化過程を個体差に配慮したベイズモデルによって調査毎にシミュレート。
- すべての話者のすべての回答（話者数×36個）を**ベルヌーイ分布に基づく回帰分析**で予測する
- 簡単なモデルから段階を追ってモデルを複雑化させ、最後にすべてのモデルを比較検討する
- シミュレーションにはStan言語を利用

モデル	特徴
1	inv_logit 関数に与える年齢と共通語化率の関係を示す一次式の**切片も傾きも一定**（ベースラインモデル）
2	**音韻クラスごと**に一次式の切片と傾きの両方が変化するモデル
3	**語彙項目ごと**に一次式の切片と傾きの両方が変化するモデル
4	**話者ごと**に一次式の切片だけが変化するモデル。傾きは一定
5	**話者ごと**に一次式の切片と傾きの両方が変化するモデル
6	**話者ごと**に一次式の切片が変化し**語彙項目ごと**に傾きが変動するモデル
7	**話者ごと・語彙項目ごと**に切片が変化し**語彙項目ごと**に傾きが変化するモデル

モデルの評価指標

評価指標	意味	
平均予測誤差	モデルによって予測された36×話者数個のサンプルの値（0か１）と実際の観測値との差の平均	小さいほど良いモデル。0と1の間の値
F値	モデルの予測値の適合率(1と予測したもののうち実際に1であったものの割合）と再現率(実際に1であったもののうち1と予測されたものの割合）の調和平均	大きいほど良いモデル。最大で1.0
WAIC	AICを非正規分布に適用できるよう拡張した情報量基準。交差検定と漸近等価。	小さいほど良いモデル

評価：第1回調査

モデル	平均予測誤差	F値	WAIC
1 (ベースライン)	0.422	0.568	24121.1
2 (音韻クラス～切片・傾き)	0.419	0.676	21503.6
3 (語彙～切片・傾き)	0.298	0.708	20318.0
4 (話者～切片)	0.284	0.714	20154.6
5 (話者～切片・傾き)	0.283	0.716	20155.6
6 (話者～切片, 語彙～傾き)	0.179	0.821	15266.5
7 (話者・語彙～切片, 語彙～傾き)	0.175	0.823	

どの指標をみてもモデル７が最良モデル

評価：第2回調査

モデル	平均予測誤差	F値	WAIC
1 (ベースライン)	0.261	0.850	15922.6
2 (音韻クラス～切片・傾き)	0.185	0.882	12462.6
3 (語彙～切片・傾き)	0.178	0.885	11886.0
4 (話者～切片)	0.230	0.858	14981.5
5 (話者～切片・傾き)	0.229	0.858	14983.9
6 (話者～切片, 語彙～傾き)	0.143	0.907	10591.0
7 (話者・語彙～切片, 語彙～傾き)	0.131	0.914	9992.6

やはりモデル７が最良モデル。モデル４は収束に難がある

評価：第3回調査

モデル	平均予測誤差	F値	WAIC
1 (ベースライン)	0.151	0.919	11427.7
2 (音韻クラス～切片・傾き)	0.128	0.928	8720.2
3 (語彙～切片・傾き)	0.120	0.932	8318.0
4 (話者～切片)	0.141	0.922	10907.8
5 (話者～切片・傾き)	0.142	0.921	10898.9
6 (話者～切片, 語彙～傾き)	0.093	0.947	7440.0
7 (話者・語彙～切片, 語彙～傾き)	0.088	0.949	6959.7

ここでもやはりモデル7が最良モデル

まとめと今後の課題

- 鶴岡データはどの調査も二項分布としては過分散の状態にある
- 通常のロジスティック回帰分析の適用は適切でない
- 個体差に配慮したベルヌーイ分布に基づく統計モデルを作ると高い予測精度を達成できる
- 特に第1回調査では個体差要因の貢献が大きい
- 今後の課題
  - 個体差の実態解明（特に第1回調査）
  - ３回の調査をまとめて分析できるモデルの構築
  - 言語変化研究への理論的貢献

謝辞

鶴岡調査の被調査者各位・調査員各位に感謝します