Probabilistic Reconsideration on the Dialect Standardization Data of Tsuruoka Survey

鶴岡市共通語化調査データの確率論的再検討

前川 喜久雄（国立国語研究所音声言語研究領域）†

Probabilistic Reconsideration on the Dialect Standardization Data of Tsuruoka Survey

Kikuo Maekawa (National Institute for Japanese Language and Linguistics)

要旨

方言音声の共通語化の過程をリアルタイムで記録した鶴岡調査の音声項目を統計モデリングの観点から再検討する。従来、この種の共通語化得点は二項分布によって生成されたと仮定してロジスティック回帰分析にかけることがあった。しかし、少なくとも鶴岡の場合、二項分布から理論的に予想されるよりもはるかに大きな分散が生じており、上記の仮定は適切でないことが明らかである。過分散状態が生じる原因は二項分布の成功確率が一定しておらず種々の要因によって変化することにあると考えられるが、重要な要因としては、調査項目の音韻クラス、調査語彙、そして話者の個体差が考えられる。これらの要因によってベルヌーイ試行の確率が変化するベイズ回帰モデルを多数考案して各モデルの予測性能を比較検討すると、上記すべての要因が予測性能に影響を及ぼすことが確認できた。このような分析が言語変化の理論に及ぼしうる影響について若干の議論を行った。

１．はじめに

国立国語研究所が 1950 年以来、ほぼ 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（以下鶴岡調査と呼ぶ）は、方言の共通語化過程をリアルタイムで追跡したデータとして有名である（国語研 1953, 1974, 2007）。本 2017 年 5 月に、鶴岡調査データのうち第１次~第３次調査の音韻項目（36 項目）のデータベースが一般公開され、誰でも自由に分析できるようになった（関連 URL 参照）。本研究では、一般公開されたデータを用いて、鶴岡における共通語化を統計的にモデル化することを試みる。そのなかで、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、鶴岡調査では、横軸に話者の年代を、縦軸に共通語化得点を配した座標系に年代ごとの共通語化得点の平均値をプロットしたグラフを描くことによって、時間の進行に沿って共通語化が進展する様子を把握してきた。公開された第１次~第３次調査のデータから、話者の年齢を 10 年区切りの年代にまとめ、各年代における共通語化得点（後述するように音韻項目は 36 項目あるので、特定話者の共通語化得点は最低 0 点から最高 36 点の範囲に分布する）の平均値を計算してグラフ化したのが図 1 である。３本の折れ線は第１次調査（凡例は 50s）、第２次調査（70s）、第３次調査(90s）における各年代の共通語化得点平均値を示しており、年代（横軸）の 1, 2, 3… はそれぞれ 10 代、20 代、30 代...を示している。

この種のグラフは、鶴岡調査の各段階でしばしば作成されてきている。また調査時期の差だけ 3 本の曲線を横にずらして重ね合わせた結果が言語変化のいわゆる S 字カーブに酷似することから、言語変化の一般理論についての研究でもしばしば利用されている（江川 1973,

† kikuo@ninjal.ac.jp

井上ほか 2009, 横山・真田 2010）。このグラフとその派生形は日本の社会言語学でもっともよく知られたグラフのひとつであると言ってよい。

ここで、このグラフの特徴をあらためて考えてみよう。まず、このグラフは共通語化得点に影響する可能性をもつ多くの言語的・社会的要因（音韻のクラス、個々の語彙項目、話者の性別・教育程度、調査者など）のなかで、時間が最も重要な要因であるとの仮定に立脚していることは明らかである。次に、グラフの形式として棒グラフではなく線グラフが選ばれていることから、言語変化（音韻の共通語化）が時間の進行とともに滑らかに進行すると仮定していることが推測される。グラフの横軸は平均値計算の便宜上 10 年ないし 5 年刻みの年代にカテゴライズされているが、本質的には連続的な変量（すなわち時間）を表していると考えられる。最後にグラフの縦軸は、便宜的に比率や百分率に変換されていることもあるが、実際には 0 から 36 までの範囲の整数値をとる離散データである。

統計モデリングの観点からすれば、このグラフは年齢をカウントデータの平均値に回帰させようとしている。つまり、話者の年齢を知ることで共通語化得点の平均値を予測しようとしていると見ることができる。現代の統計学では、そのような目的を達成するための手段として、ロジスティック回帰分析が頻繁に利用されることは周知のとおりである（Yokoyama & Sanada 2009）。

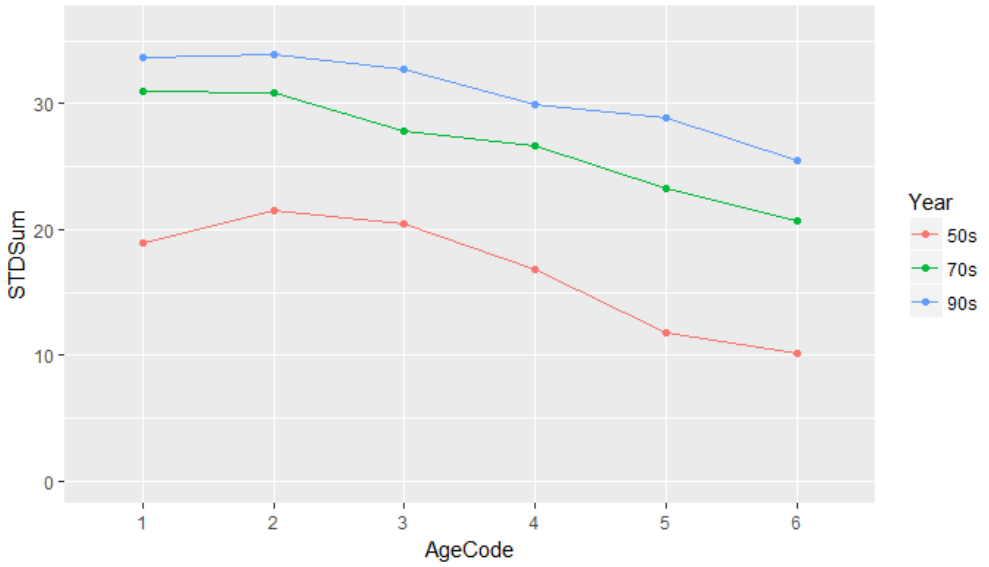

図 1：鶴岡調査（第 1 次～第 3 次）における年代別平均共通語化得点
50s は第 1 次調査, 70s は第 2 次調査, 90s は第 3 次調査を示す

2. 問題：過分散と個体差

それでは、実際、話し手の年齢を知ることで、共通語化得点をどれくらい正確に予測できるだろうか。第 1 次調査のデータを用いて、年代から共通語化得点を予測するためのロジスティック回帰分析を試みると、平均予測誤差は 7.06 となった（0～36 の範囲の値をとりうるデータに対する誤差である）[1]。これは、年代ごとに計算された平均値に対する誤差であることを考えると、精度のよい予測とはとてもいえない数字である。年代に替えて個々の話

[1] R 言語の lme4 パッケージの glm 関数を利用して試行回数 36 の二項分布から生成されたカウントデータを想定して分析した。

者の年齢（14~69 才）を用いると、平均予測誤差は 6.95 となるが、大差はない。

このように低い予測精度しか得られない原因はどこにあるのだろうか。ひとつの原因は、年齢以外の共通語化の要因を無視していることである。実際、性別・出生地・職業・学齢などの要因を用いた重回帰モデルを構築すると予測の精度は向上する。しかし平均予測誤差が若干度低下する程度の改善にとどまり、これが根本的な問題であるとは考えにくい。より根本的な原因は、統計手法の側にあるか、データと統計手法の適合性にある可能性が高い。鶴岡調査のデータにロジスティック回帰分析を適用することの是非を検討してみよう。

図 2 の左パネルは第 1 次鶴岡調査におけるすべての被験者の共通語化得点（0~36）を話者の年代ごとにプロットしたものである。どの年代でも 0 近くから 36 近くまで話者がちらばっていることがわかる。話者の年代ではなく実年齢(14~69)を用いてグラフを作ると図 2 の右パネルになる。図中には局所回帰法(LOESS)によるスムージングの結果も示しておいた。網掛けは 95%信頼区間である。

図 2 からは鶴岡調査データには年代においても年齢においても大きな分散が生じていることが確認できる。この事実は、管見のかぎり、従来の研究では指摘されていない。データの分散の大きさは、一般に予測精度の低下につながると考えられるが、二項分布に依拠した分析の場合は、さらに特殊な問題が生じる。次にその点を論じる。

鶴岡調査における個々の質問項目に対する話者の回答は、回答が共通語形かそうでないかという二者択一の観点からまとめることができる。図 1~3 もそのようにまとめられている。これは、確率論の観点からすれば、回答が標準語であれば 1 そうでなければ 0 の値をとるベルヌーイ試行とみなすことを意味している。鶴岡調査の音韻項目は全体で 36 項目あるので（分節音に関する調査項目が 31 項目あり、そのうち 5 項目についてはアクセントも調査しているので、合計 36 項目になる）、一人の話者の共通語化得点は 0-36 の範囲に分布する。確率論の観点からすれば、この得点の分布は 36 個のベルヌーイ試行から構成される二項分布に従うと考えられる[2]。

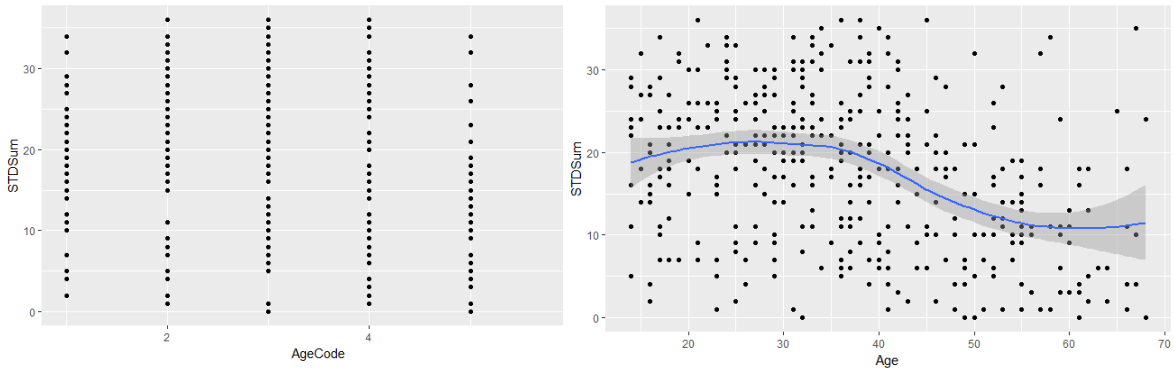

図 2：第 1 次鶴岡調査における全話者の平均共通語化得点
左は年代別、右は年齢別のプロット。右パネルには平滑化曲線を追加してある

ところで二項分布にはひとつの重要な特性がある。この分布のパラメータは試行における成功の確率 p と試行の回数 N であり、分布の分散は Np(1-p) で計算される。つまり成功の確率と試行回数 N が決まれば、分散は自動的に決まる。この点で二項分布は、正規分布のように分散（標準偏差）が分布のパラメータを構成している分布、つまり分散が自由に変化しうる分布と根本的に異なった性質をもっている。この特性に注目して、鶴岡データに実際に観察された分散の大きさを評価してみよう。

[2] Yokoyama & Sanada (2009)が示しているように正規分布とロジスティック曲線を結びつけることもできるが、鶴岡調査データに関するかぎり、正規分布を用いる必要はない。

第1次調査について、図1と図2の左パネルで利用した10年刻みの年代ごとに観察されたデータから各年代における p の推定値を求めて分散の理論値を計算し、実際にデータから計算された分散の観測値と比較すると表 1 の結果を得る。各行末の指数は分散観測値を分散理論値で除した値である。この指数から、すべての年代において理論値の 3 倍以上の分散が観察されていることが分かる。統計学ではこのような状態をデータが過分散の状態にあると言う。近年の統計解析の教科書には、過分散の状態にあるデータの解析に二項分布をそのまま適用することの問題を指摘しているものが多い(Kruschke 2015, 久保 2012 等)。

表 1：第 1 次鶴岡調査における分散の理論値と観測値

年代	N	p 推定値	分散理論値	分散観測値	過分散指数
1	46	0.528	11.464	63.156	5.509
2	46	0.597	19.492	69.553	3.568
3	123	0.568	8.835	66.183	7.491
4	88	0.469	8.965	86.042	9.598
5	63	0.329	7.952	64.382	8.097
6	32	0.283	7.305	71.899	9.843

過分散が生じる原因はデータの個体差にある。個体差とは、すべてのデータが同じベルヌーイ試行に従っていないこと、何らかの要因で成功の確率 p が変動していることを意味している。個体差というと言語学では話者の個人差を意味することが多いだろう。しかし、鶴岡調査の共通語化得点の場合、もう少し複雑な分析が必要である。

まず調査項目の言語学的な特性に由来する個体差が考えられる。鶴岡調査の音韻項目には、子音の副次調音に関する項目（有声化、唇音化、口蓋化、鼻音化)、母音イとエの調音位置に関する項目、母音の中舌化に関する項目、そしてアクセント項目から構成されている。これらを音韻クラスと呼ぶことにする。鶴岡調査の調査項目にどのような音韻クラスを認定するかはそれ自体が経験的な検討を要する問題であるが、ここでは調査票の分類をそのまま用いることにする。これらの音韻クラスにはそれぞれ複数の調査項目（調査語彙）が準備されている。

表 2 は鶴岡調査音韻項目について、まず音韻クラス別の平均共通語化率（平均共通語化得点を 36 で除した値）を示し、次いで各クラスに属する調査項目ごとに平均共通語化率を示したものである。この表からは、音韻クラスごとに平均共通語化得点が大幅に変動すること、また、有声化を除く他の音韻クラスでは、同一クラスに属する調査項目間でも 2 倍前後の平均得点差が生じていることが読みとれる。

図 3 は、鶴岡第 1 次調査について、話者の年齢と共通語化得点の平均値の散布図の上に音韻クラス別の回帰直線を描いたグラフである。図中の網掛けは、回帰直線の 95%信頼区間である。この図からは共通語化データのモデリング上重要な二つの傾向を読みとることができる。

まず、アクセント（図下部に位置する赤の直線）とそれ以外の音韻クラスの間には顕著な共通語化率の差が存在している。アクセントの回帰直線は切片が小さいだけでなく、傾きも非常に小さい（つまり年齢の影響が小さい)。一方、アクセント以外の 6 クラス間の差は、主に切片のちがいであって、直線の傾きは全般的によく似た値を示している。ただし母音イとエに関する項目（i_e 薄緑色の直線）は他の 5 クラスよりも大きな傾きを有している。

表 2：鶴岡第 1 次調査データの音韻クラス・調査項目別平均共通語化得点

音韻クラス	平均共通語化得点	調査項目	平均共通語化得点
アクセント	0.089	団扇	0.061
		烏	0.086
		背中	0.092
		旗	0.094
		猫	0.114
中舌化	0.485	地図	0.318
		知事	0.330
		辛子	0.333
		島	0.500
		烏	0.555
		団扇	0.562
		狐	0.606
		炭	0.671
イとエ	0.462	息	0.282
		駅	0.351
		煙突	0.753
唇音化	0.569	ひげ	0.359
		百	0.424
		蛇	0.477
		西瓜	0.692
		火曜日	0.898
口蓋化	0.675	税務署	0.476
		背中	0.748
		汗	0.802
前鼻音化	0.442	鈴	0.327
		帯	0.485
		窓	0.513
有声化	0.638	糸	0.590
		鳩	0.611
		蜂	0.612
		柿	0.624
		猫	0.627
		旗	0.642
		口	0.663
		靴	0.675
		松	0.700

次に調査項目の個体差を検討する。同一音韻クラス内で調査項目ごとの平均標準語化得点に差があることは先に表 2 で確認したとおりであるが、ここでは、年齢との関係に注目する。図 4～6 は、音韻クラスごとにそのクラスに属する複数の調査項目の散布図を重ね合わせ、調査項目ごとの回帰直線を描いたグラフである。回帰直線に注目して個体差の生じ方を

分類すると、図 4（音韻クラスは「イとエ」）のように、切片だけが調査項目によって変化し、傾きはほとんど変化しないタイプもあるが、図 5（「唇音化」）や図 6「中舌化」のように、切片と傾きの両方が変化するタイプが多い。

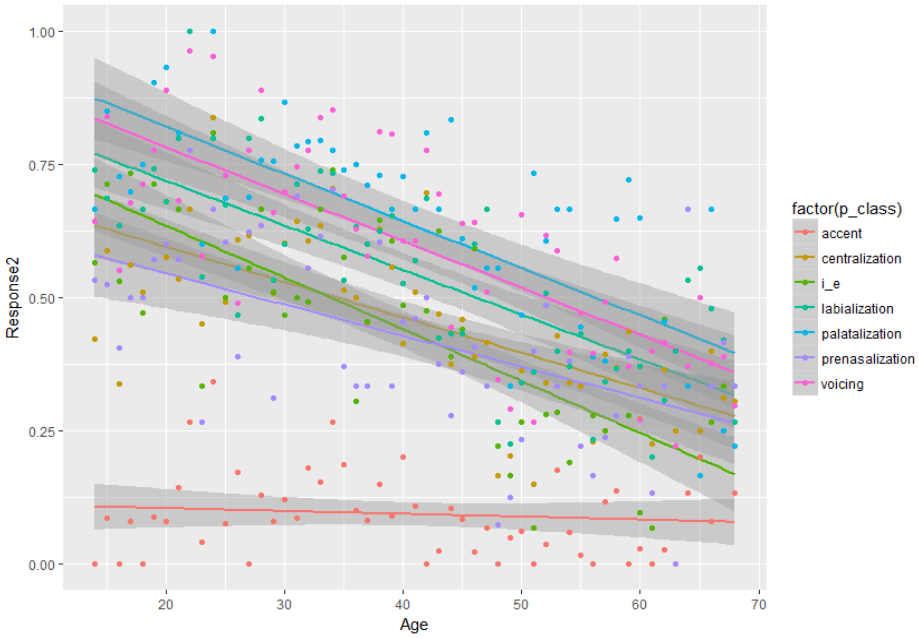

図 3：第 1 次鶴岡調査における音韻クラス別の話者の年齢（横軸）と平均共通語化得点（縦軸、満点を 1.0 に換算）の関係。直線は音韻クラス毎の回帰直線。網掛けは 95%信頼区間

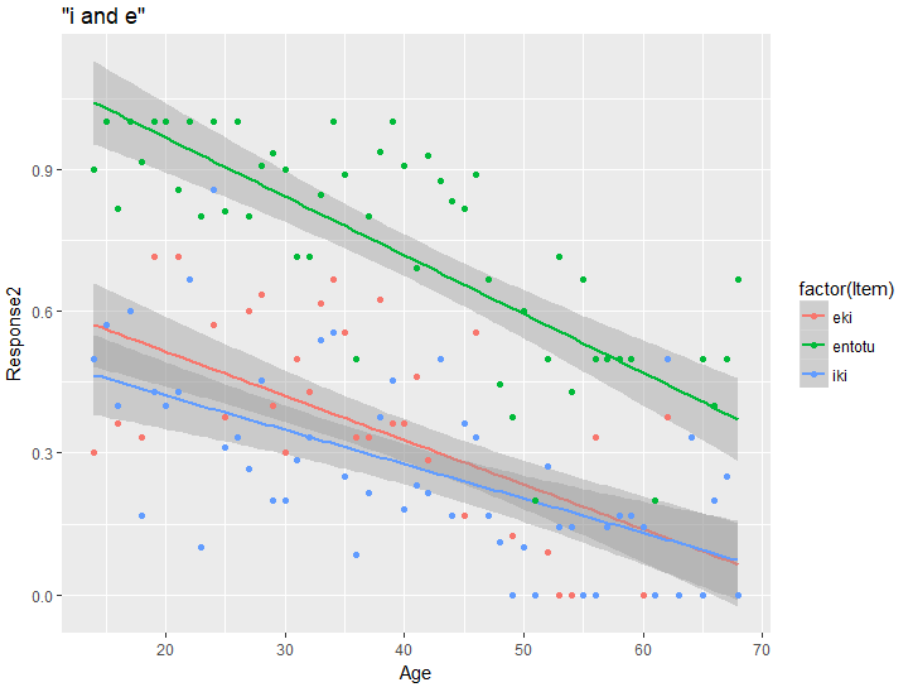

図 4：第 1 次鶴岡調査の音韻クラス「イとエ」における調査項目別の話者の年齢（横軸）と平均共通語化得点（縦軸、満点を 1.0 に換算）の関係。直線は音韻クラス毎の回帰直線。網掛けは 95%信頼区間

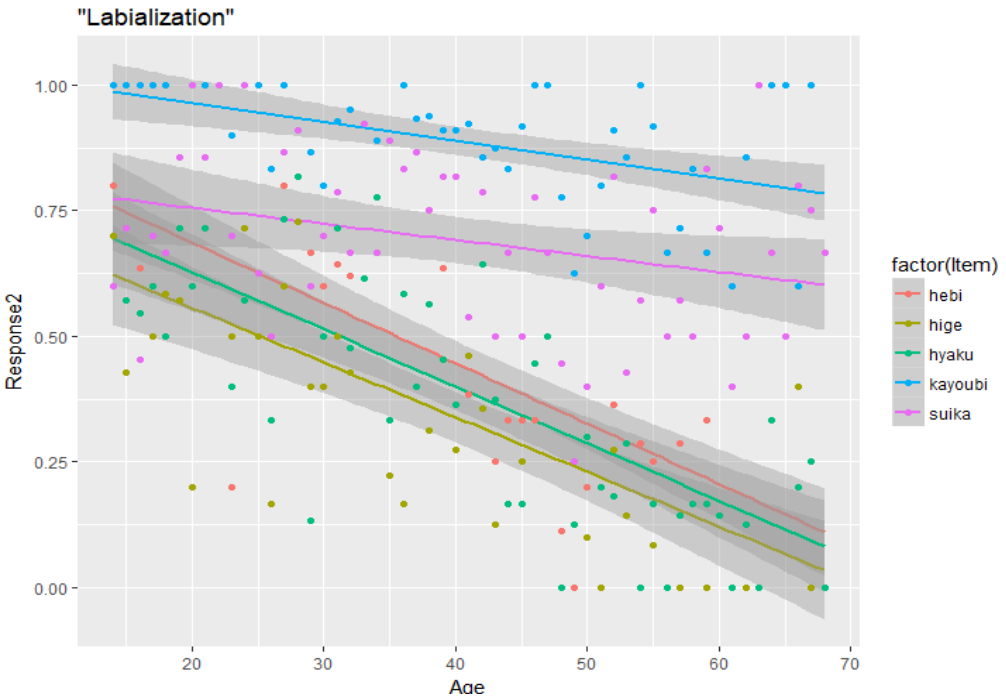

図 5：第 1 次鶴岡調査の音韻クラス「唇音化」における調査項目別の話者の年齢（横軸）と平均共通語化得点（縦軸、満点を 1.0 に換算）の関係。直線は音韻クラス毎の回帰直線。網掛けは 95%信頼区間

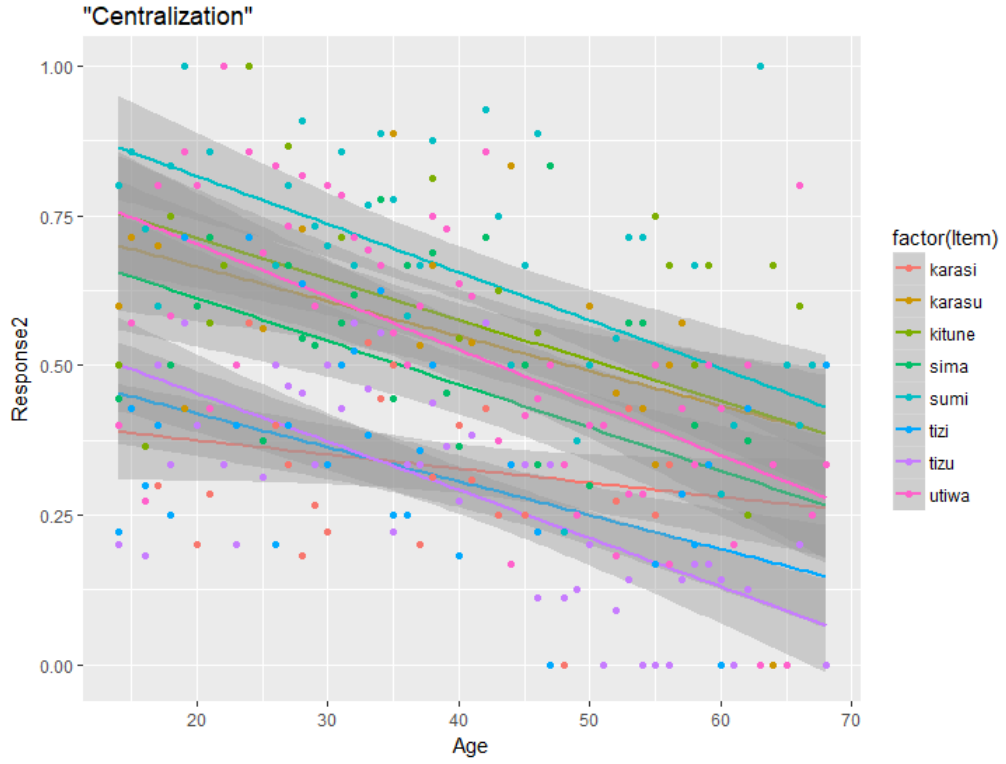

図 6：第 1 次鶴岡調査の音韻クラス「中舌化」における調査項目別の話者の年齢（横軸）と平均共通語化得点（縦軸、満点を 1.0 に換算）の関係。直線は音韻クラス毎の回帰直線。網掛けは 95%信頼区間

以上のように、第１次鶴岡調査の音韻項目は、年齢もしくは年代ごとに一定の確率をもつ二項分布によって生成されたとはみなしがたい性質を備えている。本稿では分析しないが、第２次、第３次調査のデータにも基本的に同じ性質が認められる。

このようなデータを解析するためには、話者の年齢だけでなく、音韻クラスや調査項目、さらには個々の話者に応じて、ベルヌーイ試行の確率が柔軟に変化する統計モデルが必要とされることは明らかである。そのようなモデルとして利用されるものに、一般化線形混合効果モデル(GLMM) や階層ベイズモデルがある。以下では、より柔軟性の高いベイズモデルを利用して分析を進めることにする。

３．統計的モデリング

本節ではベイズ統計モデルによって第１次鶴岡調査音韻項目の共通語化過程のシミュレーションを行う。ごく単純なモデルから次第に複雑なモデルへと段階を追って進み、最後にすべてのモデルを比較検討する。ベイズモデルは、ベイズの定理に依拠して、研究者が設定した現象の事前分布に、実際に観察されたデータから得られる尤度をかけあわせることで現象の事後分布を得る手法である。研究者が自由に事前分布を設定できるところが従来の統計モデルにない特徴だが、今回の分析では事前分布として無情報事前分布（一様分布や分散の大きな正規分布）を用いているので、事前分布の影響はほとんど認められない。GLMM によって分析してもほぼ同一の結論が得られるはずである。

以下のモデリングでは、図１や図２のように共通語化得点（話者ごとの 36 調査項目の合計点、つまりカウントデータ）を予測するのではなく、ひとつひとつの調査項目が共通語化しているかどうかを予測することにする。共通語化得点を用いたのでは、音韻クラスや調査項目の影響を析出することができないからである。具体的には、話者数×36（調査項目数）個のデータを用いて、ベルヌーイ分布（試行回数が１回の二項分布）に基づくロジスティック回帰分析を実施する。予測結果は、ある話者のある調査項目が共通語化している(1)か、していない(0)かの形で示される。

３．１　基本となる回帰モデル（モデル１）

ベイズモデルの推定は確率シミュレーション言語である Stan を利用して実行した。図 7 は、年齢だけを説明変数として共通語化の有無を予測するベイズ回帰モデルを Stan で実装したプログラムである。このプログラムの最重要部分は以下の 4 行である。

22 行は、観測されたデータ Y がベルヌーイ分布に従って生成されることを指定している。成功の確率 q はデータ１個ごとに異なる値をとる。i 番目のデータ Y[i]の成功の確率が q[i]である。データの総数(I)は話者数×36 である。

17 行は、q[i]がロジスティック関数 inv_logit に引数 b0＋b1*Age[i] を与えることで生成されることを指定している。引数は話者の i 番目のデータの話者の年齢 Age[i]の一次式で与えられており、ふたつのパラメータ b0（切片）と b1（傾き）を持っている。この一次式はパラメータ次第で－∞から＋∞までの値をとりうるが、inv_logit によって変換された出力は 0 から 1 の範囲に収まる。以下でとりあげる様々な統計モデルは、inv_logit 関数の引数を構成する一次式を複雑化させることで派生させるが、引数と出力の関係はすべて同様である。

30 行ではシミュレーションで得られたパラメータを用いて Y を予測している。予測値は y_pred[i]に格納される。以下では Y[i]と y_pred[i]の差の絶対値の平均値を平均予測誤差と呼び、モデル評価の指標のひとつとする。

最後に 31 行ではシミュレーションの過程で利用したモデルの対数尤度の値を記録している。これはモデルの評価指標である WAIC を計算するために必要だからである（次節参照）。

```
// BernlogitReg1_waic.stan
 data {
  int I;                                  //データ総数
  int<lower=14, upper=68> Age[I];         //話者の年齢
  int<lower=0, upper=1> Y[I];             // i 番目のデータが共通語か（1 と 0 で記録）
 }

 parameters {
 real b0;
 real b1;
}

transformed parameters {
  real q[I];
  for (i in 1:I)
    q[i] = inv_logit(b0 + b1*Age[i]);   //i 番目のデータの共通語化確率
}

model {
  for (i in 1:I) {
    Y[i] ~ bernoulli(q[i]);
  }
}

generated quantities {
  real y_pred[I];
  real log_lik[I];
  for (i in 1:I) {
    y_pred[i] = bernoulli_rng(q[i]);           //i 番目のデータの予測値を記録
    log_lik[i] = bernoulli_log(Y[i], q[i]);]   //対数尤度を記録(WAIC の計算用)
  }
}
```

図 7：ベイズ推定による回帰分析の Stan プログラム（モデル 1）

このプログラムを実行すると Stan 言語は 2 個のパラメータ b0 と b1 の分布（事後分布）をシミュレートする。ベイズ統計は研究者が自分の仮説をパラメータの事前分布という形で統計モデルに組み込める点に最大の特徴があるが、図 7 のプログラムでは事前分布は指定していない。その場合、Stan は大きな分散をもった正規分布を無情報事前分布として利用する[3]。

シミュレーションは 3 回実施し、各回で 2000 個のサンプルを得た。ただし各回の前半は捨ててしまうので、ひとつのパラメータの事後分布は 3000 個のサンプルから推定される。以下、この最も単純なモデルをモデル 1 と呼ぶことにする。

表 3 はシミュレーションを実行して得られた事後分布を四分位数で要約したものである。切片(b0)は 0.96 を、傾き(b1)は-0.03 をそれぞれ中央値とした左右対称の分布となっている。q[1], q[2],… はこれらのパラメータから推定された i 番目のデータに関するベルヌーイ試行の確率である。10 個目以降は省略しているが、全体で 17690 個の q[i]が推定されている[4]。Rhat はシミュレーションがうまく収束しているかどうかを判定する指標である。経験則として、Rhat が 1.1 以下であればシミュレーションは成功していると判断する。

[3] 分布の範囲が限定されていれば一様分布が利用されるが、図 7 では限定されていない。

[4] 第 1 次調査の話者数は 493 名なので総回答数は 493×36＝17748 個となるが、若干の無効回答が存在するので、これが実際のデータ数となる。

表 4 はモデル 1 によって予測の精度を評価した行列である。モデル 1 の正解率は 0.580, F 値は 0.571 となる。これが従来、共通語化の S 字カーブモデルと呼ばれてきたモデルの共通語化予測能力である。2 節でも指摘したが、決して高い数値ではない[5]。

表 3：事後分布の代表値（モデル 1）

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
b0	0.96	0.00	0.04	0.87	0.93	0.96	0.99	1.04	525	1
b1	-0.03	0.00	0.00	-0.03	-0.03	-0.03	-0.03	-0.02	566	1
q[1]	0.50	0.00	0.00	0.49	0.50	0.50	0.50	0.51	1746	1
q[2]	0.45	0.00	0.00	0.44	0.45	0.45	0.45	0.46	3000	1
q[3]	0.37	0.00	0.01	0.36	0.36	0.37	0.37	0.38	1139	1
q[4]	0.62	0.00	0.01	0.61	0.62	0.62	0.63	0.64	562	1
q[5]	0.33	0.00	0.01	0.31	0.32	0.33	0.33	0.34	911	1
q[6]	0.64	0.00	0.01	0.62	0.63	0.64	0.64	0.65	552	1
q[7]	0.52	0.00	0.00	0.51	0.51	0.52	0.52	0.52	1305	1
q[8]	0.42	0.00	0.00	0.41	0.41	0.42	0.42	0.43	3000	1
q[9]	0.50	0.00	0.00	0.49	0.49	0.50	0.50	0.50	2028	1

（以下両略）

表 4：モデル 1 による予測結果

	観測値	
予測値	**0（非共通語）**	**1（共通語）**
0（非共通語）	5327	3746
1（共通語）	3682	4935

３．２　様々なモデルとその評価

モデル 1 は変数が 1 個（年齢）だけの単回帰分析であるから、予測精度が良くないのは当然である。以下、予測モデルを少しずつ複雑化させていく。モデルは以下の 6 種類である。実際に利用した Stan のプログラムとその実行に利用した R 言語のプログラムを付録として掲載する。

モデル 1（既述）：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（以下単に一次式と呼ぶ）の切片も傾きも一定のモデル
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
モデル 3：調査項目（調査語彙）ごとに一次式の切片と傾きの両方が変化するモデル
モデル 4：話者ごとに一次式の切片が変化するモデル
モデル 5：話者ごとに一次式の切片が変化し調査項目ごとに傾きが変動するモデル
モデル 6：話者ごと・調査項目ごとに切片が変化し調査項目ごとに傾きが変化するモデル

これらのモデルについていくつか注意を述べる。モデル 2 とモデル 3 は独立したモデルとして扱っているが、実際には調査項目がわかれば音韻クラスの情報は自動的に判明するので、両モデルは完全には独立していない。

モデル 3,4,5 では、話者の個体差が一次式の切片に影響する可能性だけを考慮して、傾きに影響する可能性を排除している。特定の話者の年齢はひとつに決まっていることを考えればこの処理の妥当性は自明である。

これらのモデルの良さを評価するために 3 種の指標を利用する。モデルによる予測値と観測値の差の絶対値の平均である平均予測誤差、予測における Precision と Recover の調和平均として定義される F 値（F-measure）、およびモデルの対数尤度から計算される WAIC

[5] ただし 2 節で報告したのはカウントデータの分析、今回はベルヌーイ試行データに対する分析である点が異なっている。

の 3 種である。

平均予測誤差はでたらめな予測を行った場合は 0.5 となり完璧な予測においては 0.0 になると考えられる[6]。F 値は完璧な予測であれば 1.0 となる。WAIC は情報科学の領域で普及しつつある指標で、線形モデルの評価に用いられる AIC 同様、値が小さいほどよいモデルを意味している。WAIC による評価は交差検証(cross validation)と漸近等価であるとされている（関連 URL 参照)。

表 5 の評価指標は一貫してモデル 6 が最良モデルであることを示している。またモデル 5 が次善である点でも評価は一致している。つまり話者に起因する個体差と調査項目に起因する個体差の両方を考慮したモデルが最良のモデルであると考えられる。ただし、先に述べたように調査項目に関する個体差は、その一部として音韻クラスに関する個体差をも含んでいるから、結局、先に想定した 3 種類の個体差のすべてが予測精度の向上に貢献していると考えられる。

表 5：モデルの評価（鶴岡第 1 次調査データ）

	平均予測誤差	F 値	WAIC
モデル 1 （年齢のみ）	0.420	0.571	23951
モデル 2（音韻クラスの個体差が関与）	0.338	0.676	21329
モデル 3 （調査項目の個体差が関与）	0.296	0.710	20144
モデル 4 （話者の個体差が関与）	0.286	0.713	20088
モデル 5 （話者と調査項目が関与）	0.180	0.819	15215
モデル 6 （話者と調査項目が関与）	0.174	0.823	14636

4．議論

第 1 次鶴岡調査音声項目に観察される共通語化の実態は、3 種類の個体差に配慮したベイズモデルによって、かなりの程度まで（F 値が 0.8 を超える程度まで）正確にモデリングすることができることがわかった。この事実は言語研究にとってどのような意味をもつのだろうか。

言語変化については従来ふたつの理論的立場が対立してきている。ひとつは、言語変化は言語的な条件に従って、例外なく進行するとみる立場であり、もうひとつは、言語変化は心的辞書の内部でひとつの語彙項目から別の語彙項目へと漸次拡大していくことによって成立するとみる立場である。前者は neogrammarian principle、後者は lexical diffusion などと呼ばれる（Labov 1994)。

鶴岡方言に見られる共通語化を言語変化（音韻変化）の一例とみなすならば、今回の分析で確認された 3 要因のうち、音韻クラスの要因は前者に属し、調査項目の要因は後者に属する（ただし、繰り返しになるが、調査項目と音韻クラスは包含関係にあることに注意)。そして話者の個体差にはどちらの理論も積極的な注意を払っていない。

ところが、表 5 における平均予測誤差ないし F 値の差分を仮に各要因の貢献度とみなすならば、最も貢献が大きいのは話者の個体差である。もちろん話者の個体差の問題は、従来の言語学的分析でも等閑視されてきたわけではない。性別、学歴、成育歴等の話者の個体差に関する要因は、社会言語学において常に分析要因として重視されてきた。これらの要因が共通語化に関与していることは第 1 次鶴岡調査の報告書にも明瞭に指摘されている。

しかしながら、ここで、これらの社会的要因を考慮にいれても、話者の個体差情報の必要性が解消されるわけではないことを指摘しておかねばならない。上では議論を単純化する

[6] 第 2 節冒頭で述べた平均予測誤差（7.06 および 6.95）は 0～36 の値をとる対象に対する値であることに注意。ここでは 0 か 1 かのベルヌーイ試行に関する分析を行っている。

必要性から議論を省略したが、性別と出生地の情報がともに年齢の効果の一次式の切片を変化させるベイズモデルを用いたシミュレーションの結果は、平均予測誤差が 0.403, F 値が 0.586, WAIC 値が 23637 であり、モデル５,６に及ばないことはもちろん、モデル２の水準にも及ばないものであった[7]。

この結果は、共通語化を含む言語変化現象の分析において、従来社会言語学でとりあげられてきた種類の話者の社会的属性とは別に、いわば話者の純粋な個体差を想定することの必要性を強く示唆していると解釈できる。このような結論は、近年、心理学・社会学・生態学など、集団の挙動を問題とする科学領域におけるデータの統計分析において、いわゆるランダム効果の重要性が広く認識されている事実と軌を一にするものと考えられる。

ただし、鶴岡調査データの場合、今回最良のモデルにおいても正しく予測されなかったデータ変動のなかには、確率現象が本来的に有する誤差の他に、実はモデルによって説明可能な変動が含まれている可能性も否定できない。例えば、いわゆるスタイル差・場面差に起因する変動は今回のモデルではモデル化されていない。その理由は、調査そのものがスタイル差に配慮して設計されていないことに尽きるが、将来、例えば調査者の情報等が公開されれば、それをモデルに組み込むことなどで若干の改良は可能であろう。

最後にもうひとつ指摘しておく必要があるのは、（これも議論を省略したのだが）鶴岡調査の場合、表１に示された過分散の指標は第１次調査において最大であり、爾後、調査を繰り返す度に減少する傾向を見せていることである[8]。調査を繰り返す度に共通語化が進展していることを考えれば、これは当然の傾向であるが、話者の個体性の重要性もおそらく、これと相関しながら減少しているものと予想される。その意味では今回の分析は、話者の個体差の重要性を端的に示すデータの分析になっていると考えられる。

５．今後の課題

この研究では二つの事実を明らかにした。第一に、鶴岡方言の共通語化データはそれを二項分布から生成されたデータとみるかぎり過分散の状態にあり、二項分布を想定したロジスティック回帰による従来の共通語化モデルの予測精度は F 値で 0.6 以下にとどまること。第二に、過分散状態を生み出す原因と考えられるベルヌーイ試行の成功確率の変動は、調査項目と話者の個体差を考慮したベイズモデルを用いることで相当程度（F 値で 0.8 以上）まで説明できること、その際、わけても話者の個体差の影響が大きいと考えられることの二つである。

共通語化における話者の個体差の重要性を定量的に把握できたことが本研究の成果であるが、今回の分析は予備的と称すべき性格のものであり、本格的な分析は今後の課題である。

具体的な課題としては、①第２次および第３次調査データの分析と相互比較、②性別、出生地など既に公開されている話者の属性情報を組み込んだモデルの検討、③スタイル差・場面差などの変動を部分的にせよモデル化する試み、などがただちに思いうかぶ。

さらに、ちかく公開が予定されているパネル調査データの音韻項目を利用することで、年齢という仮想的な時間軸上ではなく、実時間上の変化をモデリングすることも興味深い課題である。また、将来公開されるであろう文法項目・言語生活項目に音韻項目と同様のモデル化が可能かどうかも検討すべき課題である。

謝　辞

鶴岡調査の被調査者・調査者各位に感謝します。本稿の草稿に対して横山詔一さん、淺原正幸さん、松田謙次郎さんからコメントをいただきました。記して感謝します。浅原さんからは

[7] 今回この種の分析を詳細に実施しなかったもう一つの理由は、現在公開されている鶴岡調査データの話者の属性情報には一部問題があると思われたことである。

[8] ただし第 3 次調査の段階でも過分散はほぼすべての世代に生じている。

Stan 言語の効率的な実行環境についても助言をもらいました。本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度)の成果であり、フィージビリティスタディ型共同研究プロジェクト「コーパスの時代に即した柔軟なデータ分析手法の導入に関する可能性評価研究」(2014-15 年度)の成果も反映しています。

文　献

井上史雄・江川清・佐藤亮一・米田正人.「音韻共通語化の S 字カーブ―鶴岡・山添 6 回の調査から―」計量国語学, 26 (8), 269-289, 2009.

江川清.「最近二十年間の言語生活の変容―鶴岡市における共通語化について―」言語生活, 257, 1973.

久保拓弥.『データ解析のための統計モデリング入門：一般化線型モデル・階層ベイズモデル・MCMC』岩波書店, 2012.

国立国語研究所.『地域社会の言語生活―鶴岡市における実態調査―』（国立国語研究所報告 5）秀英出版, 1953.

国立国語研究所.『地域社会の言語生活―鶴岡市における 20 年前との比較―』（国立国語研究所報告 52）秀英出版, 1974.

国立国語研究所.『地域社会の言語生活―鶴岡における 20 年間隔 3 回の継続調査―』国立国語研究所, 2007.

松浦健太郎.『Stan と R でベイズ統計モデリング』共立出版, 2016.

横山詔一・真田治子.「言語の生涯習得モデルによる共通語化予測」日本語の研究, 6 (2), 31-45, 2010.

Kruschke, John, K. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). NY: Academic Press, 2015.

Lobov, William. *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell, 1994.

Yokoyama, Shoichi, and Haruko Sanada. “Logistic regression model for predicting language change.” In R. Kohler (ed.) *Studies in Quantitative Linguistics 5, Issues in Quantitative Linguistics*, 2009, 176-192, RAM-Verlag, Germany, 2009.

関連 URL

鶴岡調査データベース：http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/waic2011.html

WAIC：http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/waic2011.html

付録：本稿で検討した統計モデルの Stan 言語による実装例と実行用の R 言語スクリプト
Stan プログラム中の//からその行の終わりまでは注釈。R スクリプト中の#で始まる部分も注釈。

■モデル１：年齢(Age)だけを独立変数としたベイズ単回帰分析

本文の図 7 参照

■モデル２：音韻クラス毎に一次式の切片と傾きの両方が変化するモデル

```
// BernLogitRegHie2_waic.stan
data {
```

```
  int I;
  int Npc;                                        //音韻クラスの総数
  int<lower=14, upper=68> Age[I];
  int<lower=0, upper=1> Y[I];
  int<lower=1, upper=Npc> Pclass[I];          //i 番目のデータの音韻クラス ID
}

parameters {
  real ap[Npc];                                   //音韻クラスごとに異なる係数
  real bp[Npc];
}

transformed parameters {
  real a[I];
  real b[I];
  real<lower=0, upper=1> q[I];
  for (i in 1:I) {
    a[i] = ap[Pclass[i]];                      //i 番目のデータの一次式の切片
    b[i] = bp[Pclass[i]];                      //i 番目のデータの一次式の傾き
    q[i] = inv_logit(a[i] + b[i]*Age[i]);
  }
}

model {
  for (i in 1:I){
    Y[i] ~ bernoulli(q[i]);
  }
}

generated quantities {
  real y_pred[I];
  real log_lik[I];
  for (i in 1:I) {
    y_pred[i] = bernoulli_rng(q[i]);
    log_lik[i] = bernoulli_log(Y[i], q[i]);
  }
}
```

■モデル３：調査項目毎に一次式の切片と傾きの両方が変化するモデル

```
// BernLogitRegHie3_waic.stan
data {
  int I;
  int Nitm;                                       //調査項目数
  int<lower=14, upper=68> Age[I];
  int<lower=0, upper=1> Y[I];
  int<lower=1, upper=Nitm> Item[I];           //i 番目のデータの調査項目 ID
}

parameters {
  real<lower=-5, upper=5> ap[Nitm];           //調査項目ごとに異なる係数
  real<lower=-0.1, upper=0.1> bp[Nitm];
}

transformed parameters {
  real a[I];
  real b[I];
```

```
  real<lower=0, upper=1> q[I];
  for (i in 1:I) {
    a[i] = ap[Item[i]];
    b[i] = bp[Item[i]];
    q[i] = inv_logit(a[i] + b[i]*Age[i]);
  }
}

model {
  for (i in 1:I){
    Y[i] ~ bernoulli(q[i]);
  }
}

generated quantities {
  real y_pred[I];
  real log_lik[I];
  for (i in 1:I) {
    y_pred[i] = bernoulli_rng(q[i]);
    log_lik[i] = bernoulli_log(Y[i], q[i]);
  }
}
```

■モデル４：話者ごとに一次式の切片が変化するモデル

```
# BernLogitRegHie3_5_waic.stan
# Subject2 をハイパーパラメータとして切片が変化する Bernoulli ロジスティック回帰
data {
  int I;
  int Nsbj;
  int<lower=14,upper=68> Age[I];
  int<lower=0, upper=1> Y[I];
  int<lower=1, upper=Nsbj> Subject[I];
}

parameters {
  real<lower=-5, upper=5> as[Nsbj];
  real b
}

transformed parameters {
  real a[I];
  real q[I];
  real<lower=0, upper=1> q[I];
  for (i in 1:I) {
    a[i] = as[Subject[i]];
    q[i] = inv_logit(a[i] + b*Age[i]);
  }
}

model {
  for (i in 1:I){
```

```
    Y[i] ~ bernoulli(q[i]);
  }
}

generated quantities {
  real y_pred[I];
  real log_lik[I];
  for (i in 1:I){
    y_pred[i] = bernoulli_rng(q[i]);
    log_lik[i] = bernoulli_log(Y[i], q[i]);
  }
}
```

■モデル５：話者ごとに一次式の切片が変化し、調査項目ごとに傾きが変動するモデル

```
// BernLogitRegHie6_waic.stan
data {
  int I;
  int Nitm;
  int Nsbj;                                          //被験者総数
  int<lower=1, upper=36> Item[I];
  int<lower=14,upper=68> Age[I];
  int<lower=0, upper=1> Y[I];
  int<lower=1, upper=Nsbj> Subject[I];          //i 番目の被験者 ID
}

parameters {
  real<lower=-5, upper=5> as[Nsbj];              //被験者ごとに異なる係数
  real<lower=-0.2,upper=0.1>  bs[Nitm];          //調査項目ごとに異なる係数
}

transformed parameters {
  real a[I];
  real b[I];
  real<lower=0, upper=1> q[I];
  for (i in 1:I) {
    a[i] = as[Subject[i]];
    b[i] = bs[Item[i]];
    q[i] = inv_logit(a[i] + b[i]*Age[i]);
  }
}

model {
  for (i in 1:I){
    Y[i] ~ bernoulli(q[i]);
  }
}

generated quantities {
  real y_pred[I];
  real log_lik[I];
  for (i in 1:I){
    y_pred[i] = bernoulli_rng(q[i]);
    log_lik[i] = bernoulli_log(Y[i], q[i]);
```

```
  }
}
```

■モデル６：話者ごと・調査項目ごとに切片が変化し調査項目ごとに傾きが変化するモデル

```
// BernLogitRegHie7_waic.stan
data {
  int I;
  int Nitm;
  int Nsbj;
  int<lower=1, upper=36> Item[I];
  int<lower=14, upper=68> Age[I];
  int<lower=0, upper=1> Y[I];
  int<lower=1, upper=Nsbj> Subject[I];
}

parameters {
  real<lower=-5, upper=5> as[Nsbj];             //被験者ごとに異なる係数
  real<lower=-5, upper=7> ai[Nitm];             //調査項目ごとに異なる係数
  real<lower=-0.2, upper=0.1> bi[Nitm];         //調査項目ごとに異なる係数
}

transformed parameters {
  real a[I];
  real b[I];
  real<lower=0, upper=1> q[I];
  for (i in 1:I) {
    a[i] = as[Subject[i]] + ai[Item[i]];   //被験者と調査項目がともに影響する
    b[i] = bi[Item[i]];
    q[i] = inv_logit(a[i] + b[i]*Age[i]);
  }
}

model {
  for (i in 1:I){
    Y[i] ~ bernoulli(q[i]);
  }
}

generated quantities {
  real y_pred[I];
  real log_lik[I];
  for (i in 1:I) {
    y_pred[i] = bernoulli_rng(q[i]);
    log_lik[i] = bernoulli_log(Y[i], q[i]);
  }
}
```

■Stan プログラムを R 環境で実行するためのスクリプト（モデル１の場合）

```
# BernLogitReg1_waic.R
# R のパッケージを読む
library(rstan)        # R から Stan を実行するライブラリ
library(loo)          # WAIC の計算に必要なライブラリ

# Stan に渡すデータをリスト形式で作る。R 上の dat1 というデータフレームに被験者の年
齢と回答がそれぞれ Age, Response2 という名前で記録されていると想定
```

```
bernLogitReg1_dat <- list(I=nrow(dat1), Age=dat1$Age,
Y=dat1$Response2)

# Stan でシミュレーションを 3 回実施
# “BernLogitReg1_waic.stan”が実行する Stan プログラムの名前
bernLogitReg1_waic_fit <- stan(file='BernLogitReg1_waic.stan',
data=bernLogitReg1_dat, seed=1234, chains=3)

# 推定結果を表示
bernLogitReg1_waic_fit

# 事後分布を抽出
bernLogitReg1_waic_mcmc <- rstan::extract(bernLogitReg1_waic_fit)

# WAIC を表示
loo::waic(bernLogitReg1_waic_mcmc$log_lik)

# 平均予測誤差とその標準偏差, Accuracy, Precision, Recall, F 値を計算
print("MeanPredError")
mean(abs(round(apply(bernLogitReg1_waic_mcmc$y_pred, 2, mean)) -
dat1$Response2))
print("SD of MeanPredError")
sqrt(var(abs(round(apply(bernLogitReg1_waic_mcmc$y_pred, 2, mean)) -
dat1$Response2)))
temp <- cbind(round(apply(bernLogitReg1_waic_mcmc$y_pred, 2, mean),0),
dat1$Response2)
temp_tab <- xtabs(~V1+V2, data=temp)
temp_tab
print("Accuracy")
sum(diag(temp_tab))/sum(temp_tab)
print("Precision")
precision <- temp_tab[2,2]/sum(temp_tab[2,])
precision
print("Recall")
recall <- temp_tab[2,2]/sum(temp_tab[,2])
recall
print ("F-measure")
2*precision*recall/(precision + recall)

# 事後分布を保存
save(bernLogitReg1_waic_fit, file=‘bernLogitReg1_waic_fit.R_obj’)
```