寄稿論文

AGLM：アクチュアリー実務のための データサイエンスの技術を用いたGLMの拡張

藤田 卓[*]　　田中 豊人[†]　　岩沢 宏和[‡]

2019年3月15日

概要

近年，機械学習やデータサイエンスにおける「予測モデリング」が非常に注目されており，アクチュアリーとしてそれを実務に応用することは重要課題の一つとなっている．しかし，アクチュアリーが主に扱う保険データの特殊性や，アクチュアリーとして意思決定を行う際の優先事項（たとえば，説明可能性や規制要件）などの制約から，そのまま実務に使えるケースは少なく，その結果，最先端技術を使いたいのになかなか使えず，頭を抱えている実務家も少なくないと考える．本研究では，「いかに既存手法を実務に落とし込むのか」のではなく，「アクチュアリーとしてどのような手法が求められているか」というスタンスから，一般化線形モデル（GLM）に対してデータサイエンス分野の技術を取り入れて発展させたモデリング手法である，Accurate GLM（AGLM）を提案する．

キーワード：回帰分析，一般化線形モデル（GLM），離散化，ダミー変数，正則化

1 背景

昨今，機械学習をはじめとしたデータサイエンスの手法[1]の発展は目覚ましく，アクチュアリーにとって，これらの手法を実務で活用することは重要課題の一つになっている．しかし，これらの分野のモデルや手法は，以下の点でアクチュアリーの実務には馴染まない面があり，これらが活用に向けた障壁となっている．

- 保険データの損害額分布などは，一般的に裾が長く，正規分布が適切でない場合が多い．また，リスクが既知のエクスポージャに比例すると考えられる場合があり，データそのものに特殊性がある．
- 結果の信頼性を確保するため，分析を行ったアクチュアリー以外の第三者による確認・検証や監査を受けることを求められることがあり、このとき特に結果の再現性が重要となるが，データサイエンスの手

[*] 所属：Guy Carpenter Japan, Inc.　連絡先：suguru.fujita@guycarp.com
[†] 所属：東京海上日動火災保険株式会社　連絡先：toyoto.tanaka@tmnf.jp
[‡] 連絡先：iwahiro@bb.mbn.or.jp

[1] データサイエンスの手法や技術の中でも，GLMなどの伝統的な統計学の手法以外のものを指す．

法は，一つのモデル構築に多くの手間と時間がかかる場合があり（多くのパラメータチューニングが必要な点など），分析者のバイアスが生じやすく，結果の再現性が担保できない懸念がある．

- プライシングやアンダーライティング業務では，当局や消費者に判断の根拠の説明が必要になるため，最終的な予測結果は，意思決定者に正しく理解されるものである必要があるが，モデルの枠組み自体が，予測精度という「結果」を重視するものであり，結果の根拠や，それに至る過程の説明は容易とは限らない．

以上からアクチュアリーの実務では，柔軟性があり，扱いやすく，説明しやすい手法が求められているが，ただ単に既存の手法を落とし込むだけでは限界がある．アクチュアリーの視点に立ち返り，これらを両立する手法を確立する必要がある．

そこで本研究では回帰問題に焦点を当て，**一般化線形モデル**（GLM: Generalized Linear Model）をベースとして，昨今のデータサイエンス分野の手法でかつ，アクチュアリーの実務と親和性が高いものを融合させた**Accurate GLM**[2]（以下，AGLM）を提案する．GLM は，上記の求められている条件を一定満たすものとして，アクチュアリーの実務で広く用いられている手法の一つである．また，従来アクチュアリーが行ってきた分析について，GLM のなかでも，変数選択や数値型特徴量の離散化ではプロットなどを利用した人手を介した判断に頼ってきたが，データサイエンスの手法により自動化または簡易化する方法を模索するのも，本研究における目的の一つである．

AGLM とは具体的には，GLM に**離散化**（Discretization）と O **ダミー変数**（Ordinal Dummy Variables）および**正則化**（Regularization）という三つの手法を取り入れたものである．これら三つの手法は，いずれもデータサイエンス分野において既存のシンプルな手法だが，GLM にこれらすべての組み合わせて自動化した手法は一般には用いられていない．これについては後述する．

本稿では，まず提案手法の基礎となる GLM に加えて，以降の数値実験で比較対象とするモデルについて概要を述べる．次に上記の三つの手法について整理し，今回の提案モデルである AGLM について説明する．そして，数値実験により既存手法と対比する形で AGLM の効果（簡便性，説明可能性および予測精度の両立）を確認する．最後に，今回得られた結果を踏まえて，アクチュアリアルモデリングとデータサイエンスの融合の可能性について考察する．

2 問題設定

この章では，AGLM の基礎となる GLM に加えて，以降の数値実験で AGLM の比較対象とする既存モデル（一般化加法モデル（GAM: Generalized Additive Model）と決定木（CART: Classification And Regression Tree））を紹介する．

2.1 一般化線形モデル（GLM）

GLM は，線形モデルを拡張した手法であることから，まず線形モデルの概要について説明し，その後 GLM について説明する．

2.1.1 線形モデル

線形モデル[3]とは，回帰モデルのもっともシンプルな形として，以下のとおり書けるモデルである．

---

[2] 「Accurate」には「予測精度や汎化性能が高い」という意味を込めている．
[3] たとえば文献 [3]や [22]などを参照されたい．

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (i = 1, 2, \cdots, n) \tag{1}$$

ここで，nはデータのレコード数，y_iは応答変数（目的変数），$x_{i1}, x_{i2}, \cdots, x_{ip}$は特徴量[4]（$p$は特徴量の数）である（以降も同様とする）．また，$\varepsilon_i$は誤差項であり，期待値が 0 で標準偏差が$\sigma$（等分散性）の正規分布に従う．

線形モデルの係数$\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^{\mathrm{T}}$は，以下で定義される残差平方和$\mathrm{RSS}(\beta_0, \cdots, \beta_p)$の最小化により算出する（これを最小二乗法とよぶ）．

$$\mathrm{RSS}(\beta_0, \cdots, \beta_p) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right)^2 \tag{2}$$

線形モデルは，応答変数にもっともよくあてはまる直線を算出し，残った誤差を回帰直線からのランダムかつ均質な変動として，正規分布でモデル化するものといえる．線形モデルは，各特徴量と応答変数の関係が明確であり，説明可能性に優れている．

2.1.2 GLM

線形モデルに対して，以下の二つの拡張を加えたモデルが GLM である．

① 線形性の拡張

GLM [1]は，応答変数の分布が指数型分布族に属するという仮定のもとで，その期待値の構造について，以下の関係を導入することで定式化されるモデルである．

$$\mathrm{E}[y_i] = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \quad (i = 1, 2, \cdots, n) \tag{3}$$

ここで，gはリンク関数とよばれる，特徴量の線形結合$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$と応答変数$y_i$の期待値を結びつけるものである．リンク関数$g$は，応答変数$y_i$の値域や，特徴量と応答変数の関係性などを踏まえて適切に選択する（ただし，gは微分可能かつ狭義単調である必要がある）．たとえば，対数リンク関数$g(\mu) = \ln \mu$（$g^{-1}(\eta) = e^{\eta}$）は，応答変数が正値で，大きな値を取り得て，特徴量の影響が乗法的にはたらくと想定される場合に有効とされている．表 2.1 にリンク関数の例を記載した．

表 2.1 リンク関数の例

	$g(\mu)$	$g^{-1}(\eta)$	g^{-1}の値域
恒等関数	μ	η	$(-\infty, \infty)$
対数関数	$\ln \mu$	e^{η}	$(0, \infty)$
ロジット関数	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\frac{e^{\eta}}{1+e^{\eta}}$	$(0,1)$
プロビット関数[5]	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	$(0,1)$

[4] 本稿では，説明変数とほぼ同義の「特徴量」という用語を主に使用する．ただし，説明変数の選択を話題にするときは，（特徴量選択とはいわず）「変数選択」といった表現を用いる．説明変数はモデルの式やモデルの分析の際に項として現れるもの，特徴量は説明変数の候補になるものとして，データに生で現れたり，データから加工（いわゆる特徴量エンジニアリング）して作られるもの，とする．

[5] Φは標準正規分布の分布関数

② 等分散性の拡張（応答変数の分布についての拡張）

線形モデルでは，y_iの分散は同一であると仮定している（等分散性）．しかし，たとえば応答変数y_iが事故件数などで，ポアソン分布に従うことが想定される場合は，レコードごとのモデル化で期待値と分散が等しい，という仮定をおき，またレコード間で異なる分散を仮定する必要があるため，等分散ではないモデルが望ましい．GLM では線形モデルを拡張して，応答変数y_iの分散を$\mathrm{E}[y_i]$（および，iによらない共通のパラメータ）の関数として表現することが可能である．また GLM は，応答変数の分布について指数型分布族を前提としており，アクチュアリーがクレーム分析においてよく扱うポアソン分布，ガンマ分布，Tweedie 分布などもカバーしている．さらに，オフセット項でエクスポージャの大きさを考慮できるという特長を持つ．

線形モデルは前述のとおり，応答変数にもっともよくあてはまる直線を導出するモデルであった．一方 GLM では，応答変数について指数型分布族を仮定した上で，期待値に対する特徴量の非線形な効果や，分散の不均質性をモデル化することができる．

以上を踏まえると，GLM はアクチュアリーの実務で求められる事項（応答変数を特徴量の乗法モデルで記述したい，応答変数の確率分布を正規分布ではなくポアソン分布やガンマ分布で仮定したい，など）に応えられることに加えて，線形モデルと同様，各特徴量と応答変数の関係が明確であり，説明可能性に優れているモデルといえる．

2.2 そのほかの既存モデル

ここでは，後の章の数値実験で比較対象となる既存モデルとして，一般化加法モデル（GAM）と決定木（CART）を紹介する．

2.2.1 一般化加法モデル（GAM）

GAM [2]は，応答変数の分布が指数型分布族に属するという仮定のもとで，その期待値の構造について，以下の関係を導入することで定式化されるモデルである．

$$g(\mathrm{E}[y_i]) = \alpha_i + \sum_{j=1}^{p} f_j(x_{ij}) \quad (i = 1, 2, \cdots, n) \tag{4}$$

ここで，gは微分可能でかつ狭義単調なリンク関数である．

GAM は加法的な構造を保ちつつも，各特徴量の寄与度を表すβ_jの代わりに，より一般的な関数f_jを使用する．またリンク関数を用いることで，応答変数についてさらなる自由度を与えるモデルである．

関数f_jは，一般的には三次平滑化スプラインが用いられる．すなわち，特徴量の定義域をいくつかの領域に分割し（観測値をそのまま用いることが多い），各節点を三次多項式で滑らかに接合する．ここで言う「滑らか」とは，各節点で連続であり，かつ一次導関数と二次導関数が連続となることを意味する．また，このほかに各観測点の近傍で線形回帰を行い，その結果をもとに平滑化する手法（Lowess 平滑化，カーネル平滑化など）もある．

目的関数は，予測値と観測値の適合度を評価する部分と，f_jの全体的な滑らかさを評価部分に分かれる．各f_jは，jについて逐次的に推定していくアルゴリズム（バックフィッティングアルゴリズム）により推定される．また，目的関数は適合度と滑らかさのトレードオフに対するハイパーパラメータを持つ．この決定方法の一つとして一個抜き法のクロスバリデーションがあるが，計算量が大きくなるため，実務上はこれを近似する手法として，一般化クロスバリデーション尺度（GCV）を用いる [3].

GLM と比較すると，GAM はモデルのあてはめにおいて柔軟性を発揮し，予測精度が向上する場合もあるが，その直感的な解釈は容易でなく，説明可能性の部分で劣後する．また，一般的に計算量が大きくなる傾向

があるため，実務ではその点についても留意する必要がある．

2.2.2 決定木（CART）

決定木 [4]は回帰モデルおよび分類器の一種で，データをある基準に従ってもっともよく分割する特徴量を検出し，二分木を作成するモデルである．線形モデルとは異なり，特徴量の交互作用などの非線形性を自動的に捉える特長があり，結果の可視化および特徴量と応答変数の関係の解釈が容易である [5]. その有用性から，モデリングの前段階のデータ処理においても多用される．

具体的には，以下の二つのステップにより決定木を構成する．

① 特徴量の空間を互いに重なりがないように分割する（ここでは$R_1, R_2, \ldots, R_J$とする）．
② 各特徴量の観測値に対して，その観測値が属するR_jに対応する応答変数の値を予測値として割り当てる．

たとえば，野球選手と年俸の関係について決定木により回帰させることを考える．ここで特徴量は，各選手の経験年数（Years）と年間のヒット数（Hits）の二つとする．

まず，{Years $<$ 4.5}と{Years $\geq$ 4.5}の領域に特徴量を分割する．さらに，{Years $\geq$ 4.5}の領域を{Years $\geq$ 4.5 かつ Hits $<$ 117.5}と{Years $\geq$ 4.5 かつ Hits $\geq$ 117.5}の領域に分割することを考える．新たなデータが与えられた場合，以下の図 2.2 のような木に沿って予測値を割り当てる（割り当て値は，たとえば訓練データのうち，各領域に属するものの平均値など）．

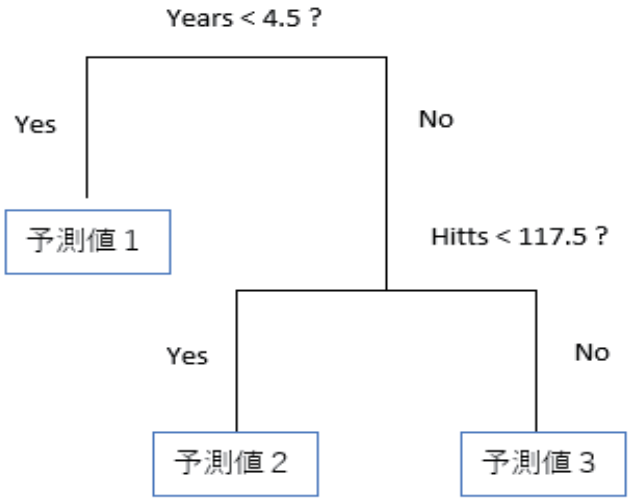

図 2.2 決定木の構成例

基本的な学習手法は，Recursive Binary Splitting Algorithm（再帰的 BSA 法）とよばれるものである．これは貪欲的に，つまり学習（特徴量の分割）においてその都度，二乗誤差を最小にする領域を探索する．なお，基本的には$\{X_j < s\}$という短冊型の分割しか考えず，学習の際はX_jとsの組み合わせを，その分割により二乗誤差がもっとも小さくなるものを選ぶ．ただし，貪欲的であり過学習（オーバーフィッティング）の可能性があるため，これを Pruning（枝刈り）する．具体的には，木の大きさ（ノード数）に対して何らかの罰則をつけて学

習を行う．

予測精度の向上方法として，バギング（Bagging），ブースティング（Boosting） [6]，ランダムフォレスト（Random Forest） [7]などがある．バギングはブートストラップサンプルに基づき，木を大量に構成し，平均をとる手法である．ランダムフォレストは，特徴量の一部をランダムに使用した木を構成し，最後に平均化する手法である．ブースティングは，ある木に対する誤差項を別の木で学習させて小さくすることを逐次的に行う手法である．これらは複数の弱学習器を組み合わせることで，予測精度を高める特長を持つ[6]．

決定木の特長として説明可能性があるが，上記のバギングなどの手法は，モデルを混ぜ合わせるため，説明可能性が損なわれることがアクチュアリーの実務に用いる上での問題である．

3 提案手法 Accurate GLM（AGLM）

この章では，GLM に組み合わせる三つの手法（離散化，O ダミー変数，正則化）について紹介し，新たなモデリング手法である AGLM を提案する．

3.1 離散化

アクチュアリーの実務では，死亡率などの応答変数を年齢などの数値型特徴量で表現するような問題がよく出現する．たとえば，死亡率を年齢で回帰する問題では，両者の関係が年齢群に応じて異なる（出生直後の死亡率，交通事故に起因する 20 歳前後の死亡率上昇，60 歳以降の死亡率急上昇など）といった非線形性を有しており，単純に GLM でモデル化することは難しい．そのほかにも，数値の取りうるレンジのごく狭い範囲にデータが集中している変数を，単純にそのまま特徴量として採用すると，一部の範囲にデータが集積してしまうことなどにより，レンジ全体を通した高い予測精度は見込めず，望ましくない場合がある．

そこで，上記の例における年齢や給与のような順序型特徴量[7]について，それぞれの区間ごとに応答変数への影響度を評価するため，いくつかのビンに分割する，**離散化**（Discretization）とよばれる手法を用いる．離散化することで，上記のような問題起因のいわゆる未学習（アンダーフィッティング）を回避し，より柔軟なモデリングが期待できる．

ただし一般的に，離散化する特徴量に対して，数値の取り得る値をそれぞれ個別のビンに割り振ると，ビン数が膨大となり，結果の解釈や計算自体が困難となることが懸念される．そのため，ビン数（もしくはビン幅）については，目的に応じて適切なレベルを設定する必要がある．

3.2 O ダミー変数

次に O ダミー変数とよばれるものを導入する．まず，通常のダミー変数の概要について説明する．

3.2.1 ダミー変数

カテゴリー型特徴量をモデルに反映するための基本的な手法として，ダミー変数化とよばれる手法がある．本研究では，ダミー変数化を順序型特徴量に適用して，順序型特徴量と応答変数との非線形な関係を捕捉することを考える．

例として，順序型特徴量Xについて，取りうる値をm個のレベル$\{1, 2, \cdots, m\}$に分割することを考える．この場合，ダミー変数化とは，各特徴量を下式で定義されるダミー変数$d_1, d_2, \cdots, d_m$に変換することを意味する．

$$d_j = \begin{cases} 1 & \text{if } X\text{が属するレベル} = j \quad (j \in \{1, 2, \cdots, m\}) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

---

[6] バギングおよびブースティングは決定木に限定した手法ではなく，そのほかの回帰モデルなどでも使用される手法である．

[7] 数値型特徴量を含む，順序を持つ特徴量に対して，本稿では「順序型特徴量」という用語を用いる（数値型も順序があるため）．

（例）特徴量がレベル$\{1,2,3,4\}$の値をとる場合

X	d_1	d_2	d_3	d_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

このダミー変数は，個々の特徴量の特徴を独立に表すため，パラメータの意味を解釈しやすいが，特徴量の順序関係という重要な情報を失うというデメリットもある．これを解消するため，本研究では次に述べる O ダミー変数を導入する．

3.2.2 O ダミー変数

O ダミー変数 (Ordinal Dummy Variables)[8]とは，特徴量を次のようにダミー変数化したものである．ここで，対象の特徴量Xは，m個のレベルに分割されており（前節の離散化による），属する特徴量の値の小さいレベルから，レベルを$\{1,2,\cdots,m\}$とする．

設定したレベルに基づき，m個のダミー変数$d_1,d_2,\cdots,d_m$を作る．各レコードのデータに対し，d_jは対象の特徴量の値が属するレベルが，jよりも小さければ 1 とし，j以上であれば 0 とする．これを数式で表すと次のようになる．

$$d_j=\begin{cases}1 & \text{if } X\text{の属するレベル} < j \quad (j\in\{1,2,\cdots,m\}) \\ 0 & \text{otherwise}\end{cases} \tag{6}$$

表 3.1 は，特徴量を年齢（10 歳～89 歳）とし，レベルを 10 歳刻みで設定（10 代をレベル 1，20 代をレベル 2，以降同様）した場合において，通常のダミー変数と O ダミー変数を比較したものである．

表 3.1 通常のダミー変数と O ダミー変数の比較例

・　通常のダミー変数の場合

年齢	レベル	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
35	3	0	0	1	0	0	0	0	0
45	4	0	0	0	1	0	0	0	0

・　O ダミー変数の場合

年齢	レベル	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
35	3	0	0	0	1	1	1	1	1
45	4	0	0	0	0	1	1	1	1

[8] 回帰分析の文脈では「split-coding」 [28]とよぶことがある．また，ニューラルネットワークの文脈では「Thermometer Encoding」 [25]とよぶこともある．

O ダミー変数に対して行う変数選択は，隣接するレベル間を切り離すか，統合するかの選択に直接対応する．たとえば，上記の例（表 3.1）でd_4を特徴量として採用するかしないかは，30 代と 40 代で年齢の説明力に定量的な差をつけるか否かを定めることに対応する．

すなわち O ダミー変数は，通常のダミー変数では失われてしまう順序関係を維持することができる．そのため，順序型特徴量（年齢や給与など）を離散化する場合など，同一変数内で順序関係があるときの影響をモデルに反映する場合に有効であることが期待される．アクチュアリーの実務ではしばしば，隣接するリスク群団間の関係（群団 A は隣接する群団 B に比べて何倍リスクが大きいか，など）に注目することがあるため，順序関係を踏まえた手法は望ましい．

3.3 正則化

最後に正則化とよばれる手法について述べる．

3.3.1 正則化の効果

正則化（Regularization）とは，モデルの目的関数について，誤差項にモデルの複雑さを表す正則化項を加え，これを最小化する手法を指す．これにより，以下の効果が期待できる（導入する正則化項の種類（後述）に応じて，得られる効果は異なる）．

- 変数選択（多数の特徴量から予測に有効なものを機械的に選択する）
- ロバスト性（特徴量同士に強い相関がある場合（特に，多重共線性が疑われる場合）でも計算結果を安定させる）
- 分散の減少（訓練データへの過学習を回避する）
- 狭義の正則化（特徴量の数がレコード数より大きい場合でもモデリングが可能である）

アクアチュリーが扱うモデルにおいても，上記の効果が得られることが望ましい．特に一つ目の変数選択の効果は，アクチュアリーもスパースな（特徴量は多いが，予測に有効なものは少ない）データを扱う場面が増えていることを踏まえると，有用であるといえる．すなわち正則化により，これまでアクチュアリーが個人の経験と試行錯誤により行ってきた変数の選択を，機械的かつ高速に行うことが期待できる．

さらに本研究の観点では，順序型特徴量について，前述の離散化および O ダミー変数化を施すことで，特徴量の数が多くなり，データへ過学習することが予想されるが，正則化を導入することで，これを抑えて予測に有効な特徴量を自動的に選択する効果が期待できる．

3.3.2 定式化

正則化を導入するためには，回帰モデルの係数$\boldsymbol{\beta}$を算出するための目的関数に，正則化項$p_\lambda(|\boldsymbol{\beta}|)$を追加する．線形モデルの場合，目的関数は次のようになる．

$$\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\right)^2+p_\lambda(|\boldsymbol{\beta}|) \tag{7}$$

正則化項を加えることで，残差（第一項）の最小化という条件を一定程度緩和し（バイアスを許し），代わりに係数$\boldsymbol{\beta}$を縮小する（0 になる場合，変数選択により説明変数から除外されたことを意味する）．

具体的な正則化の手法としては，Ridge [8]や Lasso [9] ，Elastic Net [10]が有名である．それぞれの正則化項は，以下のとおりである（切片項は正則化対象とならないため，jは 1 から始まる）．

$$
\begin{aligned}
&\text{Ridge} && p_\lambda(|\boldsymbol{\beta}|) = \lambda \sum_{j=1}^{p} |\beta_j|^2 \\
&\text{Lasso} && p_\lambda(|\boldsymbol{\beta}|) = \lambda \sum_{j=1}^{p} |\beta_j| \\
&\text{Elastic Net} && p_\lambda(|\boldsymbol{\beta}|) = \lambda_1 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|
\end{aligned}
\tag{8}
$$

係数λ（λ_1, λ_2）はハイパーパラメータとよばれ，非負の値として分析者が設定する．一般的には，λの値を変えた複数パターンのモデルを構築した上で，交差検証法（クロスバリデーション）とよばれる手法で，それぞれのモデルの性能を評価比較してλを決定する. なお, 上記の正則化の手法で変数選択が期待されるのは Lasso と Elastic Net である．

また GLM に正則化を施す場合は，最尤推定法に使用する対数尤度$l(\boldsymbol{\beta})$に正則化項を導入する．すなわちパラメータの推定は下式のように行う．

$$
\max_{\boldsymbol{\beta}} \{ l(\boldsymbol{\beta}) - p_\lambda(|\boldsymbol{\beta}|) \} \tag{9}
$$

3.4 Accurate GLM（AGLM）

3.4.1 AGLM の特長

冒頭でも述べたように，AGLM とは，前節までで取り上げた離散化，O ダミー変数，および正則化の三つの手法を，GLM に取り入れたモデリング手法である．三つのうち特に離散化と O ダミー変数は，基礎となる GLM の枠組み自体を変えるものではなく, むしろモデルの入力である特徴量を加工するものであることに注目されたい．これはデータサイエンス分野のモデリング過程における，いわゆるデータの前処理や，特徴量エンジニアリング，**探索的データ解析**（EDA: Exploratory Data Analysis）とよばれるステップに密接に関係する．後の章でも触れるが，これらに関連する手続きでモデルの改善が図れることは，アクチュアリアルモデリングとデータサイエンスの融合の可能性が示唆される一つの事項だと考える．

各手法の説明でそれぞれの効果を段階的に示したとおり，三つの手法は独立に効果を発揮する訳ではなく，三つが組み合わさることではじめてその効果を発揮する．三つの手法の効果については，後の数値実験でも確認する．GLM とそれぞれの一手法もしくは二手法の組み合わせに関する先行研究[9]，および線形モデルやロジスティック回帰モデル，比例ハザードモデルに三つすべての組み合わせを取り入れた研究 [11] [12]は存在するが，アクチュアリーが広く用いている GLM に三つすべてを応用したものは一般的には確認されていない．

AGLM についてまとめると，以下のとおりとなる．

- GLM をベースとして，順序型特徴量に対して「離散化」と「O ダミー変数」を組み合わせることで，未学習を回避しつつ，同一特徴量内での順序関係も考慮に入れた，表現力のある柔軟なモデリングを可能とする．
- さらに「正則化（特に Lasso など）」の導入により，上記の O ダミー変数を含めた特徴量に対して変数選択を行うことで，過学習を回避しつつ，汎化性能の高いモデルを実現する．実はこれは Fused Lasso という枠組みに相当している[10]．

---

[9] たとえば文献 [24]は GLM と正則化， [27]は回帰問題と離散化， [26]は回帰問題と O ダミー変数， [28]は O ダミー変数と正則化の組み合わせである． Fused Lasso との関係は，後述．

[10] ただし，Fused Lasso では，O ダミー化に相当する部分はユーザーには見えない形で実現されている．Fused Lasso その

このように，未学習および過学習を抑制したバランスのとれた平滑化モデリングにより，モデルの簡便さや説明可能性を向上させるだけでなく，予測精度の改善も図るものが AGLM である．

ここで例として，再び特徴量が年齢（10 歳～89 歳）の場合を考える．さきほどと同様に，離散化によりレベルを 10 歳刻みで設定し，O ダミー変数により順序関係が設定されたとする（表 3.2）．

表 3.2 O ダミー変数の例

年齢	レベル	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
15	1	0	1	1	1	1	1	1	1
25	2	0	0	1	1	1	1	1	1
35	3	0	0	0	1	1	1	1	1
45	4	0	0	0	0	1	1	1	1

そして正則化（たとえば Lasso）により，O ダミー変数の中から影響度が高いものを機械的に選定した結果，d_3, d_6, d_8が選択されたとする．これはすなわち 10 代 20 代のグループ，30 代から 50 代のグループ，60 代 70 代のグループ，および 80 代のグループに分割されたことを意味し，ある種のリスク区分が定量的に自動的に行われていることがわかる．

このように AGLM では，離散化，O ダミー変数および正則化を組み合わせることで，（ある種の目的関数を最小化するという意味で）ビンの結合の最適化が自動的に実現されるため，最適化する前のビン数やビン幅などの初期値の設定は，全般に十分に細かくなっていれば，どのような方法で決めても，原理的には最適化後の結果にほとんど違いはない．一方で予測の観点からは，より適切な離散化手法があり得る[11]．本稿では簡単のため，等距離法（Equal Width Method）とよばれるもっともシンプルな手法の一つを採用し，ビン数を極力細かくすることを考える．

3.4.2 既存モデルとの定性的比較

AGLM は，GLM に対して特徴量の非線形な動きをより柔軟に捉えられるよう枠組みを拡張したものであるので，従来の GLM より予測性能が高いことが見込まれる．

次に，AGLM にはリスク区分を自動的に行う特長があるため，特徴量がどこで分割されているかを把握できるという点で，決定木と同様の性質を有しているともいえる．そのため，モデリングするためのデータの前処理（モデルを走らせた結果を見てからの再吟味も含む）へも活用できる．

また，従来は膨大な計算量の懸念から，交互作用を導入する際，特徴量の組み合わせの選択に十分な吟味を要していたが，ある程度特徴量の数が多くなっても，正則化により特徴量は自動的に選別されるため，交互作用項の取捨選択プロセスを省略できるというメリットもある．

最後に，GAM と比較した場合，GAM の特長である（スプライン関数などによる）特徴量の非線形な動きを捉える平滑化は，AGLM の三つの手法で実現されている．一方で GAM には変数選択する能力はなく，前述のとおり GLM に比べて説明可能性の部分で劣後する．

以上により，定性的な観点から AGLM と既存手法を比較すると，AGLM は既存手法の特長を，別のわかりやすい形で再現している．次章ではその効果を確認するために，数値実験を通して定量的な比較を行う．

ものについては文献 [21]を参照されたい．

[11] たとえば文献 [23]などを参照されたい．

4 数値実験

この章では，AGLM の効果をさまざまな数値実験を通して定量的に示す．

4.1 数値実験の内容

これまで述べてきたように AGLM は，アクチュアリーの実務において，説明可能性および予測精度の観点から，有用であることが期待できる．一方で，すべてのデータや問題に対して最良の（予測性能が良い）モデルは存在しない（いわゆるノーフリーランチ）．そのため，アクチュアリーの実務で扱うようなデータを想定して，AGLM の効果を検証するため，以下の三つの数値実験を実施した．

① AGLM の三つの手法の個別効果と，同時に組み込んだ場合の効果の検証
既存の公開データ（以下，外部データ）の一つである自動車保険データを用いて，AGLM を構成する三つの手法（離散化，O ダミー変数，正則化）の個別の効果と，それらを同時に組み込んだ場合（すなわち AGLM）の効果について，予測誤差による比較を行うことで，予測精度（逸脱度，以下同様）の観点で，三つすべてを同時に組み込んだ AGLM が，最良の結果を与えることを検証する．

② 人工データを用いた生成モデルに対する各モデルの評価結果の比較検証
本研究のために作成したデータ（以下，人工データ）（自動車保険を想定したデータ）を用いて各候補モデルのあてはめを行い，人工データの生成に使用したモデルのパラメータと，あてはめた結果を比較することで，モデルのパラメータの再現度や説明可能性の観点から AGLM を評価する．

③ 複数の外部データを用いた各モデルの予測精度の比較検証
複数の外部データを用いて，各候補モデルのあてはめを行った上で，各モデル間で予測誤差を比較することで，予測精度の観点から AGLM を評価する．

ここでは，アクチュアリー実務のうちプライシングに焦点を当て，損害保険のプライシング手法の一つである，純保険料法（いわゆる FD 法（Frequency Damageability Method））をベースに数値実験を行う．これは頻度と損害規模のモデルをそれぞれ構築し，その期待値の積として純保険料を計算するものである．扱ったデータは，頻度（クレーム件数）および損害規模（クレーム単価）の両方（もしくは損害規模のみ）を有するものである．

4.2 使用したデータ

各数値実験で使用したデータについては，以下のとおり．いずれのデータについても，頻度および損害規模（ただし一部の外部データは損害規模のみ）が応答変数となっている．これらは，欠損値などは含まない，比較的綺麗なデータである．

① AGLM の三つの手法の個別効果と，同時に組み込んだ場合の効果の検証
表 4.1＜外部データ＞の Data 1 を使用

② 人工データを用いた生成モデルに対する各モデルの評価結果の比較
表 4.1＜人工データ＞Data 0 を使用

③ 複数の外部データを用いた各モデルの予測精度の比較
表 4.1＜外部データ＞の Data 1 から Data 8 を使用

表 4.1 使用したデータの内容と出典

	データの内容	出典
＜人工データ＞		
Data 0	想定自動車保険クレームデータ	本研究で作成（詳細は 4.4.2）
＜外部データ＞		
Data 1	自動車保険クレームデータ	Predictive Modeling vol.2 [13]の Chap.1「sim-modeling-dataset.csv」
Data 2	オートバイ保険クレームデータ	NL Ins Pricing with GLM [14] Case Study の「mccase.txt」 [15]
Data 3	自動車保険クレームデータ	MACQUARIE University [16]の「car.csv」
Data 4	対人賠償保険クレームデータ	R パッケージ CASdatasets [17]の「freMTPLfreq/sev」
Data 5	自動車保険クレームデータ	R パッケージ CASdatasets の「freMPL1」
Data 6	傷害保険クレームデータ	MACQUARIE University の「persinj.xls」
Data 7	自動車保険クレームデータ	kaggle の Automobile Dataset [18]
Data 8	医療費データ	kaggle の Medical Cost Personal Datasets [19]

各データにおけるレコード数（頻度および損害規模が有効なもの）は，表 4.2 のとおり．外部データの Data 6 から Data 8 については，損害規模のみのデータであるため，頻度のレコード数は NA となる．

なお，Data 2 のデータには Duration が 0 のレコード（すなわちエクスポージャが 0）が存在しているが，分析対象はあくまでもエクスポージャを有するものとしているため，これらのレコードについては事前に除外した上で，分析を行った．

表 4.2 各データのレコード数

	頻度 （全データ）	損害規模 （クレーム有のデータ）
Data 0	50,000	8,543
Data 1	40,760	3,169
Data 2	64,548	670
Data 3	67,856	4,624
Data 4	413,169	15,390
Data 5	30,595	3,265
Data 6	NA	22,036
Data 7	NA	164
Data 8	NA	1,338

4.3 各数値実験における前提条件

4.3.1 データの分割

「②人工データを用いた生成モデルに対する各モデルの評価結果の比較」の検証では全レコードを用いて分析する．一方で「①AGLM の三つの手法の個別効果と，同時に組み込んだ場合の効果の検証」と「③複数の外部データを用いた各モデルの予測精度の比較」については，予測精度の評価を行うため，全レコードの半分

（50%）を訓練データとして用いて，モデルのあてはめを行った上で，残りの半分のレコードをテストデータとして用いて検証を行い，各モデルの予測精度を検証した．

4.3.2 予測対象

各数値実験における予測対象は，以下のとおりとする．

① AGLM の三つの手法の個別効果と，同時に組み込んだ場合の効果
頻度

② 人工データを用いた生成モデルに対する各モデルの評価結果の比較
頻度および損害規模

③ 複数の外部データを用いた各モデルの予測精度の比較
頻度，損害規模，およびそれらを掛け合わせた総損害額（ただし 4.2 のとおり，Data 6 から Data 8 については，損害規模のデータのみしか存在しないため，それらについては予測対象を損害規模のみとする）

なお，上記③の総損害額については，たとえば Tweedie 分布による AGLM を用いて直接推定することも可能である．4.2 の＜外部データ＞の Data 1 に対してこの手法を用いた予測結果について，付録1に記載している．

4.3.3 比較対象とするモデル

比較対象とするモデルは，頻度および損害規模のいずれについても，以下のとおりとした（ただし「①AGLM の三つの手法の個別効果と，同時に組み込んだ場合の効果の検証」は検証対象が三つの手法に関連するパターンに限られるため，この限りではない）．

- GLM
- 正則化 GLM（Ridge/Lasso/Elastic Net）
- AGLM（正則化なし）
- AGLM（Ridge/Lasso/Elastic Net）
- GAM
- CART

いずれのモデルについても，頻度，損害規模はそれぞれポアソン分布，ガンマ分布を前提とし，リンク関数は対数リンク関数を採用した．予測精度の評価は，頻度はポアソン逸脱度，損害規模はガンマ逸脱度，そして総損害額は Tweedie 逸脱度を採用した（小さいほど良い成績であることを意味する）．なお，Tweedie 逸脱度について，Tweedie 分布の分散関数のべき指数pは，複合ポアソン・ガンマ分布に相当する場合の，特に$p = 1.5$を採用した．また，レコード数が異なる場合でも予測精度を比較できるように，各逸脱度の数値はレコード数で除した平均値としている．

各候補モデルの具体的な仕様は，表 4.3 のとおりとする．「Full」は正則化を行わない手法を，「Penal」は正則化を行う手法を指す．またαは Elastic Net のパラメータであり，$\alpha = 0$は Ridge を，$\alpha = 1$は Lasso をそれぞれ表す．数値実験では，Ridge と Lasso の間に位置する$\alpha = 0.5$の Elastic Net も候補とした．

なお，「AGLM/Full」は正則化なしの AGLM となり，正確には AGLM ではない点に注意されたい（AGLM は離散化，O ダミー変数，正則化を同時に組み込んだモデルを指すため）．以降の節において各モデルに言及する際は，表 4.3 に従い，Freq①から Freq⑩，および Sev①から Sev⑩と表記する．

表 4.3 候補モデルの仕様

	頻度（Freq）/損害規模（Sev）
①	GLM/Full
②	GLM/Penal/$\alpha = 0$
③	GLM/Penal/$\alpha = 0.5$
④	GLM/Penal/$\alpha = 1$
⑤	AGLM/Full
⑥	AGLM/Penal/$\alpha = 0$
⑦	AGLM/Penal/$\alpha = 0.5$
⑧	AGLM/Penal/$\alpha = 1$
⑨	GAM/Full
⑩	CART/Full

4.3.4 そのほかの前提条件

AGLM において，離散化や O ダミー変数化を行う際，本数値実験では使用するデータの各特徴量（特にカテゴリー型特徴量）の取る値を既知としてモデリングする[12]．ここで離散化する手法として，等距離法を採用する（ビン数はmin(順序型特徴量のレベルの数, 100)により定める）．

モデルの特徴量について，裾が長いと判断した特徴量については，適宜対数変換を施した．一方で，データから新たな特徴量生成は行わず，データに含まれる特徴量をそのままモデルに用いた．つまり，各特徴量の高次項や交互作用項は導入せず，それぞれの特徴量が単独に応答変数に影響するものとして，モデル化している．ただし交互作用については，4.2 の＜外部データ＞Data 1 と Data 2 についてのみ，数値的な検証を別途実施した．その概要は付録 2 に記載している．

また，正則化項のハイパーパラメータについては，交差検証法によって各逸脱度が最小となるものを選択した．

4.4 数値実験の結果と考察

4.4.1 AGLM の三つの手法の個別効果と，同時に組み込んだ場合の効果

本節では，AGLM を構成する三つの手法（離散化，O ダミー変数，正則化）の個別の効果と，それらを同時に組み込んだ場合（すなわち AGLM）の効果の検証を行う．具体的には，AGLM の有効性（三つの手法を同時に導入）を確認するため，4.2 の＜外部データ＞Data 1 の頻度の予測を例に，以下の三つのパターンのブレイクダウンを行った．

- パターン 1　GLM＋離散化＋O ダミー変数＋正則化の順
- パターン 2　GLM＋正則化＋離散化＋O ダミー変数の順
- パターン 3　GLM＋離散化＋正則化＋O ダミー変数の順

テストデータに対するポアソン逸脱度の比較を行った結果，以下の表 4.4 のとおり，各要素を独立に導入した場合，および各要素を二つ組み合わせた場合よりも，三つの手法を同時に導入する AGLM の方が優位な結果となることが確認できた（ここで，正則化項は Lasso を採用した）．なお，パターン 1 の「左記＋O ダミー変数」

[12] 本来，テストデータは未知であり，モデリングに使用しなかった特徴量に対する処理方法を事前に検討しておく必要がある（たとえば「others」という変数を用意するなど）．

でポアソン逸脱度が非常に大きい値となっているが，これは離散化と O ダミー変数により訓練データへの過学習を引き起こしているためと考えられる（最終的には正則化により緩和されている）．

表 4.4 Data 1 に対する各パターンのポアソン逸脱度

				AGLM
パターン1	GLM	左記＋離散化	左記＋Oダミー変数	左記＋正則化（Lasso）
ポアソン逸脱度	0.39128	0.39684	699147833.77139	0.39037

パターン2	GLM	左記＋正則化（Lasso）	左記＋離散化	左記＋Oダミー変数
ポアソン逸脱度	0.39128	0.39063	0.39553	0.39037

パターン3	GLM	左記＋離散化	左記＋正則化（Lasso）	左記＋Oダミー変数
ポアソン逸脱度	0.39128	0.39684	0.39553	0.39037

4.4.2 人工データを用いた生成モデルに対する各モデルの評価結果の比較

本節では，人工データの生成モデルに対する，各候補モデルの再現性確認を行う．具体的には，頻度および損害規模について，それぞれポアソン分布およびガンマ分布を仮定して人工データを生成し，それらに対して表 4.3 の各モデル[13]をあてはめて，人工データで仮定した分布のパラメータと，当該あてはめた結果の比較を行い，各候補モデルについて人工データの生成モデルに対する再現性を検証した．表 4.5 に人工データの概要をまとめる．

表 4.5　人工データの概要

変数名		内容	変数の種類
特徴量	gender	性別（Male/Female）	カテゴリー型
	area	地域（Urban/Rural）	カテゴリー型
	age	年齢（20 歳から 79 歳）	数値型（O ダミー変数化対象）
応答変数	freq	事故件数	
	sev	損害額	

なお，人工データの生成モデルの前提（頻度および損害規模）は以下のとおりである．

- 頻度
  ポアソン分布（パラメータ: λ）を仮定し，$\lambda = 0.16$を基準とした上で，それぞれの区分（性別/地域/年齢）に応じてλを調整した．具体的には表 4.6 の λ のとおり．たとえば Male/Urban/25 歳のλは，$0.22(= 0.16 + 0.02 + 0.02 + 0.02)$となる．
- 損害規模
  ガンマ分布（パラメータ: α, β）を仮定し，$(\alpha, \beta) = (1, 0.02)$を基準とした上で，それぞれの区分（性別/地域/年齢）に応じてαのみ調整した（βは固定）．具体的には表 4.6 のαのとおり．

[13] ただし，AGLM/Full は正則化を導入していないモデルであり，本来的に AGLM ではないことから，本検証では候補から除外した．

表 4.6 人工データの生成モデルのパラメータおよび期待値

			Freq ~Poisson(λ)		Sev ~Gamma(α, β)		
			λ	期待値 λ	α	β	期待値 α/β
Base			0.16	0.16	1		50
gender	性別：Male/Female	M	+ 0.02	+ 0.02	+ 0		+ 0
		F	+ 0	+ 0	+ 0		+ 0
area	地域：Urban/Rural	U	+ 0.02	+ 0.02	+ 0.20		+ 10
		R	+ 0	+ 0	+ 0		+ 0
age	年齢：20-79歳	20-29	+ 0.02	+ 0.02	+ 0	0.02	+ 0
		30-69	+ 0	+ 0	+ 0		+ 0
		40-49	+ 0	+ 0	+ 0.2		+ 10
		50-59	+ 0	+ 0	+ 0.2		+ 10
		60-69	+ 0	+ 0	+ 0		+ 0
		70-79	+ 0.02	+ 0.02	+ 0		+ 0
データ数			50,000		8,543		

頻度について，人工データを用いて各候補モデルのあてはめを行い，真のモデルから導かれる各区分の頻度と，各モデルから得られる平均頻度との比較を行った結果，表 4.7 のとおりとなった（ここではパラメータの再現性を確認することが目的のため，モデルのあてはめについては，全データを用いた（すなわちテストデータ検証は行っていない））．また，表 4.7 に基づき，それぞれの区分の差異を集計した結果は，表 4.8 のとおり（たとえば性別の差異については，性別以外（地域/年齢）が同条件のときの男女の差分値をとり，その値を地域/年齢の全体で平均したものを記載している．他の区分の差異についても同様）．

表 4.8 より，AGLM および GAM は，他のモデルと対比して，それぞれの区分間の差異を捕捉できていることが分かる（特に順序型特徴量である（O ダミー変数化の対象となる）年齢の差異について，捕捉できているところがポイントである）．さらに，モデルをあてはめた結果の逸脱度も踏まえると，「AGLM w/ Ridge ($\alpha = 0$)」が最良という結果が得られた（ただし，これは全データを用いてモデルのあてはめを行った結果に対する逸脱度であり，必ずしも予測精度が最良とは言えない）．

損害規模についても同様に，各候補モデルのあてはめを行い，真のモデルから導かれる各区分の損害規模と，各モデルから得られる平均損害規模との比較を行った結果，表 4.9 のとおりとなった．また，表 4.9 に基づき，それぞれの区分の差異を集計した結果は表 4.10 のとおり（計算方法は頻度の場合と同様）．

表 4.10 により，AGLM および GAM は，他のモデルと対比して，それぞれの区分間の差異を捕捉できていることが分かる（頻度と同様，特に順序型特徴量である年齢の差異がポイント）．さらに，モデルをあてはめた結果の逸脱度も踏まえると，「AGLM w/ Lasso ($\alpha = 1$)」が最良という結果が得られた（ただし，これも頻度と同様で，必ずしも予測精度が最良とは言えない点には注意）．

頻度と損害規模いずれについても，AGLM および GAM ではリンク関数について対数リンク関数を採用しているため，加法的な仮定をおいている人工データの生成モデルを完全に再現することはできないが，それでも一定程度の再現力があることが分かる．

表 4.7 人工データ（頻度）に対するモデルのあてはめの結果（その 1）

頻度			TRUE	GLM	GLM＋正則化			AGLM			その他	
性別	地域	年齢		Freq①	Freq② Ridge	Freq③ Enet (α:0.5)	Freq④ Lasso	Freq⑥ Ridge	Freq⑦ Enet (α:0.5)	Freq⑧ Lasso	Freq⑨ GAM	Freq⑩ CART
Female	Rural	20-29	0.180	0.172	0.173	0.178	0.178	0.181	0.178	0.178	0.177	0.188
		30-69	0.160	0.170	0.171	0.178	0.178	0.170	0.172	0.172	0.167	0.188
		70-79	0.180	0.168	0.169	0.178	0.178	0.179	0.178	0.178	0.175	0.188
	Urban	20-29	0.200	0.194	0.194	0.193	0.193	0.199	0.197	0.198	0.200	0.188
		30-69	0.180	0.192	0.192	0.193	0.193	0.187	0.191	0.191	0.189	0.188
		70-79	0.200	0.190	0.190	0.193	0.193	0.197	0.197	0.197	0.198	0.188
Male	Rural	20-29	0.200	0.180	0.181	0.180	0.180	0.189	0.184	0.184	0.186	0.188
		30-69	0.180	0.178	0.179	0.180	0.180	0.178	0.177	0.177	0.175	0.188
		70-79	0.200	0.176	0.177	0.180	0.180	0.187	0.183	0.183	0.184	0.188
	Urban	20-29	0.220	0.204	0.203	0.194	0.194	0.207	0.203	0.203	0.209	0.188
		30-69	0.200	0.201	0.201	0.194	0.194	0.195	0.196	0.196	0.198	0.188
		70-79	0.220	0.199	0.199	0.194	0.194	0.206	0.202	0.202	0.207	0.188

表 4.8 人工データ（頻度）に対するモデルのあてはめの結果（その 2）

頻度			Freq ~Poisson(λ)	
			λ	期待値 λ
Base			0.16	0.16
gender	性別：Male/Female	Male	＋0.02	＋0.02
		Female	＋0	＋0
area	地域：Urban/Rural	Urban	＋0.02	＋0.02
		Rural	＋0	＋0
age	年齢：20-79歳	20-29	＋0.02	＋0.02
		30-69	＋0	＋0
		70-79	＋0.02	＋0.02

GLM	GLM＋正則化			AGLM			その他	
Freq①	Freq② Ridge	Freq③ Enet (α:0.5)	Freq④ Lasso	Freq⑥ Ridge	Freq⑦ Enet (α:0.5)	Freq⑧ Lasso	Freq⑨ GAM	Freq⑩ CART
＋0.01	＋0.01	＋0.00	＋0.00	＋0.01	＋0.01	＋0.01	＋0.01	▲ 0.00
＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0
＋0.02	＋0.02	＋0.01	＋0.01	＋0.02	＋0.02	＋0.02	＋0.02	▲ 0.00
＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0
＋0.00	＋0.00	＋0.00	＋0.00	＋0.01	＋0.01	＋0.01	＋0.01	＋0.00
＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0
▲ 0.00	▲ 0.00	＋0.00	＋0.00	＋0.01	＋0.01	＋0.01	＋0.01	＋0.00

ポアソン逸脱度	0.67402	0.67402	0.67423	0.67423	0.67281	0.67350	0.67379	0.67379	0.67380

表 4.9 人工データ（損害規模）に対するモデルのあてはめの結果（その 1）

損害規模			TRUE	GLM	GLM＋正則化			AGLM			その他	
性別	地域	年齢		Sev①	Sev② Ridge	Sev③ Enet (α:0.5)	Sev④ Lasso	Sev⑥ Ridge	Sev⑦ Enet (α:0.5)	Sev⑧ Lasso	Sev⑨ GAM	Sev⑩ CART
Female	Rural	Others	50	55	55	55	55	52	51	51	51	60
		40-59	50	55	55	55	55	62	61	61	61	60
	Urban	Others	60	64	64	63	63	59	59	59	60	60
		40-59	60	64	64	63	63	70	70	71	71	60
Male	Rural	Others	50	54	54	55	55	52	51	51	51	60
		40-59	50	54	54	55	55	62	61	61	60	60
	Urban	Others	60	63	63	63	63	59	59	59	59	60
		40-59	60	63	63	63	63	70	71	71	70	60

表 4.10 人工データ（損害規模）に対するモデルのあてはめの結果（その 2）

損害規模			Sev ~Gamma(α, β)			GLM	GLM＋正則化			AGLM			その他	
			α	β	期待値 α/β	Sev①	Sev② Ridge	Sev③ Enet (α:0.5)	Sev④ Lasso	Sev⑥ Ridge	Sev⑦ Enet (α:0.5)	Sev⑧ Lasso	Sev⑨ GAM	Sev⑩ CART
Base			1		50									
gender	性別：Male/Female	Male	＋0		＋0	▲ 1.0	▲ 0.9	▲ 0.0	＋0.0	▲ 0.4	▲ 0.0	▲ 0.0	▲ 1.0	＋0.0
		Female	＋0		＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0
area	地域：Urban/Rural	Urban	＋0.2	0.02	＋10	＋9.2	＋8.9	＋7.2	＋7.2	＋7.7	＋8.4	＋8.5	＋9.5	＋0.0
		Rural	＋0		＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0
age	年齢：20-79歳	Others	＋0		＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0	＋0
		40-59	＋0.2		＋10	▲ 0.0	▲ 0.0	＋0.0	▲ 0.0	＋10.6	＋10.7	＋10.8	＋10.3	▲ 0.0
ガンマ逸脱度						0.90130	0.90130	0.90158	0.90157	0.89096	0.89097	0.89096	0.89212	0.90677

4.4.3 複数の外部データを用いた各モデルの予測精度の比較

本節では，外部データ（4.2 を参照）を用いて，頻度，損害規模，およびそれらを掛け合わせた総損害額の三つ（ただし 4.2 のとおり，Data 6 から Data 8 については損害規模データのみしか存在しないため，それらについては予測対象を損害規模のみとする）を予測対象として，各モデルの逸脱度を計算し，予測精度の観点からモデルを評価する．

頻度の予測については表 4.11 のとおり，「AGLM w/Elastic Net ($\alpha = 0.5$)もしくは Lasso ($\alpha = 1$)」（表 4.11 の Freq⑦と Freq⑧に対応）が，各データに対する予測誤差の平均順位[14]の観点で，予測結果が良いモデルという結論となった．

表 4.11 外部データ（頻度）に対する各モデルの予測精度

頻度モデル		Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8	平均順位
Freq①	GLM/full	7	8	8	10	NA	頻度データが存在しないため対象外			8.3
Freq②	GLM/Penal/α=0	5	7	4	9	1				5.2
Freq③	GLM/Penal/α=0.5	3	6	5	8	4				5.2
Freq④	GLM/Penal/α=1	4	5	6	7	3				5.0
Freq⑤	AGLM/full	10	10	9	5	NA				8.5
Freq⑥	AGLM/Penal/α=0	8	3	1	4	2				3.6
Freq⑦	AGLM/Penal/α=0.5	1	2	3	2	6				2.8
Freq⑧	AGLM/Penal/α=1	2	4	2	3	5				3.2
Freq⑨	GAM/full	6	1	7	1	8				4.6
Freq⑩	CART/full	9	9	10	6	7				8.2

テストデータに対してポアソン逸脱度が最小となるモデルを 1 位，最大となるモデルを 10 位とした．なお，NA はモデルが回らなかった（結果が発散した）ケースを表す．Data 6 から Data 8 に関しては，頻度データが存在しないため，空白となっている．

AGLM と競合する GLM との比較のため，GLM と AGLM の差異となる「正則化項の有無」と「AGLM（離散化＋O ダミー変数＋正則化）の導入の有無」の各要素について，予測精度の観点から評価を行なった結果は表 4.12 のとおり．

表 4.12 予測精度向上（頻度）に対する AGLM の各要素の評価

項目	評価
正則化項	Freq①と Freq②から④，もしくは Freq⑤と Freq⑥から⑧を比較すると，正則化項を導入した場合の方が，導入しない場合と比べて，平均順位が高い結果となっていることがわかる．
AGLM	Freq②から④と Freq⑥から⑧を比較すると，正則化項のみを導入したモデルに比べて AGLM の方が，平均順位が高い結果となっていることがわかる．

[14] 単純平均としている．

損害規模の予測については表 4.13 のとおり，「GLM もしくは AGLM w/ Ridge ($\alpha = 0$)」（表 4.13 の Sev②と⑥に対応）が，平均的に予測精度が良いモデルとなった.

表 4.13 外部データ（損害規模）に対する各モデルの予測精度

損害規模モデル		Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8	平均順位
Sev①	GLM/full	8	9	7	9	8	5	8	5	7.4
Sev②	GLM/Penal/α=0	1	6	1	7	2	4	3	1	3.1
Sev③	GLM/Penal/α=0.5	4	5	2	3	3	7	6	3	4.1
Sev④	GLM/Penal/α=1	5	4	2	3	3	6	2	4	3.6
Sev⑤	AGLM/full	10	10	10	10	10	3	9	10	9.0
Sev⑥	AGLM/Penal/α=0	7	1	6	2	1	2	1	7	3.4
Sev⑦	AGLM/Penal/α=0.5	3	8	2	3	3	9	5	6	4.9
Sev⑧	AGLM/Penal/α=1	2	3	2	3	3	9	4	9	4.4
Sev⑨	GAM/full	9	2	9	1	9	1	NA	2	4.7
Sev⑩	CART/full	6	7	8	8	7	8	7	8	7.4

テストデータに対してガンマ逸脱度が最小となるモデルを 1 位，最大となるモデルを 10 位とした．なお，NA はモデルが回らなかった（結果が発散した）ケースを表す.

頻度と同様に「正則化項の有無」と「AGLM（離散化＋O ダミー変数＋正則化）の導入の有無」の各要素について，予測精度の観点から評価を行なった結果は，表 4.14 のとおり．AGLM が GLM 対比で必ずしも優れているとはいえない結果となったが，これはたとえば，以下が要因として考えられる.

- クレームが発生しているレコードが少なく，モデル間の優劣を評価できるほど，予測結果が安定していない可能性がある.
- AGLM では順序型特徴量についても非線形な動きを捕捉することができ，モデルの表現力が高い反面，外れ値に対する頑健性に課題がある可能性がある．この場合は，後述の「AGLM への通常の（離散化＋O ダミー変数化前の）数値型特徴量の追加」により，AGLM のクラスを従来の GLM を包含するようにすることで，問題を解消できる可能性がある.
- AGLM が，他のモデル対比で高いパフォーマンスを発揮すると考えられるのは，「順序型特徴量が応答変数に対して非線形な動きを持つ場合」であるが，損害規模の予測では，通常の名義特徴量の説明力が大きく，AGLM の有効性が十分に発揮できていない可能性がある．この場合に関しても，後述の「AGLM への通常の数値型特徴量の追加」により，AGLM の表現力をさらに向上させることで，問題を解消できる可能性がある.
- 本稿ではモデルそのものの評価に重点を置いており，事前の EDA が十分でない可能性がある.

表 4.14 予測精度向上（損害規模）に対する AGLM の各要素の評価

項目	評価
正則化項	Sev①と Sev②から④，もしくは Sev⑤と Sev⑥から⑧を比較すると，正則化項を導入した場合の方が，導入しない場合と比べて，平均的に良い予測結果となっていることがわかる.
AGLM	Sev②から④と Sev⑥から⑧を比較すると，正則化項のみを導入したモデルに比べて AGLM の方が，必ずしも良い予測結果が得られるとはいえない結果となっている.

総損害額の予測については，上記の頻度と損害規模の推定結果を掛け合わせて計算する．頻度および損害規模と同様に，予測精度の観点から比較を行った結果は，表 4.15 のとおり（計 100 通り（＝10 通り×10 通り）のうち，逸脱度の小さい上位 15 個の組み合わせについて結果を記載している）.

表 4.15 外部データ（総損害額）に対する各モデルの予測精度

総損害額モデル（上位15個）	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8	平均順位
Freq⑦*Sev⑧	19	4	26	24	26	頻度データが存在しないため対象外			19.8
Freq⑧*Sev⑥	37	11	39	17	5				21.8
Freq⑧*Sev⑧	17	3	30	29	30				21.8
Freq⑦*Sev⑥	42	12	38	16	4				22.4
Freq⑦*Sev②	3	32	44	28	9				23.2
Freq⑦*Sev⑦	20	21	26	24	26				23.4
Freq⑧*Sev②	2	30	45	33	11				24.2
Freq⑥*Sev⑧	51	7	34	18	13				24.6
Freq⑦*Sev③	28	20	26	24	26				24.8
Freq⑩*Sev⑥	75	2	3	23	21				24.8
Freq⑧*Sev⑦	18	19	30	29	30				25.2
Freq⑦*Sev④	29	23	26	24	26				25.6
Freq⑧*Sev③	21	18	30	29	30				25.6
Freq⑧*Sev④	22	22	30	29	30				26.6
Freq⑥*Sev⑥	56	27	41	15	2				28.2

テストデータに対して Tweedie 逸脱度が最小となるモデルを 1 位, 最大となるモデルを 100 位とした.
Data 6 から Data 8 に関しては，頻度データが存在しないため，空白となっている.

総損害額では頻度は「AGLM w/ ENET ($\alpha = 0.5$)もしくは Lasso ($\alpha = 1$)」 （表 4.15 の Freq⑦と⑧に対応）のモデル，損害規模は「AGLM w/ Ridge ($\alpha = 0$)もしくは Lasso ($\alpha = 1$)」（表 4.15 の Sev⑥と⑧に対応）が，平均的に予測精度が良いモデルであるという結果となった.

5 結語と今後の課題

5.1 結語

本稿では，アクチュアリー業務において広く使われている GLM に対して，データサイエンス分野の技術を取り入れた新たな手法である AGLM を提案し，その有用性を定性的かつ定量的に示すことで，アクチュアリーにおけるデータサイエンスの活用可能性の一例を示した.

順序型特徴量に O ダミー変数化を行い，順序型特徴量の非線形な効果をモデルに追加するとともに，特徴量が増えることで生じる過学習の可能性を，正則化を組み合わせて自動的に緩和することで，AGLM は，わかりやすく予測性能の高いモデルを生成できる手法であることを示した．

外部データによる数値実験を通じて，予測精度や汎用性の面で，AGLM はその基礎となる GLM だけではなく，他の既存手法（GAM および CART）よりも優れていることを示した．また人工データに対する数値実験では，人工データの生成モデルに対する AGLM の再現能力を検証することで，各区分間の較差を適切かつ，説明可能な形で捕捉できることを示した．

AGLM は，基礎となる GLM の枠組み自体を変えるものではなく，三つの手法（離散化，O ダミー化，正則化）により，特徴量の処理方法を工夫することで実現できる手法である．データサイエンス技術で前提となる特徴量に対する処理を，アクチュアリー業務で多用する GLM に応用して改善を図ることで，既存手法を起点としたデータサイエンスの活用可能性を例示することができた．

5.2 今後の課題

今後の課題としては，以下の点が挙げられる．

- 離散化手法の見直し

  AGLM では順序型特徴量の離散化手法として等距離法（ビン数はmin(順序型特徴量のレベルの数, 100)）を採用しているが，その他のビン数の適用（たとえば，ビン数もハイパーパラメータとみなして交差検証法により決定するといった，ビン数の決定プロセス自体の見直しも含む）や，他の離散化手法についても検討する余地がある．

  ただし AGLM については，本質的には単純にビン数を多くすればするほど表現力が向上する一方，正則化が有効にはたらいて不要なビンを自動的に除外することが期待されるため，単純な予測精度の比較だけではなく，たとえばデータ容量と計算時間の関係や，得られた結果の説明可能性など，実務面も踏まえた観点から評価を行うことが重要である．

- 損害規模の予測結果に関する追加の考察

  頻度の予測については，AGLM は，その基礎となる GLM と対比して，予測精度の観点からも優位性を示すことができた．一方で，損害規模の予測については，必ずしも AGLM の方が優れているとはいえない結果となった．4.4.3 にいくつかの考えられる要因を挙げたが，これらを追究することで，AGLM のさらなる改善に繋げられる可能性がある．

- 問題の性質に応じた効果的な正則化項（Ridge や Lasso など）の検証

  AGLM 同士での比較では，外部データと人工データの数値実験の結果に関して，特に損害規模の予測において Ridge と Lasso の優劣が整合していないため，この要因についても検証する必要がある．たとえば，外部データについて異常値に対する対応が十分にできておらず，正しくモデルを評価できていない可能性があるため，EDA を追加で行うなどの精査も検討する．

  また頻度の予測や損害規模の予測といった，応答変数間での正則化項の優劣についても分析し，どのような場合にどのような正則化が有効なのかを考察することで，与えられた問題に AGLM を適用する際の最適な正則化項を，事前に特定することができるかもしれない．

- モデル評価への交差検証法（クロスバリデーション）の活用

  今回はモデルの評価について，全データの半分（50%）を訓練データ，残りの半分をテストデータとするホールドアウト法を採用したが，データを複数に分割し，いわばホールドアウト法を繰り返し行うことにより安定的な評価が期待できる k 分割交差検証法の活用も検討する．

・　AGLM への通常の数値型特徴量の追加

AGLM への順序型特徴量の反映は，現在はすべて O ダミー変数化したもののみとしているが，なんの変換も施さない通常の数値型特徴量も合わせてモデルに反映して，AGLM のクラスが GLM を包含するものとすることで，GLM 対比でのさらなるモデルの改善が図れる可能性がある．

・　未知の入力データに対する予測の実務的な対応方法の検討

本稿における数値実験では，訓練データだけではなくテストデータも用いた全データに基づいて，離散化や O ダミー変数化を行なっている．一方で現実の予測問題では，訓練データでは想定しないようなデータについても対応しなくてはならず，数値型特徴量であれば，値の外挿方法，名義特徴量であれば，新たなカテゴリーに関する評価方法，などについて検討する必要がある．

・　EDA 手法の見直し

本研究は，従来アクチュアリーが行ってきた分析について，「データサイエンスの手法により自動化または簡易化する方法を模索する」ということが一つの目的であることも踏まえて，現時点の EDA や諸前提については，簡単な One-way 分析，一部の特徴量の対数変換，等距離法におけるビン数の固定運用など，簡便なものに留めていた．

ただし Two-way 分析などを通じて，分布形状の多峰性などのデータ特性の詳細まで踏み込んだ分析を行うことも考えられるため，分析者のバイアスが生じぬよう簡便さにも配慮しながら，手法の改善に繋がる EDA プロセスの確立に向けた検証を行う余地がある．

AGLM は，本稿を通して説明したとおり，「アクチュアリーとしてどのような手法が求められているか」という観点を踏まえ，アクチュアリー業務で広く使われる GLM を出発点として，データサイエンス技術の活用を試みたモデルであり，アクチュアリアルモデリングとデータサイエンスの融合可能性を示すことを意図したものである．本研究が，多くのアクチュアリーにとってデータサイエンス技術を取り入れる上での大きなヒントになれば幸いである．

謝辞

本研究のために有益なサポートを頂いた，公益社団法人日本アクチュアリー会 ASTIN 関連研究会の以下の方々（五十音順）に感謝いたします．

加藤 奈々 氏　　小島 睦月 氏　　後藤 陽介 氏　　近藤 健司 氏　　齋藤 知輝 氏
佐野 誠一郎 氏　　関口 健太郎 氏　　新居 悠輝 氏　　平松 雄司 氏　　向田 裕人 氏
渡辺 重男 氏

以上の方々には，研究の進め方への助言，既存手法の調査ならびに検証用データの整備に関してご協力を頂きました．

付録 1 Tweedie 分布による総損害額の予測

ここでは, Tweedie 分布を用いた AGLM により総損害額を予測した場合の効果を検証するため, R の「h2o」パッケージ [20]を用いて，以下のパターンに基づき総損害額の予測を実施した（表 A1.1）.

表 A1.1 Tweedie 分布を用いた AGLM のパターン

	総損害額（Aggregate Loss）
Twd①	AGLM/Penal/α = 0/VarPower= 1.5/LinkPower= 0
Twd②	AGLM/Penal/α = 0.5/VarPower= 1.5/LinkPower= 0
Twd③	AGLM/Penal/α = 1/VarPower= 1.5/LinkPower= 0

ここで VarPower は，Tweedie 分布の分散関数のべき指数pである（たとえば $p = 1$のときはポアソン分布, $p = 2$のときはガンマ分布を表す）．ここでは，総損害額モデルでよく用いられる複合ポアソン・ガンマ分布の場合の，特に$p = 1.5$を採用した[15]．また LinkPower は，リンク関数を表すパラメータ[16]であり，たとえば LinkPower = 0のときは対数リンク関数を，LinkPower = 1のときは恒等リンク関数を表す．ここでは対数リンク関数を採用した.

「4.4.3 の分析と同様の候補モデル（Freq①から⑩×Sev①から⑩）」＋「上記の三つの Tweedie 分布の AGLM」の計 103 通りのモデルに基づき，Data 1 を例に，テストデータに対する逸脱度（$p = 1.5$の Tweedie 逸脱度．以下同様）を比較した結果，表 A1.2 のとおりとなった（1 位は逸脱度最小，103 位は逸脱度最大を表す）.

結果，Tweedie 分布による AGLM については，全モデルの組み合わせの中で，50 位から 70 位程度となり，頻度と損害規模を別々に予測するモデルと対比して，予測精度の観点から優位性は確認できなかった.

表 A1.2 Tweedie 分布による総損害額の予測結果

総損害額モデル	Data1
Freq①*Sev⑩	1
Freq⑧*Sev②	2
Freq⑦*Sev②	3
Freq②*Sev⑩	4
Freq③*Sev⑩	5
:	:
Twd③	50
Twd②	51
:	:
Twd①	73
:	:
Freq③*Sev⑤	103

テストデータに対してポアソン逸脱度が最小となるモデルを 1 位, 最大となるモデルを 103 位とした.

[15] 本来，pは事前にテストしておくべきハイパーパラメータであるが，本稿では割愛する.

[16] 具体的には，LinkPower = lとしたとき，

$$g(\mu) = \begin{cases} \mu^l & l \neq 0 \\ \ln \mu & l = 0 \end{cases}$$

付録 2 交互作用項追加による効果の検証

ここでは，交互作用項の追加による予測精度への効果を確認する目的で，4.2＜外部データ＞の Data 1（「性別/年齢」の一つのパターン）と Data 2（「性別/年齢」と「地域/年齢」の二つのパターン）を例に分析を行った．

ここでは，AGLM における交互作用項追加の効果を検証することが目的であるため，O ダミー変数の要素が含まれた交互作用項を構成するべく，いずれのパターンにおいても，O ダミー変数化する「年齢」を，交互作用項の対象に含めている．従来のモデルであれば，離散化した順序型特徴量に対して交互作用項を導入すると，組み合わせの数が爆発的に増えてしまい，過学習（オーバーフィッティング）することが容易に想定されるが，これが正則化によってどの程度抑制され，さらには AGLM の改善可能性があるのかを検証することがねらいである．

表 A2.1 は，交互作用項の有無の両パターンについて，モデルをあてはめた結果（訓練データに対する逸脱度）およびテストデータに対する逸脱度を比較したものである（色をつけたものが，同一のモデル内で，逸脱度が小さかった方を表す）．

モデルのあてはめにおける逸脱度の比較から，交互作用項の導入により，多くの場合，モデルのあてはめに用いた訓練データに対する逸脱度が向上していることが分かる（ただしこれは後述のとおり，過学習を意味する）．

一方で，テストデータに対する逸脱度によると，多くの場合で交互作用項がない方が良い結果を与えているため，交互作用項ありの場合は，なしの場合と対比して，訓練データに対して過学習しているといえる．

ただし，Data 2 の②交互作用項あり（地域/年齢）については，テストデータの検証において，交互作用項なしよりも良い予測精度を与えるケースもいくつか得られたため，EDA などを通じて交互作用項を適切に定めれば，交互作用項なし対比でさらに良い予測精度が得られる可能性がある．

表 A2.1 交互作用項が予測精度に与える影響

訓練データへのあてはめ			Data1			Data2				
			A 交互作用あり (性別/年齢)	B 交互作用なし	A - B	A 交互作用あり (性別/年齢)	B 交互作用あり (地域/年齢)	C 交互作用なし	A - C	B - C
頻度	⑥	AGLM/Penal/α=0	0.38216	0.38300	-0.00084	0.08934	0.08907	0.09050	-0.00116	-0.00142
	⑦	AGLM/Penal/α=0.5	0.38553	0.38530	0.00023	0.08988	0.09031	0.09041	-0.00053	-0.00010
	⑧	AGLM/Penal/α=1	0.38535	0.38514	0.00021	0.08965	0.09020	0.09044	-0.00079	-0.00023
損害規模	⑥	AGLM/Penal/α=0	1.98658	1.99462	-0.00804	1.36413	1.46746	1.48505	-0.12092	-0.01759
	⑦	AGLM/Penal/α=0.5	2.05511	2.05536	-0.00025	2.00815	2.00975	2.00975	-0.00160	0.00000
	⑧	AGLM/Penal/α=1	2.05485	2.05516	-0.00030	1.83596	1.76212	1.76385	0.07211	-0.00173
総損害額	Freq⑦*Sev⑥		44.59357	44.74592	-0.15235	89.74707	91.39521	101.50363	-11.75657	-10.10843
	Freq⑦*Sev⑦		45.14091	45.24955	-0.10864	99.29756	99.12002	98.68468	0.61288	0.43534
	Freq⑦*Sev⑧		45.13905	45.24806	-0.10901	97.11465	96.27169	94.03225	3.08240	2.23944
	Freq⑧*Sev⑥		44.86455	44.91775	-0.05320	90.15844	92.88333	92.90790	-2.74947	-0.02457
	Freq⑧*Sev⑦		45.42065	45.39929	0.02136	100.03816	101.38425	101.08992	-1.05176	0.29432
	Freq⑧*Sev⑧		45.41866	45.39770	0.02096	97.81352	98.05799	98.19633	-0.38281	-0.13834
	Freq⑨*Sev⑥		44.84317	44.89627	-0.05310	89.87033	92.68684	92.94178	-3.07146	-0.25495
	Freq⑨*Sev⑦		45.39720	45.37579	0.02141	99.61975	101.11303	101.11321	-1.49346	-0.00018
	Freq⑨*Sev⑧		45.39522	45.37420	0.02101	97.42304	97.83014	98.22968	-0.80663	-0.39954

テストデータへの検証			Data1			Data2				
			A 交互作用あり (性別/年齢)	B 交互作用なし	A - B	A 交互作用あり (性別/年齢)	B 交互作用あり (地域/年齢)	C 交互作用なし	A - C	B - C
頻度	⑥	AGLM/Penal/α=0	0.39447	0.39388	0.00059	0.09373	0.09247	0.09202	0.00172	0.00046
	⑦	AGLM/Penal/α=0.5	0.39061	0.39036	0.00025	0.09229	0.09209	0.09200	0.00029	0.00009
	⑧	AGLM/Penal/α=1	0.39061	0.39037	0.00025	0.09242	0.09211	0.09202	0.00040	0.00009
損害規模	⑥	AGLM/Penal/α=0	2.01792	2.01411	0.00381	2.43968	1.68306	1.67780	0.76188	0.00526
	⑦	AGLM/Penal/α=0.5	1.99926	1.99880	0.00046	1.98822	1.98862	1.98862	-0.00039	0.00000
	⑧	AGLM/Penal/α=1	1.99932	1.99879	0.00053	1.84177	1.76680	1.76726	0.07451	-0.00046
総損害額	Freq⑥*Sev⑤		45.66566	45.49691	0.16875	161.76243	101.35207	103.12158	58.64085	-1.76951
	Freq⑥*Sev⑥		45.60423	45.47934	0.12489	117.25209	104.11897	100.03318	17.21892	4.08579
	Freq⑥*Sev⑦		45.60413	45.47878	0.12535	115.75463	100.31819	103.36808	12.38654	-3.04989
	Freq⑦*Sev⑤		45.25833	45.17121	0.08712	114.03695	99.50373	101.35604	12.68091	-1.85231
	Freq⑦*Sev⑥		45.22398	45.14990	0.07408	103.35072	103.64091	102.23761	1.11310	1.40330
	Freq⑦*Sev⑦		45.22381	45.14922	0.07459	101.31708	99.46792	98.98870	2.32839	0.47922
	Freq⑧*Sev⑤		45.25378	45.16704	0.08674	115.88791	99.27903	101.30924	14.57867	-2.03021
	Freq⑧*Sev⑥		45.21774	45.14381	0.07393	103.66439	103.32983	102.15077	1.51362	1.17906
	Freq⑧*Sev⑦		45.21757	45.14313	0.07445	101.72327	99.23842	98.91924	2.80403	0.31918

上の表は訓練データにあてはめた結果，下の表はテストデータの検証結果を表す．頻度はポアソン逸脱度，損害規模はガンマ逸脱度，総損害額はTweedie 逸脱度に基づいて算出している．

参考文献

[1] P. McCullagh and J. A. Nelder, "Generalized Linear Models," CRC Press, 1972.

[2] R. T. J. and T. H. J., "Generalized Additive Models," Chapman & Hall/CRC, 1990.

[3] E. W. Frees, R. A. Derrig and G. Meyers, Predictive Modeling Applications in Actuarial Science, Volume 1. Predictive Modeling Techniques, Cambridge University Press, 2014.

[4] B. Leo, F. J. H., O. R. A. and S. C. J., "Classification and regression trees," Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[5] G. Witten, D. Hastie, R. Tibshirani and R. James, "An Introduction to Statistical Learning: with Applications in R," Springer, 2013.

[6] J. R. Quinlan, "Bagging, Boosting, and C4.5," AAAI/IAAI, 2006.

[7] L. Breiman, “Random Forests,” Springer, 2001.

[8] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, 1970.

[9] R. Tibshirani, "Regression Shrinkage and Selection via the lasso," Journal of the Royal Statistical Society, 1996.

[10] H. Zou and T. Hastie, "Regularization and variable selection via the Elastic Net," Journal of the Royal Statistical Society, 2005.

[11] J. Gertheiss and G. Tutz, "Sparse modeling of categorial explanatory variables," The Annals of Applied Statistics, 2010.

[12] W. D. Stephen, A. R. Feinstein and C. K. Wells, "Coding ordinal independent variables in multiple regression analyses," no. American Journal of Epidemiology, 1987.

[13] E. W. Frees, R. A. Derrig and G. Meyers, Predictive Modeling Applications in Actuarial Science, Volume 2. Case Studies in Insurance, Cambridge University Press, 2016.

[14] E. Ohlsson and B. Johansson, Non-Life Insurance Pricing with Generalized Linear Models, Springer, 2015.

[15] "Case studies and examples: data and SAS programs and comments," [Online]. Available: http://staff.math.su.se/esbj/GLMbook/case.html. [Accessed 2019].

[16] "MACQUARIE University," 2012. [Online]. Available: http://www.businessandeconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets.

[17] "Package CASdatasets," 2016. [Online]. Available: http://cas.uqam.ca/.

[18] R. Srinivasan, "kaggle - Automobile Dataset," 2019. [Online]. Available: https://www.kaggle.com/toramky/automobile-dataset/home.

[19] M. Choi, "kaggle - Medical Cost Personal Datasets," 2019. [Online]. Available: https://www.kaggle.com/mirichoi0218/insurance.

[20] 2019. [Online]. Available: https://cran.r-project.org/web/packages/h2o/index.html.

[21] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight, "Sparsity and Smoothness via the Fused lasso," Journal of the Royal Statistical Society, 2005.

[22] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2016.

[23] D. J., K. R. and S. M., "Supervised and unsupervised discretization of continuous features," Machine Learning: Proceedings of the Twelfth International Conference, 1995.

[24] J. Friedman, T. Hastie and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," J Stat Softw, 2010.

[25] S. Garavaglia and A. Sharma, "A SMART GUIDE TO DUMMY VARIABLES: FOUR APPLICATIONS AND A MACRO," Proceedings of the Northeast SAS Users, 1998.

[26] R. Adamczak, A. Porollo and J. Meller, "Accurate prediction of solvent accessibility using neural networks–based regression," PROTEINS, 2004.

[27] D. D. Rucker, B. B. McShane and K. J. Preacher, "A researcher's guide to regression, discretization, and median splits of continuous variables," Journal of Consumer Psychology, 2015.

[28] J. Gertheiss, S. Hogger, C. Oberhauser and G. Tutz, "Selection of Ordinally Scaled Independent Variables," Applied Statistics 62, 2009.

AGLM: an extension of GLM for actuarial practice using data science techniques

Suguru Fujita*, Toyoto Tanaka†, Hirokazu Iwasawa‡

Abstract

Predictive modeling in machine learning and data science is paid much attention in recent years, and it is now one of the most important tasks for actuaries to apply it in practice. However, due to the characteristic features of insurance data and such priorities as the interpretability of models and regulatory requirements in their decision-making, most actuaries may find difficulties in using those advanced techniques. The aim of our research is not to apply existing methods per se into actuarial practice, but rather to construct methods to fulfill the actuary's original needs. We propose, from this standpoint, Accurate GLM (AGLM), a modeling method that is developed by combining data science techniques with the generalized linear model (GLM).

Keywords: Regression Analysis, Generalized Liner Model (GLM), Discretization, Dummy Variable, Regularization

* Guy Carpenter Japan Inc., Mail: suguru.fujita@guycarp.com
† Tokio Marine & Nichido Fire Insurance Co., Ltd., Mail: toyoto.tanaka@tmnf.jp
‡ Mail: iwahiro@bb.mbn.or.jp