1E4-2in2

高次元データの回帰分析結果検証のための可視化手法

A Scatterplot-based Visualization Tool for Regression Analysis

鈴木千絵*1　伊藤貴之*1　梅津圭介*2　本橋洋介*3
Chie Suzuki　Takayuki Itoh　Keisuke Umezu　Yousuke Motohashi

*1 お茶の水女子大学大学院 人間文化創成科学研究科　*2 *3 日本電気株式会社
Ochanomizu University　NEC

Regression analysis has been widely applied to various academic and industrial fields. Applications of regression analysis include medical problems such as health estimation, environmental problems such as disaster prediction and energy consumption estimation, and business/economic analytics. Here, accuracy and quality of regression analysis strongly depend on relevancy between input explanatory variables and actual objective functions. It often happens that several explanatory variables are well correlated with objective functions while others do not well correlated, and therefore accuracy of regression analysis may improve by removing unnecessary explanatory variables. This paper presents a scatterplot-based regression analysis tool. This tool visualizes the distribution of errors between actual and estimated values of objective functions, and provides user interfaces to explore the relationships between explanatory variables and the errors. This paper introduces examples of the visualization results using the presented tool with actual and estimated revenues at a store.

1. はじめに

回帰分析を用いた予測は自然科学や社会科学に関する幅広い分野で活用されている．特に複数の説明変数を入力情報とする重回帰分析重回帰分析において，予測に大きく寄与する説明変数と大きく寄与しない説明変数，また予測値と実測値の誤差につながる説明変数を特定することが，回帰分析の性能を向上するために重要である．

以下，小売店での商品の日々の販売を例題として議論する．日常的に販売される商品の売上は，その日の気温や曜日など，さまざまな要因に左右される．販売競争の激しい近年において，過剰発注や完売を防ぐために，適切な在庫数を保つことが必要である．取得データを解析し，販売数をある程度予測するため，その日の販売データを蓄積している企業は少なくない．しかし取得するデータの中には予測結果にほとんど影響しない要因も存在しており、それを予測に用いることでノイズを生み，予測値と実測値との誤差の要因になることがある．よって，入力情報がどのように予測値に寄与しているかを理解することは重要だが，情報の複雑化によってその理解が難しくなっている場合も多い．

そこで本報告では，回帰分析による予測値と実績値との誤差を可視化する一手法を提案する．本手法では回帰分析の対象となる標本群の可視化に3次元散布図を採用しており，説明変数群のうち2つを選んでX,Y軸に割り当て，予測値または実測値をZ軸に割り当てて各標本をプロットする．さらに予測値と実測値の誤差をプロットの色に割り当てることで，誤差が大きくなる標本が3次元空間中のどこに集中しているかを視認しやすくする．

ここで説明変数が非常に多い問題の場合，どの説明変数をX,Y軸に割り当てるかによって可視化の効果は大きく変わってしまう．そこで前処理として，各説明変数について予測値への寄与や誤差への要因を評価し，各説明変数の興味深さを定量的にユーザに提示することが有用であると考えられる．その定量評価手段の一例として本報告では，赤池情報量基準(Akaike's Information Criterion; AIC)をもとに各説明変数を評価した例を紹介する．

連絡先：鈴木千絵，お茶の水女子大学大学院　人間文化創成科学研究科，〒112-8610東京都文京区大塚2-1-1，chie@itolab.is.ocha.ac.jp

2. 関連研究

回帰分析や予測問題の可視化は有用であると考えられるが，それを目的として可視化システムを開発した研究事例は少ない．代表的な例としてThomasら[Muhlbacher　2013]は，複雑さが最小限に抑えられるモデルおよびそれに寄与する説明変数を推薦し，その選択が精度の高い回帰分析につながることを視覚的に表現する可視化システムを提案している．予測問題の評価基準としてAICを用いている事例もいくつか発表されている．山口ら[山口　2004]は飲食店の売上げデータの解析においてモデル選択の際にAICを用いており，現実の売上データにおいてもAICが有効であることを示している．

3. 提案手法

本章では我々が提案する可視化手法と，その説明変数評価手段の一例としてのAICの利用方法を示す．図1に我々が開発中の可視化ツールの画面キャプチャを示す．

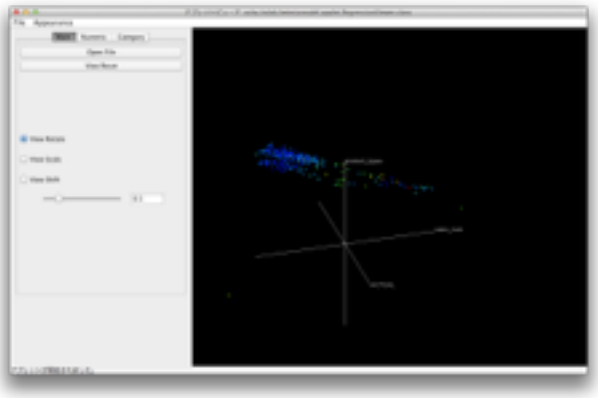

図1　本報告で提案する可視化ツール

3.1 データ構造

本研究では以下のデータ構造を想定する．

$$\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$$

$$x_i = \{v_{i1}, \ldots, v_{im}, c_{i1}, \ldots, c_{il}, p_i, a_i\}$$

ここでXは標本群，nは標本数，x_iはi番目の標本を表す．またmは説明変数の個数，v_{ij}はi番目の標本におけるj番目の説明変数値，lは後述するカテゴリ変数の個数，c_{ij}はi番目の標本におけるj番目のカテゴリ変数値，p_iはi番目の標本における予測値，a_iはi番目の標本における実測値である．

商品販売情報の回帰分析を例にすると，販売日の気温など，実数で表現される情報を説明変数として扱い，曜日など，実数値で表されない情報をカテゴリ変数とする．

3.2　3次元散布図による可視化

前節で示したデータ構造を可視化するために，我々の実装では3次元散布図を採用している．この可視化ツールでは，m個の説明変数群の中から2個を選んでx軸およびy軸に割り当て，実績値または予測値をz軸に割り当てることで，3次元散布図を実現する．また，各標本における予測値と実測値の誤差を色で表現する．我々の実装では，誤差の小さい要素を寒色系の色相で，誤差の大きい標本を暖色系の色相で描画する．またアフィン変換による拡大縮小・回転・平行移動の各操作を搭載しており，可視化結果をどの角度からも描画することができる．

可視化ツールの画面左側には4つのタブにGUI部品が搭載されている．1つ目のタブにはファイル操作や描画調節のためのGUI部品が搭載され，2つ目のタブにはx,y,z軸に割り当てる変数を選択するためのラジオボタンがm行にわたって搭載されている(図2).

3つ目および4つ目のタブにはカテゴリ変数選択のためのラジオボタンおよびチェックボックスを搭載している(図3)．3つ目のタブにはカテゴリ変数の種類を選択するラジオボタンがl列にわたって搭載される．ユーザがそのうちの1個を選択すると，4つ目のタブにはそのカテゴリ変数の選択肢となりえる変数値を選択するチェックボックスが搭載される．例えば3つ目のタブで「曜日」というカテゴリ変数を選択すると，4つ目のタブには「日曜」から「土曜」までの7個のチェックボックスが搭載される．

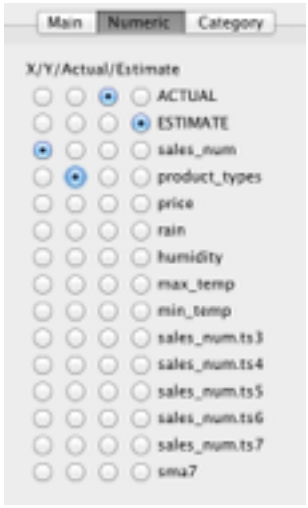

図2　説明変数選択タブ

4つ目のタブに搭載されたチェックボックス群のうち，チェックされているカテゴリ変数値をもつ標本は高彩度の色で描画され，チェックされていないカテゴリ変数値を持つ標本は灰色で描画される．この機能により，誤差分布とカテゴリ変数値の関係を表現可能にしている．

3.3　AICによる説明変数の評価

AICは，データとモデルの当てはまりの悪さを数値化したもので，次の公式で表される．

$$\mathrm{AIC} = -2 \log L + 2k$$

上式においてLは最大尤度，kは自由パラメータの数である．AIC値が最小のものを選択することで，多くの場合，良質な予測を実現できるモデルが選択できることが知られている．

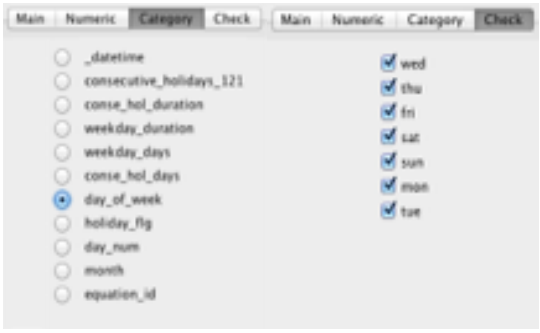

図3　カテゴリ変数選択タブ

4.実行結果

本章では販売情報の回帰分析結果について本手法を適用した．入力データには，実績値と予測値に加え，12個の説明変数，8個のカテゴリ変数が含まれていた．

入力データに対してAIC値を求め，AIC値が小さくなるモデルを選択した結果を表1に示した．AIC値が小さくなる3つの説明変数（説明変数A～C）から任意の2軸を選択し，それらをx軸およびy軸に割り当てて可視化した．図4~6にその可視化結果を示す．いずれの可視化結果においても実績値が小さい方に寒色点が多くあり，暖色点は寒色点の集合から離れた位置にプロットされている．

表1　各説明変数を用いたモデルのAIC値

説明変数	AIC値
説明変数Aのみ	113.76
説明変数A～B（2個）	115.76
説明変数A～C（3個）	117.84
説明変数A～D（4個）	119.88
説明変数A～E（5個）	121.92
説明変数A～F（6個）	123.93
説明変数A～G（7個）	125.96
説明変数A～H（8個）	127.96
説明変数A～I（9個）	129.95
説明変数A～J（10個）	131.95
説明変数A～K（11個）	133.95

説明変数Aをx軸に，説明変数Bをy軸に割り当てた可視化結果を図4に示す．この結果から，説明変数Aと説明変数Bの双方が大きいときに誤差が大きくなり，また実績値は大きいときに誤差が大きい傾向があることがわかる．説明変数Aをx軸に，説明変数Cをy軸に割り当てた可視化結果（図5）からも同様に，説明変数の双方が大きいときに誤差が大きくなる傾向が観察された．

説明変数Bをx軸に，説明変数Cをy軸に割り当てた可視化結果（図6）においても暖色系の点と寒色系の点はある程度分離して見えるが，図4,5に比べるとxy平面全体に渡って暖色系および寒色系の点が広く分布しているのがわかる．よって，説明変数B,Cを2軸としたときの誤差の説明性は下がることが示唆される．以上より，説明変数Aが回帰分析による予測において特に重要な説明変数であると考えられる．

5.まとめと今後の課題

本報告では，回帰分析の結果検証のための3次元散布図ツールを提案した．本手法では説明変数のうち2つを選んでx軸とy軸に割り当て，予測値または実測値をz軸に割り当てることで3次元散布図を実現する．そして予測値と実測値の誤差を色で表現することで，誤差の大きい標本の分布を視覚的に観察できる．ここでx軸とy軸に割り当てる説明変数を選択するための一手段として，我々はAICを用いて説明変数の評価を実施した．その結果として有用な説明変数を選択して3次元散布図に用いることで，誤差の小さい要素と誤差の大きい要素を視覚的に分離できるような3次元散布図を実現できた．

今後の課題として，AICを用いた説明変数評価手法の改善があげられる．AICは非線形性を有する標本群に対して，非線形性ゆえの数値特性を誤差として扱ってしまい，複雑なモデルを選択してしまう場合がある．そこで深田らの提唱するモデル次数決定手法[3]を適用した可視化結果や，ベイズ情報量基準(Schwarz's Bayesian Information Criterion; BIC) [4] といった他のモデル選択基準を適用した可視化結果を比較し，より精度の高い結果が得られる手法を適用することも検討する．また説明変数の選択基準として，AIC以外の評価手法を適用することも考えられる．一例として，誤差の大きい要素が散布図上でより固まる2軸を選ぶ手法を開発したい．これを実現するために我々は，xy平面を格子状に分割し，各区画における誤差の傾向を統計的に解釈する手法を実装する予定である．

説明変数の選択だけでなく，興味深い誤差分布を発見するためのカテゴリ変数評価手法の開発も課題としてあげられる．カテゴリ変数値を効果的に選択できるようになれば，誤差の要因や各変数間の関係をさらに詳細に理解できると期待される．

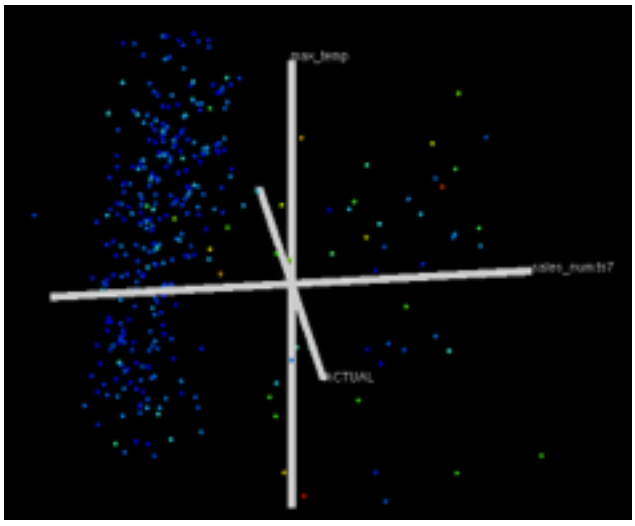

図4　説明変数A,Bを用いた可視化結果

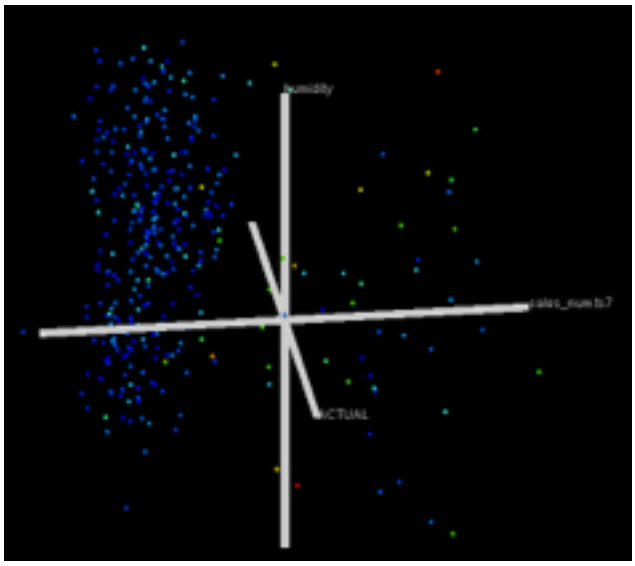

図5　説明変数A,Cを用いた可視化結果

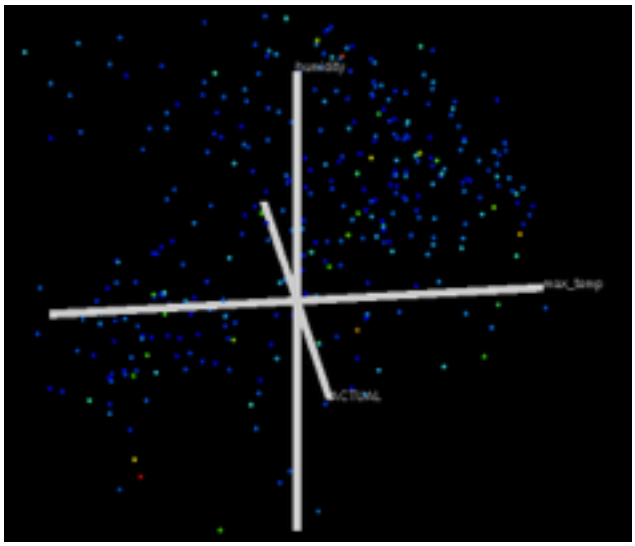

図6　説明変数B,Cを用いた可視化結果

参考文献

[Muhlbacher 2013]　Thomas Muhlbacher and Harald Piringer: A Partition-Based Framework for Building and Validating Regression Models，EEE Trans Vis Comput Graph 125, pp. 1962-1971，2013.

[山口　2004]　山口類，土屋映子，樋口知之: 論文タイトル，状態空間モデルを用いた飲食店売上の要因分解，2004.

[深田　2006]　深田健太，鷲尾隆，矢田勝俊，元田浩: 広告・販促効果に関する外部入力付自己回帰モデル解析，The 20th Annual Conference of the Japanese Society for Artificial Intelligence，2006.

[山口　2008]　山口健太郎: 統計学におけるモデル：情報量基準の観点から，科学哲学科学史研究，pp. 43-59，2008.