Department of Social Systems and Management

Discussion Paper Series

No. 1213

最小絶対値回帰分析を利用した 落札金額マッピングシステムの提案

by

高野 祐一, 水野 圭, 山菅 和人, 佐藤 齊行, 小川 直哉

July 2008

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

最小絶対値回帰分析を利用した落札金額マッピングシステムの提案

高野 祐一 * 　水野 圭 † 　山菅 和人 ‡ 　佐藤 齊行 * 　小川 直哉 *
* 筑波大学 システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
† 日本電気株式会社 第一キャリアソリューション事業本部　〒 108-8001 東京都港区芝 5-7-1
‡ 東日本電信電話株式会社 東京支店総務部　〒 108-8019 東京都港区港南 1-9-1

平成 20 年 7 月 31 日

概要　回帰モデル構築の際には, なるべく説明変数を減らしてシンプルなモデルを作成したいと考えられるが, 大量のサンプルや説明変数を用いて分析を行う場合には計算時間や説明変数の組合せ数が増大し, 回帰式に使用する説明変数の選択は困難な問題となる. そこで本研究では, 説明変数の選択を組み込んだ最小絶対値回帰分析を 0-1 混合整数線形計画問題として効率的に求解し, 中古車オークションにおける車両落札金額予測モデルを構築する. 予測モデルには, 多重共線性の排除, CVaR を用いたリスク回避, プロスペクト理論の考慮といった 3 種類の工夫を組み込む. さらにサンプルのセグメンテーションを行い, セグメントごとに回帰式を構築することで回帰式の精度を向上させる. また, 落札金額予測サービスは入力項目が多いとユーザーが使いにくいと考えられるが, 本研究で提案する落札金額マッピングシステムでは最適変数選択を利用することでユーザーの入力の負担を軽減させている.

キーワード： 最小絶対値回帰分析, 最適変数選択, Conditional Value-at-Risk, 中古車オークションデータ

1　はじめに

本研究では,「平成 19 年度データ解析コンペティション」において提供された中古車オークションデータを用いて落札金額の回帰分析を行い, 落札金額予測サービスを提案する. 提供されたデータは, 2005 年 6 月から 2007 年 6 月までの約 24ヶ月間の車両出品データ（サンプル数 125,880）であり, オークションは 17 会場で 860 回開催され, 入札は一度きりで他者の入札額は知らされず最高額の入札者が落札する第一価格秘密入札方式がとられている.

まず, 本中古車オークションの収入体系を確認する. このオークションの主な収入源は, 出品者からの出品料, 成約料, 再出品料（1 台あたり定額）と, バイヤーからの年会費（1 社あたり定額）, 落札料（1 台あたり定額）であり, 落札金額や札数には依存しない収入体系となっている. したがって, 出品者, 出品台数の増加が収入の増加に直結することが分かる. また, 出品者は落札金額に満足できなければ成約せずに流札を選択することも可能だが, このオークションでは成約率が 95.8%と非常に高い. 上記 2 点より, 出品者, 出品台数の増加が高い確率で, 成約料, 落札料の増加につながることも分かる.

こういった背景から, 出品車両の落札金額を前もって見積もることで出品者の不安を取り除き, より多くの出品を促進することは，オークション主催者にとっても関心があると考えられる.

また, 現在では, インターネットによる通信販売や POS システムの充実化によって, 大規模データを容易に手に入れることができる. 一方で, その規模故に分析が難航することが少なくなく, その理由としてデータ処理や計算に時間を要したり, モデルが複雑になりその解釈が難しくなってしまうことなどが考えられる. 特に回帰モデル構築の際には, なるべく説明変数を減らしてシンプルなモデルを作成したいと考えると, 大規模データの下では使用する説明変数の選択は困難な問題となる.

そこで本研究では Konno-Yamamoto [6] を参考に, 説明変数の選択を組み込んだ最小絶対値回帰分析を 0-1 混合整数線形計画問題として定式化し, 使用する説明変数の選択を回帰分析と同時に行う. また, 最小絶対値法の利点を活用し, 多重共線性の排除, Conditional Value-at-Risk を利用して予測を大きく外すリスクを回避, プロスペクト理論を考慮し予測額より実際の落札金額が下がることを避ける, といった工夫を行う. さらに, サンプルを距離と排気量でセグメンテーションし, セグメントごとに回帰式を構築することで回帰式の精度を向上させる. そして最適変数選択を利用して, ユーザーの入力の負担を減らした落札金額予測サービスである落札金額マッピングシステムを提案し, その実装案を示す.

本論文の構成は次の通りである. 第 2 節でデータの概要を説明し, 第 3 節では最適変数選択・最小絶対値回帰分析について説明する. 第 4 節では回帰式を構築し, 3 種類の工夫を提案する. 第 5 節ではサンプルのセグメンテーションを行って, セグメントごとに回帰式を構築し, 回帰式の精度評

価を行う．第 6 節では落札金額マッピングシステムを提案し, 第 7 節では結論と今後の課題について述べる．

2　データ概要

今回分析に用いたデータについて説明する．回帰分析に使用した説明変数は次の通りである:

質的変数（25 種）: { エアコン, パワーステアリング, パワーウィンドウ, サンルーフ, 革シート, カーナビ, 左ハンドル, W タイヤ, 冷凍冷蔵装置（機械式）, 冷凍冷蔵装置（畜冷式）, 抹消区分, 保証書, 整備手帳, 記録, 修復歴, ヤブレ, 穴, 亀裂 } の有無, 会場コード, 形状記号, 燃料, 駆動方式, ドア数, 距離区分, レンタカー歴.

量的変数（7 種）: 排気量, 看板面数, 距離, 全長, 全幅, クレーン数, 新車価格.

次は提供されたデータを利用して我々が独自に作成した説明変数である:

作成した変数（9 種）:

・累計月数（出品日は 2005 年 6 月から何ヶ月目か, 量的）
・会場ごと累計月数（例えば「会場 1 累計月数」は会場 1 で出品された車両には累計月数を入力し, それ以外には 0 を入力した変数, 量的）
・出品月（出品日が何月か, 質的）
・登録月数（初年度登録と出品日から計算, 量的）
・車検残存（出品日における車検残存の有無, 質的）
・車検残存月数（出品日と車検期限年月から計算し, 残っていない場合は 0, 量的）
・メーカ名（サンプル数上位 9 メーカ, その他国内, 海外の 11 値, 質的）
・ミッション記号（段数とミッションでグループ化, 質的）
・損傷ポイント（車両パネル損傷データから計算, 量的）

質的変数については適宜ダミー変数化を行い, 説明変数の数は 133 個となる．

本研究では, 車検期限年月, 損傷ポイント, 新車価格のいずれかに欠損のあるサンプルを除去した 82,510 サンプルを使用した．ただし, パラメータ設定のための実験（図 4, 図 6）と第 6 節の変数選択では計算の都合上, 各会場から 500 サンプルずつ選び, 計 8,010 サンプルで計算を行った．なお, サンプル数が 500 に達しない会場があるためにサンプル数の合計が 500×17 とは一致していない．

被説明変数は, 新車価格の何%の金額で落札されたかを示す落札金額率（単位: %）を用いた:

(落札金額率):=100×(落札金額)/(新車価格).

予備的な実験において, 落札金額率の予測値に (新車価格)/100 をかけて落札金額に変換すると, 落札金額を直接予測するよりも精度が良かったため, 本研究では落札金額率を被説明変数とした．なお, 入札の無かった車両（サンプル数 960）は落札金額を 0 円としている．また, 回帰分析の結果, 予測値が負になってしまう車両には価値が認められず入札がされないと考え, 後処理として予測値を 0 に変換している．

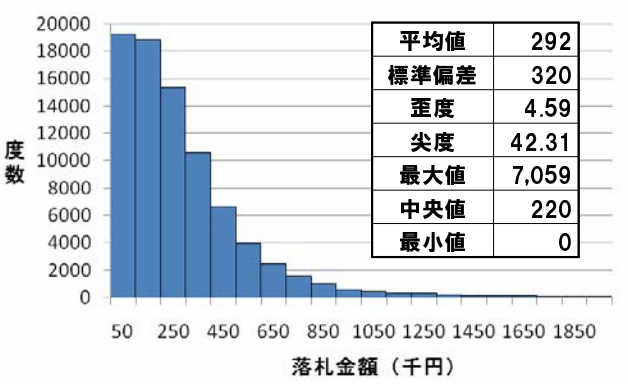

図 1: 落札金額の度数分布と基本統計量

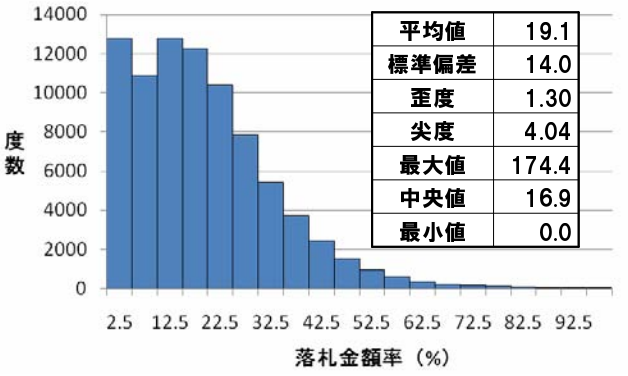

図 2: 落札金額率の度数分布と基本統計量

図 1 と図 2 に落札金額と落札金額率の度数分布と基本統計量を示す（度数分布はそれぞれ, 200 万円, 100%以上の部分については省略した）．落札金額の統計量（図 1）を見ると全サンプルの半数は 22 万円以下で落札もしくは入札無しとなっている一方で, 700 万円もの高額で落札されている車両が混在していることが分かる．同様の傾向は落札金額率（図 2）にも見られ, 新車価格の 1.7 倍で落札された車両が存在する．

3　最適変数選択・最小絶対値回帰分析

3.1　定式化

回帰分析は, 被説明変数を回帰式と呼ばれる説明変数の式で表すことによって, その間の関係をとらえる統計手法である．説明変数の線形式を用いて被説明変数との差の 2 乗和

が最小となるように回帰式を構築する最小 2 乗法が代表的な推計手法である.

しかし, 大量のサンプルや説明変数を用いて分析を行う場合には, 計算時間や説明変数の組合せ数が増大し, 回帰式に使用する説明変数の選択は困難な問題となる. その解決策として, 説明変数の選択を組み込んだ回帰分析を 0-1 混合整数計画問題として定式化し, 最適な説明変数集合を回帰式の構築と同時に求める方法がある. また, 最小 2 乗法を用いて定式化すると目的関数は 2 次関数となるが, 予測値と被説明変数の値との差の絶対値（絶対偏差）を最小化する最小絶対値法を用いると, 線形計画問題に帰着でき効率的に求解できる. 本研究では, 被説明変数の T 個の観測値 y_t, $t \in \mathcal{T} = \{1, ..., T\}$ と, J 個の説明変数の T 個の観測値 x_t^j, $j \in \mathcal{J} = \{1, ..., J\}$, $t \in \mathcal{T} = \{1, ..., T\}$ が与えられたとし, J 個の説明変数から s 個の説明変数を選択する問題を [6] を参考に次のように定式化し, 最適化ソルバー Xpress-MP (ver 2007B) を利用して計算を行った.

$$
\left|
\begin{array}{ll}
\underset{\boldsymbol{\alpha},\boldsymbol{u},\boldsymbol{v},\boldsymbol{z}}{\text{minimize}} & \displaystyle\sum_{t\in\mathcal{T}} \frac{1}{T}(u_t + v_t) \\
\text{subject to} & u_t - v_t = y_t - \Big(\alpha_0 + \displaystyle\sum_{j\in\mathcal{J}} \alpha_j x_t^j\Big), \\
 & \qquad\qquad\qquad\qquad\qquad\qquad t \in \mathcal{T} \qquad (1) \\
 & u_t \geq 0, \quad v_t \geq 0, \quad t \in \mathcal{T} \\
 & \displaystyle\sum_{j\in\mathcal{J}} z_j = s \\
 & 0 \leq \alpha_j \leq M_j z_j, \quad z_j \in \{0,1\}, \quad j \in \mathcal{J},
\end{array}
\right.
$$

ただし, M_j, $j \in \mathcal{J}$ は十分大きな正の定数とする. ここで, 前処理として各説明変数の被説明変数との相関係数を計算し, 相関係数が負の場合には説明変数の符号を反転させておく. 被説明変数との相関係数が非負だからといって偏回帰係数が必ず非負になるとは言えないが, 整数計画問題は制約式を追加することで問題が解きやすくなる場合が多く, 計算時間の短縮のために定式化 (1) では偏回帰係数を表す決定変数 α_j, $j \in \mathcal{J}$ を非負変数としている.

定式化の仕組みについて, 簡単に説明する. 被説明変数の値から予測値を引いた残差は, 非負決定変数の差 $u_t - v_t$ で表されている. 目的関数の最小性により, 残差が正（負）の場合は $v_t = 0$（$u_t = 0$）となり, u_t（v_t）に残差の絶対値が現れ, 目的関数値は平均絶対偏差となる. また, 0-1 決定変数 z_j が 0 になると, 制約式により偏回帰係数 α_j が 0 となることで, 対応する説明変数は回帰式から削除される. なお, 第 6 節の計算以外では回帰式の精度を優先し, 説明変数の選択個数に関する制約式 $\sum_{j\in\mathcal{J}} z_j = s$ は消去しておく.

3.2 最小絶対値法の利点

末吉 [11] では最小絶対値法の利点として, 次の 2 点を挙げている:

1. 残差が Laplace 分布に従う場合は最小絶対値法による偏回帰係数が最尤推定値になる [1] など, 残差が正規分布に従わない（特に外れ値を含む）場合には最小 2 乗法より信頼度の高い偏回帰係数推定値を与える.
2. 正規方程式を解いて計算を行う最小 2 乗法とは対照的に, 線形計画問題として定式化することで, 様々な目的や条件を加味できる.

そこで, 残差の正規性を調べるために正規 Q-Q プロットを行った（図 3）. 残差が正規分布に従う場合, プロットは直線状になることが期待されるが, 図 3 を見るとプロットの右端で対照直線からの系統的なずれが見られる. よって, 今回のデータでは残差が正規分布に従うとはみなせず, 最小絶対値法が有効であると言える. また, この系統的なずれは第 2 節で観察したような高額で落札されている車両の混在と関連する.

第 2 の利点に関して補足すると, 最小 2 乗法も 2 次計画問題として定式化することは可能である. しかし, 大規模な問題や整数変数を導入した問題を解くことを考えると線形計画問題として定式化可能な最小絶対値法に優位性がある.

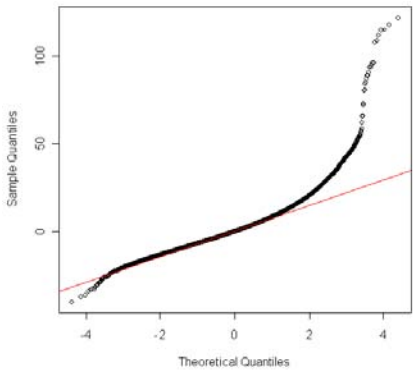

図 3: 残差の正規 Q-Q プロット

4 回帰式の構築と工夫

前節で, 最小絶対値法の利点として様々な目的や条件を加味できることを挙げたが, 本節ではその利点を活用し 3 種類の工夫を提案する.

4.1 多重共線性の排除

多重共線性とは説明変数間の強い相関関係のことであり, 多重共線性がある場合, 明らかに被説明変数に対して正の影響を与える説明変数の偏回帰係数の符号が負になってしまうなどの悪影響を回帰結果にもたらすことがある. そこで, 本研究では説明変数 $\boldsymbol{x}^i, \boldsymbol{x}^j$ $(i > j)$ 間の相関係数の絶

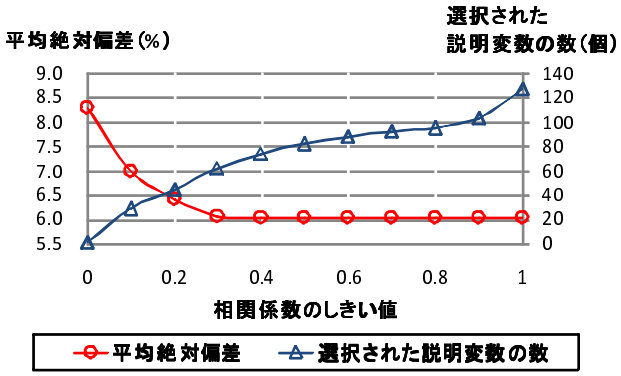

図 4: 相関係数のしきい値 c が回帰の結果に与える影響

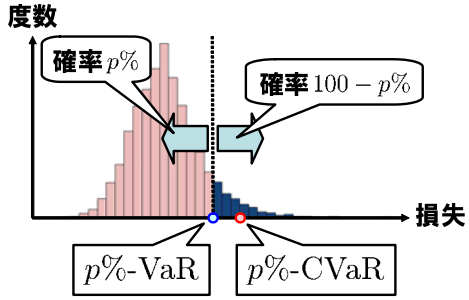

図 5: VaR と CVaR のイメージ

対値がしきい値パラメータ $c \in [0,1]$ より大きい場合に制約式 $z_i + z_j \leq 1$ を定式化 (1) に追加し, 相関の強い 2 変数の少なくともどちらか一方を削除することで多重共線性を排除した. この方法では最小絶対値基準の下で, 分析者の主観や恣意性が入ることなく多重共線性を排除することが可能である. しきい値パラメータ c の値は分析者が任意に設定できるが, 0.3 以下に設定して回帰式の精度を悪化させてしまうことは好ましくないと考え, 本研究では余裕を持って $c = 0.5$ と設定した（図 4, ただしサンプル数 8,010 で計算）.

4.2 CVaR を用いたリスク回避

多重共線性を排除する制約式を追加して落札金額率の最小絶対値回帰分析を行った結果, 921 サンプル（全サンプル中の約 1.1%）で絶対偏差が 25%以上となった. 例えば新車価格 200 万円の車両が 70 万円で落札される場合に, 120 万円, または 20 万円と予測すると絶対偏差が 25%となる. この例からも分かるように, 25%以上の絶対偏差というのは非常に大きく, 予測サービス依頼者からの信用を失うような事態も起こりかねない. こうしたリスクを避けるために, 本研究では Conditional Value-at-Risk [8]（以下, CVaR）を利用する.

CVaR の定義と回帰分析への応用 CVaR は, 損失を表す確率変数 $\tilde{X}$ を入力とするリスク尺度であり, Value-at-Risk（以下, VaR）以上という条件付きの損失の期待値として説明できる. なお本研究では, 残差の正負に関係なく予測値が観測値から大きく外れることが予測サービスの信頼性を損なうリスクであると考え, 回帰式の絶対偏差を用いて損失を定義している. まず VaR を定義しよう. 信頼水準を表すパラメータ $p \in [0, 100)$ を設定し（本研究では, $p = 99$ とした）, p%-VaR は次の式で定義される:

$$\min_{\eta \in \mathbb{R}} \{\eta \mid \Psi_{\tilde{X}}(\eta) \geq p/100\}, \tag{2}$$

ただし, $\Psi_{\tilde{X}}$ は損失 $\tilde{X}$ の分布関数, すなわち $\Psi_{\tilde{X}}(\eta) := \mathrm{Prob}(\tilde{X} \leq \eta)$ である.

確率変数 $\tilde{X}$ が連続分布に従う場合には, p%-VaR は $\tilde{X}$ の p% 分位点と一致する. この p%-VaR を超えるような損失の期待値が p%-CVaR であり, 確率変数が有限個の実現値を持つ離散分布に従う場合, p%-CVaR は p%-VaR を表す決定変数 $\eta \in \mathbb{R}$ と, $(\max\{X_t - \eta, 0\})$ を表す決定変数 $\boldsymbol{w} \in \mathbb{R}^T$ を用いた次の定式化で計算できる [8] :

$$\left| \begin{array}{ll} \underset{(\eta, \boldsymbol{w}) \in \mathbb{R} \times \mathbb{R}^T}{\text{minimize}} & \eta + \dfrac{1}{1 - p/100} \displaystyle\sum_{t \in \mathcal{T}} q_t w_t \\ \text{subject to} & w_t \geq X_t - \eta, \quad w_t \geq 0, \quad t \in \mathcal{T}, \end{array} \right. \tag{3}$$

ただし, X_t, $t \in \mathcal{T} = \{1, ..., T\}$ は確率変数 $\tilde{X}$ の実現値で, 各生起確率は q_t, $t \in \mathcal{T}$ であるとする. 本研究では, 定式化 (1) において, $X_t := u_t + v_t$, $q_t := 1/T$, $t \in \mathcal{T}$ とすることで, 損失を絶対偏差で定義した CVaR を計算している.

CVaR は大損失の期待値であり, 最小化して大損失のリスクを回避したいと考えられるが, 損失の各実現値 X_t が決定変数に対して凸関数であれば CVaR 最小化問題は凸計画問題に帰着されるという計算上好ましい性質がある [8]. また, CVaR は Coherent なリスク尺度, すなわちリスク尺度が満たすべき望ましい性質を備えていて [2], さらにリスク回避的な効用関数による期待効用最大化原理と高い整合性がある [7]. ゆえに, CVaR は計算面, 理論面の両面から好ましいリスク尺度であると言える. 回帰分析に CVaR を導入した研究として [9] などがあるが, そこでは変数選択は考慮されていない.

平均-CVaR モデル 目的関数を 99%-CVaR として計算を行うと, 問題となっていた絶対偏差が 25%以上のサンプルは減るが, 一方で平均絶対偏差の悪化が大きい. そこで本研究では, 絶対偏差の平均値と 99%-CVaR のトレード・オフ・パラメータ $\lambda \in [0, 1]$ を導入した次の式 (4) を目的関数とし, 最小化することで両方を同時に考慮する, 絶対偏差の平均-CVaR モデルを用いる,

$$(1 - \lambda) \times (\text{平均絶対偏差}) + \lambda \times (99\%\text{-CVaR}). \tag{4}$$

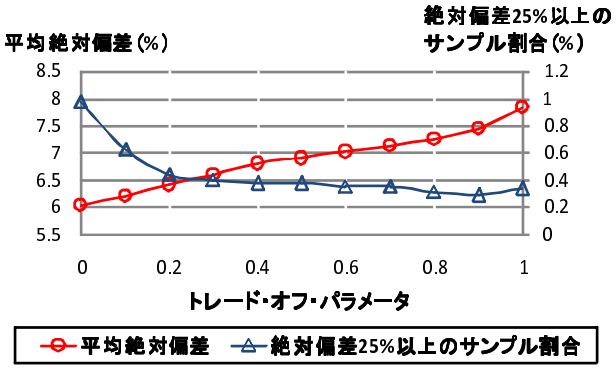

図 6: トレード・オフ・パラメータ λ が回帰の結果に与える影響

本研究では, トレード・オフ・パラメータ $\lambda = 0.2$ と設定した. 図 6（ただしサンプル数 8,010 で計算）から, 平均絶対偏差の悪化は小さく, 絶対偏差が 25%以上となるサンプルの割合は大きく減少していることが分かる.

4.3 プロスペクト理論の考慮

Kahneman-Tversky [4] の提唱したプロスペクト理論では, 人々の評価は図 7 のような価値関数によって説明され, その特徴として次の 2 点が挙げられる [12]:

1. 価値評価はある期待（参照点）からの乖離によって行われる.
2. 参照点からの乖離が同程度なら, 利得に対する満足感より損失に対する拒否感の方が大きい.

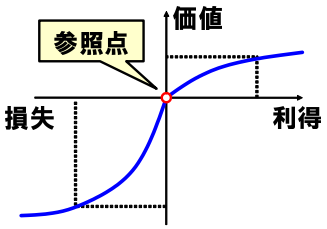

図 7: プロスペクト理論における価値関数

落札金額の予測を行い, それを出品者に示すことによって, 出品者は落札金額の予測値, すなわち参照点を持つようになる. プロスペクト理論を考慮すると, 前もって予測した落札金額より実際の落札金額が下がった場合の拒否感は, 実際の落札金額が上がった場合の満足感よりかなり大きいと考えられる. そこで予測値より観測値が小さい場合の偏差 v_t と, 予測値より観測値が大きい場合の偏差 u_t の重要度に

表 1: セグメントごとの回帰の結果（上段: サンプル数, 下段: 平均絶対偏差（単位: %））

距離 (km)	排気量 (cc) -660	661-1,500	1,501-2,000	2,001-2,500	2,501-3,000	3,001-
-5 万	6,646 5.43	3,882 5.60	2,228 7.11	621 6.24	453 7.94	571 7.09
5 万-10 万	8,711 4.56	10,822 4.28	5,626 5.12	1,721 5.75	1,156 5.80	1,189 5.69
10 万-15 万	3,197 3.70	9,921 4.63	5,813 4.85	2,126 5.57	1,268 5.74	1,105 4.86
15 万-	776 3.68	5,030 4.55	4,040 4.70	2,496 5.58	1,392 6.04	1,720 4.56

差を付けるために, 絶対偏差の計算式（定式化 (1) の目的関数）を次のように変形した:

$$\sum_{t \in \mathcal{T}} \frac{1}{T} \Big(\theta u_t + (2 - \theta) v_t \Big), \tag{5}$$

ただし, $\theta \in (0, 2)$ は偏差の相対的重要度を表すパラメータである. $\theta = 1$ に設定すると, 式 (5) は元の絶対偏差の計算式になる. 本研究では, $\theta = 2/3$ と設定し, 予測値より観測値が小さい場合の偏差には相対的に 2 倍の重みをかけることにした. このように参照点に対して非対称な偏差計算式を採用することは, 図 7 の価値関数の形からも自然な方法だと考えられ, 出品者に示す予測落札金額が下がることで流札を減らす効果も期待できる. また, 式 (5) の偏差計算による最小絶対値回帰は, Koenker-Bassett [5] によって提案された分位点回帰と見なすこともでき, 落札金額の $100 \times \theta/2$ % 分位点回帰に対応する.

5 セグメンテーションと回帰式の精度評価

5.1 セグメンテーション

第 4 節で提案した 3 種類の工夫を組み込んで回帰分析を行うと, 平均絶対偏差が 6.30% となる. 本節では距離と排気量を用いてサンプルのセグメンテーションを行い, セグメントごとに回帰式を作ることで回帰式の精度を向上させる. 距離はおおよそ 4 分位点となるような値で分類し, 排気量は税金の基準を参考に 6 種類に分類した. 表 1 は各セルが $4 \times 6 = 24$ 種類のセグメントに対応し, 各セルの上段に対応するセグメントのサンプル数, 下段には対応するセグメントで回帰式を構築した際の平均絶対偏差を記載している. ここで排気量に着目すると, 排気量 660cc 以下の車両すなわち軽自動車の回帰式は比較的精度が良く, 対照的に排気量が

表 2: セグメントごとの回帰の集計結果

被説明変数	平均絶対偏差	決定係数
落札金額率（単位: %）	4.95	0.79
落札金額（単位: 千円）	75.9	0.86

表 3: 5 分割交差検証法の結果

組合せ No.	1	2	3	4	5
平均絶対偏差 (単位: %)	5.31	5.42	5.29	5.30	5.32

2,001cc から 3,000cc の車両すなわち大型乗用車と呼ばれるような車両に関しては, 比較的精度が悪いことが分かる. また, 距離に着目すると, 走行距離が短い車両ほど回帰式の精度が悪く, 走行距離が長くなるにつれて回帰式の精度が良くなる傾向がある. これらは, 「大型乗用車の値段は高い物から安い物まで多くの種類があるために予測は難しいだろう」, 「長い距離を走っている車両の落札金額は安いに違いなく予測は簡単だろう」といった直感に近い結果と言える.

最後に, 表 2 でセグメントごとの回帰を集計した結果と, 落札金額率の予測値に (新車価格)/100 をかけて落札金額の予測とみなした結果を示す. 落札金額率は平均絶対偏差が 6.30% から 4.95% になり, 回帰式の精度が大きく改善されていることが分かる. また, 落札金額の予測とみなすと決定係数が 0.86 となり, 説明力の高いモデルが作れていることが分かる. ここまでに, 各種の工夫を回帰式に組み込み回帰式の性質を改良してきたが, セグメンテーションを行うことにより, 非常に精度の良い回帰式を得ることができた.

5.2 5 分割交差検証法

ここまでは全サンプルを利用して回帰式を構築し, 全サンプルを利用して回帰式の精度評価を行ってきたが, ここでは予測モデルとしての性能を評価するために 5 分割交差検証法を行う. まず, サンプルを 5 グループに分割し, その中の 4 グループを利用して回帰式を構築し, 回帰式の構築に利用しなかったサンプルで回帰式の精度を評価するということを 5 通りの組合せで行った（表 3）. 交差検証を行っているため, 全サンプルを利用した場合よりは結果が良くないが悪化の度合いは小さく, モデル構築用のサンプルに当てはまり過ぎてモデル評価用のサンプルでは精度が悪くなるといった過学習を生じていないことが確認できる. この結果から予測モデルとしての有効性についても実証できた.

6 落札金額マッピングシステムの提案

6.1 最適変数選択による入力項目の絞り込み

落札金額予測サービスを提供する上での問題点として, 例えどれだけ精度の良い予測でも, 入力項目が多過ぎるとユーザーが入力を面倒に感じて予測サービスの利用をためらう, といったことが考えられる. ここまでの分析では説明変数の選択個数に関する制約式 $\sum_{j\in\mathcal{J}} z_j = s$ を消去していたが, ここではこの制約式を利用して, 10 項目程度に入力項目を絞ることにする. 予測サービスでは説明変数の値が入力となるが, 多値の質的変数については選択式にすることで入力も簡単であり, 複数個のダミー変数として回帰式に組み込まれているため, 変数 1 個を採用することで回帰式の上では複数個のダミー変数を使用することに対応する. 新車価格は落札金額率を落札金額へ変換する際に必要不可欠であり. また, 直近のオークション開催日を想定して予測落札金額を出力することとして, 出品月はユーザーの入力項目とはせずに回帰式に使用する. 以上より, 質的変数であるメーカ名, ミッション記号, ドア数, 形状記号と新車価格, 出品月は, 多重共線性に関連して削除される以外は無条件で採用し, その他のユーザーが入力せざるをえない説明変数を最適変数選択（ただしサンプル数 8,010 で計算）で 5 個に絞り込んだ.

無条件で採用する変数（6 種）: メーカ名（質的, ダミー変数 11 個）, ミッション記号（質的, ダミー変数 12 個）, ドア数（質的, ダミー変数 7 個）, 形状記号（質的, ダミー変数 13 個）, 新車価格, 出品月.

最適変数選択により選択された変数（5 種）: 登録月数, 車検残存月数, 距離, 損傷ポイント, 冷凍冷蔵装置（機械式）.

上記の説明変数を使用して会場ごとに落札金額率の回帰式を構築する. なお, サンプル数が少なくなってしまうため, ここでは距離 × 排気量のセグメンテーションは行わない.

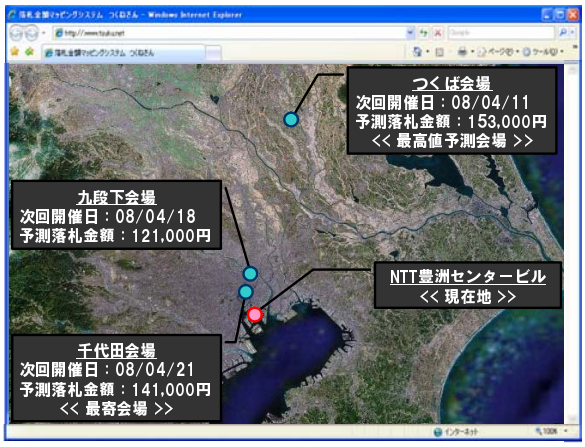

図 8: 予測落札金額出力画面イメージ

6.2 実装案

本研究で提案する落札金額マッピングシステムは会場ごとの回帰式を基に, 各会場に出品した場合の予測落札金額を地図上で表示するものである. まず, システムにログインし, 現在位置（出品車両が保管してある場所など）を入力する. 次に, 出品する車両のデータを入力する. ここで, 入力項目が出品月を除く厳選された 10 項目だけで済む点が本システムの利点であり, ユーザーにとって非常に使いやすいものとなっている. 車両のデータ入力を終えると, 地図上に近場のオークション会場と次回オークション開催日, 予測落札金額が表示される（図 8, ただし背景の地図は [13] を利用）. よって, ユーザーはどのオークション会場で, いつオークションが開催され, 落札金額はどれくらいになりそうか, といった情報を一目で理解できる.

7 おわりに

本研究では, 最小絶対値回帰分析を基に落札金額率の回帰式を構築した. 最小絶対値法には偏回帰係数の信頼度といった統計的性質の良さだけでなく, 線形計画問題として定式化することで, 様々な目的や条件を加味できる利点がある. 本研究ではその利点を活用し, 多重共線性の排除, CVaR を用いたリスク回避, プロスペクト理論の考慮といった 3 種類の工夫を組み込み回帰式の性質を改善した. さらにサンプルのセグメンテーションを行い, セグメントごとに回帰式を構築することで回帰式の精度を改善した. 最後に提案した落札金額マッピングシステムは, 最適変数選択を利用することで入力項目を 10 項目と厳選し, ユーザーの入力の負担を軽減している点が最大の利点と言える.

本研究で計算したような大規模な混合整数計画問題が最適化ソルバーを利用して解けるようになったことは最適化手法の急速な発展に依るところが大きい. また, 今回は中古車オークションデータを利用して分析を行ったが, 本研究で用いた手法は汎用性の高い方法ばかりであり, その用途は今回のデータに限ったものではない.

今回のように何らかの割合を予測する場合にはロジスティック曲線を当てはめた非線形回帰がよく応用されるが（例えば [10] など）, 今後の課題として, こうした非線形回帰に本研究で提案したような工夫を組み込んで分析を行うことが考えられる. また, 本研究では実験結果が良好であった距離 × 排気量のセグメンテーションを採用したが, 近年では整数計画法を利用してサンプルの分類と回帰分析を行う手法が提案されており [3], こうした手法を応用することも今後の課題としたい.

謝辞 研究内容に関して有意義なアドバイスをいただいた住田潮先生（筑波大学）, 後藤順哉先生（中央大学）, 生田目崇先生（専修大学）, 原稿に助言をいただいた山本芳嗣先生（筑波大学）に深く感謝を申し上げます. また, データをご提供いただき, 研究発表の際には多くの貴重なコメントをいただいたデータ解析コンペティション関係者の皆様方に心からの謝意を表します. 最後になりましたが, 問題を解くために Xpress-MP を使わせていただいた *Dash* 社に御礼申し上げます.

参考文献

[1] T.S. Arthanari and Y. Dodge, *Mathematical Programming in Statistics*, John Wiley & Sons, 1981.

[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, “Coherent Measures of Risk,” *Mathematical Finance*, **9** (1999), 203-228.

[3] D. Bertsimas and R. Shioda, “Classification and Regression via Integer Optimization,” *Operations Research*, **55** (2007), 252-271.

[4] D. Kahneman and A. Tversky, “Prospect Theory: An Analysis of Decision under Risk,” *Econometrica*, **47** (1979), 263-291.

[5] R. Koenker and G. Bassett, “Regression Quantiles,” *Econometrica*, **46** (1978), 33-50.

[6] H. Konno and R. Yamamoto, “Choosing the Best Set of Variables in Regression Analysis using Integer Programming,” Department of Industrial and Systems Engineering, Discussion Paper Series, (2007), Chuo University.

[7] W. Ogryczak and A. Ruszczyński, “Dual Stochastic Dominance and Related Mean-Risk Models,” *SIAM Journal on Optimization*, **13** (2002), 60-78.

[8] R.T. Rockafellar and S. Uryasev, “Conditional Value-at-Risk for General Loss Distributions,” *Journal of Banking & Finance*, **26** (2002), 1443-1471.

[9] A. Takeda, “Support Vector Machine Based on Conditional Value-at-Risk Minimization,” Department of Mathematical and Computing Sciences, Research Report, (2007), Tokyo Institute of Technology.

[10] 岡太彬訓, 木島正明, 守口剛 編集, 『マーケティングの数理モデル』, 朝倉書店, 2001.

[11] 末吉俊幸,『最小絶対値法による回帰分析』, *Journal of the Operations Research Society of Japan*, **40** (1997), 261-275.

[12] 古川一郎, 守口剛, 阿部誠, 『マーケティング・サイエンス入門』, 有斐閣, 2003.

[13] Google マップ, `http://maps.google.co.jp/`.