論文・事例研究

多量の入札履歴からの落札額分布推定

関　庸一，阿左美尚志

1. はじめに

ウェブ上での入札が広く行われるようになりつつあり，これに伴い，出品された商品と入札結果の情報が公開される場面も増えてきている．このような電子化された多量の入札履歴があれば，入札過程を考慮した落札額予測の可能性が広がる．

特に，落札額分布が入札前に予測できると，出品者と入札者の双方に有益な情報となる．まず，入札者にとっては，ある入札額で入札した場合，その値が最大値となる確率，つまり，落札できる確率がわかることとなる．よって，落札したい確率から入札額をいくらにすればよいかがわかることとなる．また，出品者にとっては，その落札会場に出品した場合，いくらで落札されるかが予想でき，出品すべきかどうかなどの判断に有効と考えられる．

ただし，一般には入札数と落札額のみが公表され，個々の入札額は公表されない場合が多い．そこで本研究では，競争入札について，このような落札データ中の出品商品の特徴と入札数から，落札額分布を予測するモデルとその推定方法を与え，中古車オークションの実データから提案法の有効性を検証する．

本研究の対象とする入札過程については，経済学や経営科学分野で多くの数理モデルの研究が行われている[2][4]．これらの研究の多くは，入札者に利得などの特性を考え，均衡点理論などに基づき価格特性を理論的に考察するものであり，落札額分布を具体的に与える方法は見当たらない．本研究では，入札額が二重指数分布に従うとの仮定を利用し，落札額分布を商品の特性などから個別商品ごとに具体的に推定する方法を与える．

せき　よういち，あざみ　ひさし
群馬大学 工学研究科情報工学専攻
〒376-8515 桐生市天神町 1-5-1
受付 08.7.22　採択 08.11.10

2. 極値回帰モデル

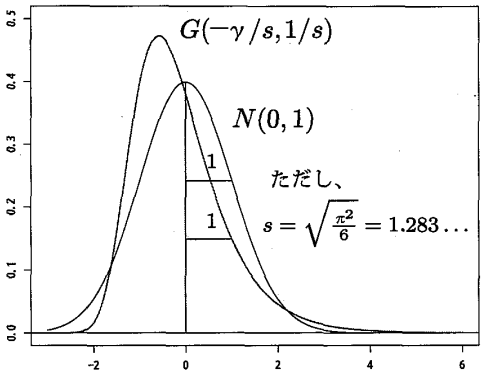
2.1 落札額の分布

第 k 回の落札 ($k=1,\cdots,K$) の落札額を Y_k，入札数を N_k とする．N_k 人中の第 i 番目の入札額を B_{ik} としたとき，落札額 Y_k は $Y_k=\max_{i=1,\cdots,N_k}B_{ik}$ と入札額の最大値となる．

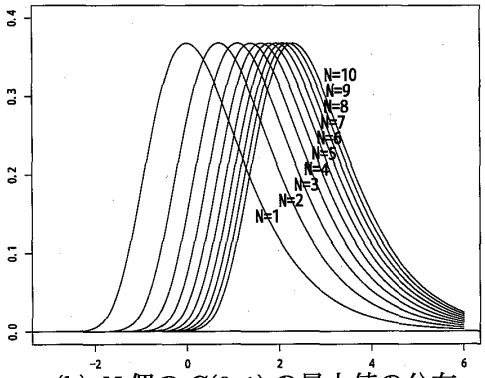
本研究では，$B_{ik}(i=1,\cdots,N_k)$ が，入札対象ごとの位置母数 μ_k と共通のスケール母数 σ をもった $G(\mu_k,\sigma)$ に，独立に従うとする．ただし，$G(\mu_k,\sigma)$ は，以下の分布関数をもつ二重指数分布である（図1(a)）．

$$\Pr(V\le v)=e^{-e^{-\frac{v-\mu_k}{\sigma}}} \tag{1}$$

この分布の期待値は，$\gamma=0.57722$（オイラー定数）として，$\sigma\gamma+\mu_k$，分散は $\pi^2\sigma^2/6$ となる[5]．



(a) 標準正規分布との平均分散を揃えての比較



(b) N 個の $G(0,1)$ の最大値の分布

図1　二重指数分布

なお，この分布の仮定では，入札者に個性がないものとしていることとなる．また，二重指数分布は指数タイプの分布[3]の極値分布であるので，各入札者の入札対象の価値評価が指数タイプの分布であり，入札時には，それぞれ同一多数回の評価結果の最大値として入札額を決定しているとすれば，入札額は二重指数分布となる．

二重指数分布は極値分布であるので，独立に同一二重指数分布に従う入札額の最大値（落札額）は，再度，二重指数分布となり，図1(c)に示すように，最大値を求める個数Nに依存して，$\log N$だけ右にシフトした分布となる．したがって，以上の分布の仮定の下で，B_{ik}の最大値であるY_kの分布関数は以下となる．

$$\Pr(Y_k \leq b)=\prod_{i=1}^{N_k}\Pr(B_{ik}\leq b)=e^{-e^{-\frac{b-\mu_k-\sigma\log N_k}{\sigma}}}$$

つまり，落札額$Y_k \sim G(\mu_k+\sigma\log N_k, \sigma)$であり，確率変数$V_k \sim G(0,1)$を用いれば，

$$Y_k=\mu_k+\sigma\log N_k+\sigma V_k \qquad (2)$$

とも表せる．その期待値と分散は$\mathrm{E}[Y_k]=\mu_k+\sigma\log N_k+\sigma\gamma$, $\mathrm{V}[Y_k]=\pi^2\sigma^2/6$となる．本研究では，この$\mu_k$が，$Y_k$の説明変数ベクトル：$\boldsymbol{x}_k=(x_{k1},\cdots,x_{kp})^t$と回帰係数：$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^t$で，$\mu_k=\boldsymbol{x}_k^t\boldsymbol{\beta}$と説明されるものとする．

2.2 モデルの推定法

前節のモデルをデータから推定するにあたっては，σと$\boldsymbol{\beta}$の2つの未知数を推定しなければならない．これには，極値分布の密度関数から尤度関数を導出し，最適化算法を用いるアプローチも考えられるが，より頑健で，大標本にも適用可能な方法として，EMアルゴリズム[1]による解法を提案する．

(2)式を変形すると，

$$Y_k-\sigma(\log N_k+\gamma)=\boldsymbol{x}_k^t\boldsymbol{\beta}+\sigma(V_k-\gamma) \qquad (3)$$

となり，右辺は期待値$\boldsymbol{x}_k^t\boldsymbol{\beta}$，分散$\pi^2\sigma^2/6$となる．そこでこの右辺を正規近似できるとみなした以下のEMアルゴリズムによって，未知数の反復推定を行う．初期値としては$V^{(0)}$としてYの標本分散などを用い，M-Step 3.から始める．

E-Step：(Expectation)

2重指数分布スケール母数$\hat{\sigma}^{(l)}$を所与として，$\tilde{Y}_k=Y_k-\sigma^{(l)}(\log N_k+\gamma)$を新たな目的変数とする．

M-Step：(Maximaization)

1. $\tilde{Y}_k$に対し$\boldsymbol{x}_k$を説明変数とした線形回帰を行い，$\hat{\boldsymbol{\beta}}^{(l)}$を定める．
2. 残差：$e_k^{(l)}=\tilde{Y}_k-\boldsymbol{x}_k^t\hat{\boldsymbol{\beta}}^{(l)}$から残差分散$V^{(l)}=\sum_k e_k^{(l)2}/(n-p-1)$を求める．
3. $V^{(l)}=\pi^2\hat{\sigma}^{(l)2}/6$を解き，スケール母数を$\hat{\sigma}^{(l+1)2}=V^{(l)}/(\pi^2/6)$と更新する．

2.3 落札額分布推定法

以上から落札額分布は$G(\boldsymbol{x}_k^t\hat{\boldsymbol{\beta}}+\hat{\sigma}\log N_k, \hat{\sigma})$と得られる．この分布の上側$\alpha$点

$$v=\boldsymbol{x}_k^t\hat{\boldsymbol{\beta}}+\hat{\sigma}\log\hat{N}_k-\hat{\sigma}\log\{-\log(1-\alpha)\}$$

で入札すれば，確率$1-\alpha$で落札できることになる．また，落札額の点推定値としては，期待値：$\hat{Y}_k=\boldsymbol{x}_k^t\hat{\boldsymbol{\beta}}+\hat{\sigma}(\log N_k+\gamma)$を用いれば良い．

予測時点で，入札数N_kが未知である場合には，データから一般化線形モデルで推定した入札数期待値をN_kとして用い，plug-in推定する．

3. 推定例

3.1 データ概要

対象データは，平成19年度データ解析コンペティションで提供された中古車オークションデータであり，2005年6月から2007年6月までの2年間に行われた125,880台の車輌の入札結果である．入札は，開札されるまで競合他社の入札額を秘密とする方法で，計614社の企業会員からの競争入札として行われていた．出品車輌は業務用に利用されたものが主であり，計105社から出品され，計17会場のいずれかに展示され，会場での入札とWeb上での入札が併用されて入札された．この入札では中古車の属性情報として，損傷状態を含む図2の車輌属性の変量群が，Web上で会員企業に公表され，Webからの入札では，この情報のみに基づいて会員企業が入札参加と値付けの判断を行ったと考えられる．なお，会員企業には，海外への輸送販売などを目的とする会社が多く，多くの場合，転売目的で入札に参加していることになる．

対象データは入札時に公開される中古車属性データと開催環境（開催日，開催会場コード，車輌の登録番号）に加え，入札結果として落札金額と入札数が与えられた．

3.2 モデル適用の方針

対象入札結果は，中古車輌に対し，専門業者が転売目的で値付けした結果であり，客観的で専門的判断のもとに入札が行われていたと想定できる．つまり，専門家にとって標準的な車輌価値が背後にあり，その均質な評価構造$G(\mu_k, \sigma)$を推定できると期待される．

落札額は，基本的にはこの残存する車輌価値で決まると考えられるが，入札数が多ければ，それだけ入札

額の最大値である落札額が高くなりやすいと考えられる（図2）．この過程について，前章までで提案した極値回帰モデルを利用することにより，落札額分布を予測する．本データの場合，Web入札者にとって入札時の判断根拠情報は，Web公表情報のみであり，以上のアプローチによりモデルを構築すれば，推定モデルに用いるデータと入札業者の値決め根拠データが同じとなり，予測力のあるモデルが期待される．

なお，入札数は，開札されれば分かる情報ではあるが，入札額の決定時点などの落札額分布が必要となる時点では未知数となる．そこで，本事例では，後述3.4節のように一般化線形モデルにより入札数の推定を行い，必要なときに入札数を与えられるようにする．なお，入札数には，車輌属性と入札時期や地域性などの開催環境で決まる需給関係のほか，オークション主催者側で変更できる要因（開催場所や落札手数料，会員増加プロモーションなど）も関連してくると考えられる．上記のアプローチによれば，入札数をオークション開催側が制御しようとした場合には，入札数を推定するモデルを修正することにより，落札額への影響も評価できることが期待される．

3.3 データクリーニング

サンプルに関しては，以下のように原データの125,880台を75,331台にクリーニングした．まず，特徴量の効果が構造的に異なると考えられるバスやトラックなどの車輌は除外した．これでサンプル数は110,194台と限定された．さらに，因子変量について，低頻度でありながら特殊な効果を持つと考えられる水準（「燃料」の水準：LPGや電気，「ミッション」の水準：CVTなど）に該当する車輌は除外した．また，利用変量に欠測がある車輌も除外した．以上で得られた75,331台のサンプルを二等分して，推定用と検証用として以下で用いた．

変量に関しては以下のようにクリーニングした．まず，車輌の色，製造メーカ，車種名など，水準数が多い因子変量は，稀にしか表れない水準を類似した水準同士統合して，水準数を減らした．また，損傷データとして，損傷位置と損傷種類の組合せ272ごとに損傷箇所数が与えられていた．これについては，入札数を予測する一般化線形モデルを暫定的に作成し，その中での各変数の影響を回帰係数から評価し，出現頻度と損傷の特徴にも配慮し，63変量に統合した．最終的に利用した変数の数を表1に示す．

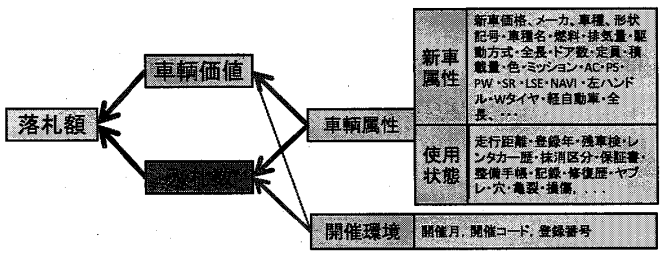


図2　変量間の関係

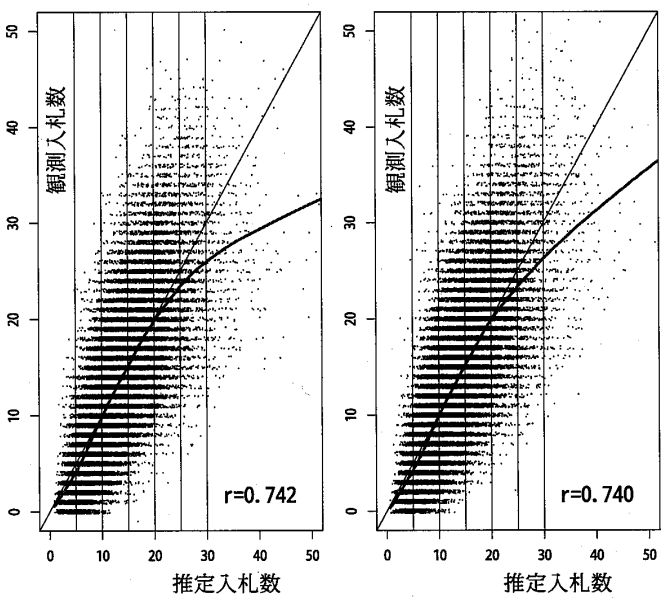
3.4 入札数予測モデル

入札数N_kを予測する一般化線形モデルでは，入札数N_kが期待値$m_k=\exp(\sum_{j=0}^{p}\alpha_j x_{kj})(k=1,\cdots,K)$のポアソン分布すると仮定し，第$j$説明変数への回帰係数$\alpha_j$を推定する．ここで，$x_{kj}$は表1の第$j$説明変数である．

表1　モデル推定に利用した変数の数

		変量数	変数数
車輌属性	新車価格	1	1
	新車属性	25	157
	使用状態	11 (3)	21(10)
	損傷状態	63	63
開催環境		3	37
定数項		1	1
計		104	280

注：変量数は、自然な変量を単位として数えた数である。変数数は、因子変量についてはダミー変数化後の変数の数で、モデルの自由度への寄与に相当する。たとえば、1つの因子変量「メーカ」は、9社を取り上げたので、変数数で8となっている。( )内は、交互作用項の数で内数



(a) 訓練データ　　(b) 検証データ

rは観測値と推定値の相関係数を示す。太線は入札数平滑化曲線

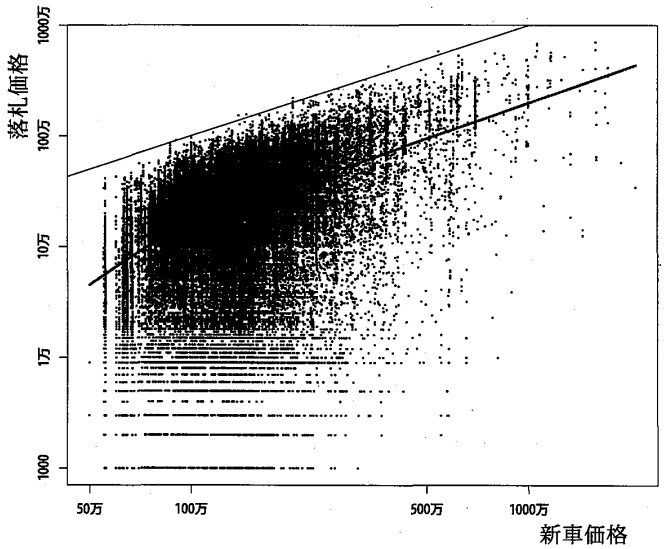
図3　入札数の推定結果

事前の予備解析から，少排気量の車輌では走行距離や登録後経過年数の増加がより強く入札数を減らす効果があることや，走行距離の増加が入札数を減らす効果がメーカにより違うことが分かった．そこで，排気量と距離，排気量と登録年，メーカと距離の交互作用効果と取り込んだモデルを利用した．以上の交互作用を含んで，用いた説明変数の数が表1である．

推定結果を図3に示す．検証データでも訓練データと遜色ない当てはまりを示している．以下では，このモデルによる推定結果 $\hat{m}_k=\exp(\sum_{j=0}^{p}\hat{a}_j x_{kj})$ を $\hat{N}_k$ として用いる．

3.5 落札額モデル

対象となる中古車は，新車価格が平均で144.7万円，最高は2,078万円，落札価格が平均で28.24万円，最高が705.9万円と，車輌間で大きな違いがあった．さらに，新車価格 p_k と落札額 Y_k の散布図（図4）を見ると，落札額の平滑化曲線は，対数軸表現で新車価格とほぼ並行であることから，落札額は，平均的には新車価格に比例した値に決まっていることがわかる（この比の全対象車輌での平均は18.9%）．そこで，落札額 Y_k が直接に二重指数分布するというモデル（G1）に追加して，落札額の新車価格比 $R_k=Y_k/p_k$ が二重指数分布するというモデル（G2）の両者を検討した．後者は(5)式を意味するが，これは(6)式なる分布を落札額に仮定することとなる．つまり，二重指数分布のスケールパラメータを p_k 倍するような効果を持つこととなる．また，比較対象のため，新車価格比 R_k に対する標準的な回帰モデル（N）も推定した．



両対数軸。太線: 落札額平滑化曲線、細線: [落札額]=[新車価格]。

図4 新車価格と落札額

それぞれのモデルと，そこでの新車価格 p_k と，入札数推定値 $\hat{N}_k$ の取扱いについては，次のようになる．

モデルG1（落札額極値回帰）：

$\boldsymbol{x}_k$ には p_k を含め，$\hat{N}_k$ を含ませず．

$$Y_k\sim G(\boldsymbol{x}_k^t\boldsymbol{\beta}+\sigma\log\hat{N}_k,\sigma) \tag{4}$$

モデルG2（新車価格比極値回帰）：

$\boldsymbol{x}_k$ には $p_k, \hat{N}_k$ を含ませず．

$$R_k=Y_k/p_k\sim G(\boldsymbol{x}_k^t\boldsymbol{\beta}+\sigma\log\hat{N}_k,\sigma) \tag{5}$$

$$Y_k\sim G(\boldsymbol{x}_k^t\boldsymbol{\beta}p_k+\sigma p_k\log\hat{N}_k,\sigma p_k) \tag{6}$$

モデルN（新車価格比標準回帰）：

$\boldsymbol{x}_k$ には p_k を含めず，$\hat{N}_k$ を含ませる．

$$R_k=Y_k/p_k\sim N(\boldsymbol{x}_k^t\boldsymbol{\beta},\sigma^2) \tag{7}$$

$$Y_k\sim N(\boldsymbol{x}_k^t\boldsymbol{\beta}p_k,(\sigma p_k)^2) \tag{8}$$

ここで，自然には入札数 $\hat{N}_k$ の効果を含まないモデルNでのみ説明変数 $\boldsymbol{x}_k$ に意識的に $\hat{N}_k$ を含ませている．上述以外の説明変数としては，入札数予測モデルと同じ表1の変数を用いた．

推定に当たっては，入札数が極端に少ない（5以下）落札事例は，推定に用いないこととした．これは次の理由による．

今回のモデルでは，入札額は二重指数分布すると仮定した．その平均つまり車輌価値は，入札参加コスト，運搬コストなどを考慮すると，非負に制限されるものではない．しかし，実際に入札を行う上では，入札システム上，正の入札額でなければ入札できない．したがって，モデルが正しいとすれば，低価値で入札数が少ない落札の事例では，落札額の二重指数分布が原点で切断された分布となっていると考えられる．落札額が負値となる事例は無札（入札札なし）となることからデータ化していない．ここで，偶然，落札された事例のみを取り上げて推定に用いると，母数推定に偏りが生ずることが予想される．そこで，落札額モデルを推定するに当たっては，少入札数の事例は，利用しないこととした．なお，少入札数として5以下という基準は，入札数を含まない説明変数で落札金額を回帰した予備解析結果の，入札数に対する残差解析から判断した．この結果では，入札数が5～10程度で，推定値が大きすぎる傾向が見られた．これが上述の効果によるものと考え，削除サンプルを少なめにしたいため5を選択した．

3.6 落札額推定分布の評価

推定された落札額分布を次のように評価する．まず，推定されたモデルから，車輌ごとに，落札額の分布 $\hat{F}_k(a)=Pr(Y_k<a|\hat{\mu}_k,\hat{\sigma},\hat{N}_k)(k=1,\cdots,n)$ が求まる．

この車輌ごとの分布の確率点と実績データ y_k を比較すると確率 p ごとに，落札価格が p 確率点を下回った割合：$H(p)=\#\{y_k \leq \hat{F}_k^{-1}(p)\}/n$ を求めることができる．この $(p, H(p))$ をプロットすると（PP-plot），分布が適切に推定できていれば，対角線となるはずである．したがって両者の解離から，予測のよさを評価する．ここでは，解離の指標として，範囲：$R=(\min_p\{H(p)-p\}, \max_p\{H(p)-p\})$ と，$|H(p)-p|$ の平均値（MA と示す）をみる．

検証用データ全中古車を対象とした評価結果を図5に示す．落札額に直接二重指数分布を仮定した極値回帰モデル（G1）では，大きなずれが生じていることが分かる．新車価格の大小で，スケール母数が変わらないと仮定しているためと考えられる．これに対し，新車価格比を用いた極値回帰モデル（G2）では，落札確率は最大3%程度のずれに収まっていることがわかる．一方，二重指数分布でなく，通常の正規分布を想定した線形回帰で推定したモデル（N）では，期待値付近で累積確率が多すぎることがわかる．これは，右裾の重い二重指数分布との分布形の相違（図1(a)参照）のためと解釈できる．

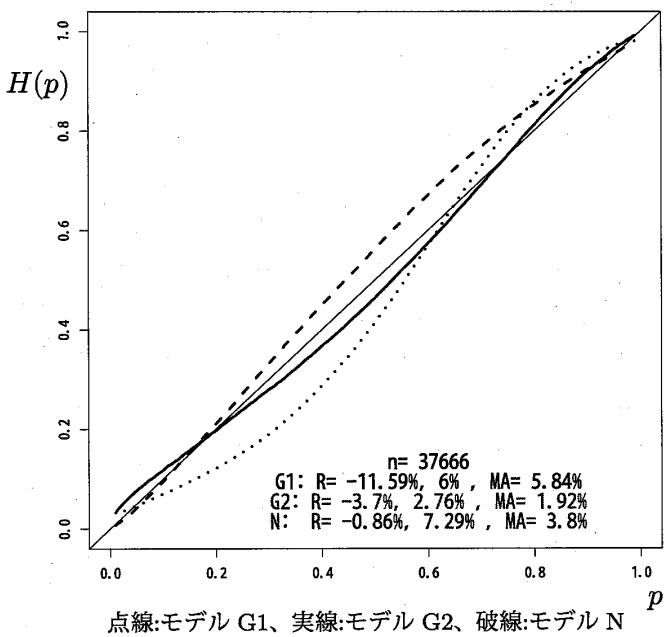
この関係を確認するため，正規分布を想定して計算した確率点での二重指数分布確率を図6に示す．図5におけるモデルNのグラフとの類似性が認められ，落札額の二重指数分布への適合性が示唆される．

以上のように，新車価格比極値回帰モデルが，対象車輌全体での当てはまりが良好であることがわかった．しかし，さまざまな中古車グループに限定しても，推定されたモデルの当てはまりに問題がないか確認する必要がある．その結果例を図7に示す．概ねどの場合も，他のモデルに比べ新車価格比極値回帰モデル（G2）で，当てはまりが良いことは変わらなかった．ただし，車種ごとにモデルの当てはまりを確認すると，3つのモデルとも当てはまりの悪い車種があった（図7(c)，(f)）．例えば，欧米車には推定された共通スケール母数 $\hat{\sigma}$ が過小であり，ADという車種にはσが過大と読み取れる．今回のモデルでは，すべてのサンプルに対してスケール母数 σ が共通であることを仮定したが，σ を適切にモデル化する必要を示唆する結果となった．

3.7 落札額モデル推定結果

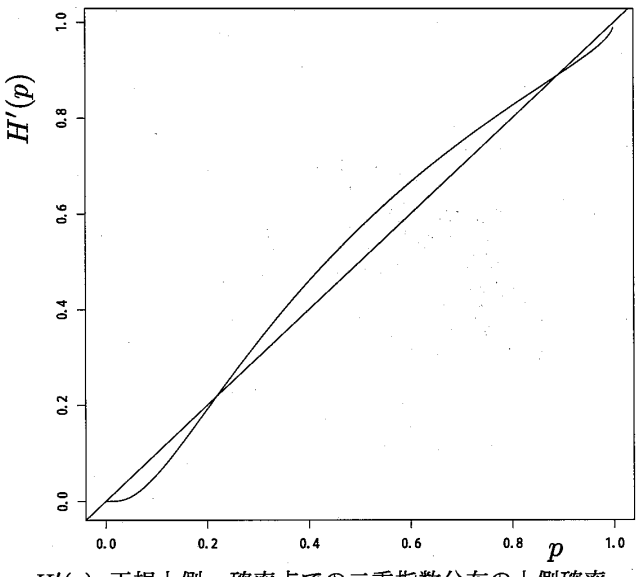
最も当てはまりのよかった新車価格比極値回帰モデルについて，その詳細を検討する．まず，新車価格比と入札数の関係を，図8に実績データと推定結果で示す．無札サンプルの多くが，推定結果では，図(b)で○プロットで示されるように負の新車価格比期待値 $\hat{R}_k=\boldsymbol{x}_k^t\hat{\boldsymbol{\beta}}+\hat{\sigma}(\log N_k+\gamma)$ をもっており，そのために無札であったと解釈される．また，入札数5以下のサンプルは，正の新車価格比期待値を持つものもあるが，負の期待値も多く，そのため正の入札額を付ける入札者が少なく，入札数が少なかったと解釈される．以上の点から，(5)式の推定において，5以下の落札事例を用いなかったのは，概ね妥当であったと考えられる．

表2に利用変数のうち，高度に有意であった説明変



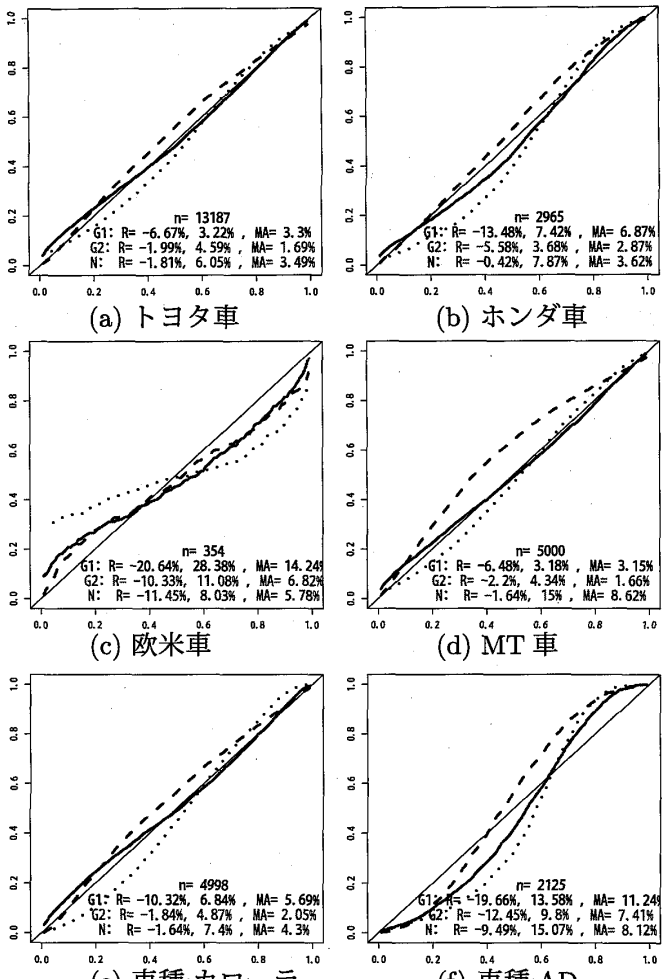
点線:モデルG1、実線:モデルG2、破線:モデルN

図5 入札額と推定入札額のPP-plot（全体）



$H'(p)$: 正規上側 p 確率点での二重指数分布の上側確率
(二重指数分布の平均分散は正規分布と等しくなるよう調整)

図6 正規分布上側 p 点での二重指数分布確率



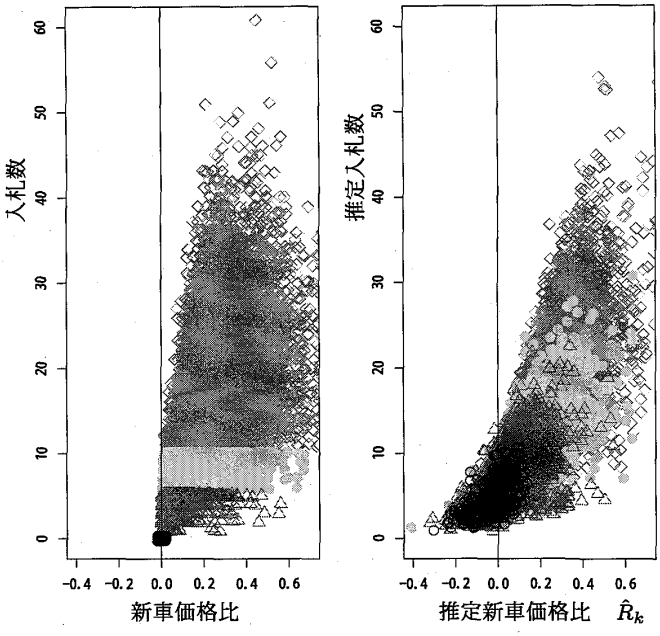
(a) トヨタ車　(b) ホンダ車

(c) 欧米車　(d) MT 車

(e) 車種:カローラ　(f) 車種:AD

点線:モデル G1、実線:モデル G2、破線:モデル N

図7　入札額と推定入札額の PP-plot（層別）



(a) 実績データ　(b) 推定結果

サンプル区分：○:無札、△:入札数 5 以下、●:入札数 10 以下 6 以上、◇:入札数 11 以上。図 (a) では、縦軸が実績値であるので、前記 4 つのプロットが下から順に並んでいる。図 (b) では、上下に拡散するとともに、左下に入り込んでいる。

図8　新車価格比と入札数

表2　モデル推定の結果で 0.1%有意な変数の数

		入札数予測モデル	新車価格比予測モデル
車輌属性	新車価格	1	-
	新車属性	104	100
	使用状態	19 (9)	14 (8)
	損傷状態	27	14
開催環境		25	9
計		176	137

( ) 内は、交互作用項の数で内数

数の数を入札数の予測モデルでの結果と対比して示す．入札数の予測では，使用した 280 変数のほとんどが高度に有意であった．一方，新車価格比モデルでは，新車時点での属性は効果のあるものが多かったが，使用状態や損傷状態，開催環境には有意でない変数も多かった．使用状態については走行距離および，そのメーカとの交互作用の効果が大きかったが，他の変数の効果は少なかった．損傷状態も入札数予測とは異なり，効果が少なくなっている．海外などへの転売価値としての車輌価値は，車輌の新車時の価値の影響が大きいが，入札時には問題リスクの少ない車に人気が集まりやすいと解釈できる．

また，開催環境においては，登録地方や会場に有意なものがほとんどなかった．ただし，開催月は，需給状況が転売価値に影響を及ぼすためか，若干有意なものがあった．

4. 終わりに

中古車事例の分析結果から，入札数の予測モデルとともに，提案した極値回帰モデルを用いることで，個々の車輌の特徴ごとの落札額分布を推定することが可能であることが示された．また，車輌属性などが，入札数や落札価格へ影響する要因効果を個別に明らかにできた．

今回のモデルでは，入札対象の価値評価の期待値が入札対象の属性等により定まる場合を扱い，価値評価の分散については，すべての入札対象で共通と仮定した．今回の事例では，新車価格比に提案モデルを当て

はめることで，この仮定の拘束を緩和することができたと考えられるが，共通の入札会場に集まるものでも，価値評価のバラツキがシステマティックに増減することが普通であると考えられる．このような現象に対しても説明力のあるようにモデルを拡張することが，今後の課題となる．また，提案した推定算法は，大標本に適用できる簡便法として，二次までのモーメントを揃えた正規近似を利用しており，近似をしない最尤法での算法の検討も今後の課題となる．

謝辞　複雑な損傷データについて，データクリーニングに努力してくれた川端聖氏に感謝する．

参考文献

[1] Dempster, A. P., Laird, N. M. and Rubin, D. B.: "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society Series B*, Vol. 39, No. 1, pp. 1-38 (1977).

[2] Engelbrecht-Wiggans, R.: "Auction and bidding models: A survey," *Management Science*, Vol. 26, No. 2, pp. 119-142 (1980).

[3] 河田, 岩井, 加瀬訳：「極値統計学」, 生産技術センター新社 (1978), (Gumbel, E. J.: *Statistics of Extremes*, Columbia University Press (1958)).

[4] Klemperer, P.: "Auction theory: a guide to the literature," *Journal of Economic Surveys*, Vol. 13, No. 3, pp. 227-286 (1999).

[5] Kotz, S. and Nadarajah, S.: "*Extreme Value Distributions*," Imperial College Press (2000).