D-016

中古車の価格順に基づく車種と年式の偏差値推定

井口 皓貴 † 髙木 寛樹 † 山岸 祐己 † 祝田 龍一 ‡ 祝田 実 ‡
† 静岡理工科大学 情報学部 ‡ 祝田石油株式会社

1 はじめに

一般に，中古車価格を精緻に推定するためには，専門家が実物を見る必要があり，事前に顧客が価格を推定することは難しい．さらに，顧客にとってはその推定価格が他の車種やメーカーなどと比較してどの程度の相場なのかが不明瞭なため，中古車の購入や売却の判断材料となる客観的な情報を提供することは重要であると言える．本研究では，中古車データを価格順にソートし，メーカーや車種，年式といった属性をカテゴリとして扱うことによって，それらの偏差値を推定するための新たな統計的指標を提案する．

2 多群 U 統計量

オブジェクトが有するカテゴリ集合を $\{1,\cdots,j,\cdots,J\}$ としたとき，全オブジェクトの平均価格を E，カテゴリ j を有するオブジェクトの平均価格を X_j，カテゴリ j を有するオブジェクトの価格の標準偏差を s_j とすれば，カテゴリ j の平均価格の z-score $z_{s,j}$ は，カテゴリ j を有するオブジェクト数 m_j を考慮した標準化 [1] によって

$$z_{s,j} = \frac{X_j - E}{s_j/\sqrt{m_j}}, \tag{1}$$

のように算出することができる．以下ではこの標準化 z-score に対し，順位和 z-score [2] を拡張したカテゴリ j の価格指標を提案する．

あるデータにおける観測順序集合と，それらが有するカテゴリ集合をそれぞれ $\mathcal{N}$ と $\mathcal{J}$ とする．ここで，それぞれの要素数は $N = |\mathcal{N}|$ と $J = |\mathcal{J}|$ とし，各要素は整数と同一視されるとする．つまり，$\mathcal{N} = \{1,\cdots,n,\cdots,N\}$ および $\mathcal{J} = \{1,\cdots,j,\cdots,J\}$ である．なお，オブジェクト n は，ある基準における最下位のものが 1，最上位のものが N となるよう並んでいるものとする．このとき，観測順序 n がカテゴリ j を有する場合は 1，それ以外の場合は 0 となっている J 行 N 列の行列を Q $(q_{j,n} \in \{0,1\})$ とすると，オブジェクト n が有するカテゴリ数は

$$t_n = \sum_{i=1}^{J} q_{i,n}, \tag{2}$$

観測順序 n までのカテゴリ j の出現数は

$$I_{j,n} = \sum_{i=1}^{n} q_{j,i}, \tag{3}$$

観測順序 n までの全カテゴリの総出現数は

$$I_n = \sum_{i=1}^{J} I_{i,n}, \tag{4}$$

のように表せる．いま，オブジェクトに付随してカテゴリが出現するとし，以降では，オブジェクト出現からカテゴリ出現へと視点を変える．このとき，オブジェクト n が唯一のカテゴリのみ有する $t_n = 1$ の場合では，オブジェクト n に付随して出現したカテゴリ j の出現順位は $r_n = I_{n-1} + 1$ であるが，複数のカテゴリを有する $t_n > 1$ の場合では，平均順位を考えなければならないため，その出現順位は $r_n = I_{n-1} + (1 + t_n)/2$ となる．ここでの目的は，観測順序とカテゴリの集合が与えられたとき，大きい（上位の），または逆に小さい（下位の）観測順序が有意に多く含まれるカテゴリを定量的に評価する指標の構築である．

Mann-Whitney の二群検定で用いられる U 統計量 [2] を多群に拡張し，カテゴリの出現順位に適用する方法について述べる．いま，カテゴリ j に着目すれば，このカテゴリに属する集合 $\{n \in \mathcal{N} : q_{j,n} = 1\}$ と，このカテゴリに属さない集合 $\{n \in \mathcal{N} : q_{j,n} = 0\}$ の二群に分割することができる．よって，Mann-Whitney の U 統計量に従い，次式により，カテゴリ j に対し U 統計量の z-score を求めることができる．

$$z_j = \frac{u_j - \mu_j}{\sigma_j}. \tag{5}$$

ここで，統計量 u_j, 出現順位の平均 μ_j，および，その分散 σ_j^2 は次のように計算される．

$$u_j = \sum_{i=1}^{N} r_i q_{j,i} - \frac{I_{j,N}(I_{j,N}+1)}{2}, \tag{6}$$

$$\mu_j = \frac{I_{j,N}(I_N - I_{j,N})}{2}, \tag{7}$$

$$\sigma_j^2 = \frac{I_{j,N}(I_N - I_{j,N})}{12}\left((I_N+1) - \sum_{i=1}^{N}\frac{t_i^3 - t_i}{I_N(I_N-1)}\right). \tag{8}$$

すなわち，u_j は順位和に基づくカテゴリ j の U 統計量であり，その平均と分散が μ_j と σ_j^2 である．ただし，各オブジェクトが複数のカテゴリを有し得ないケース

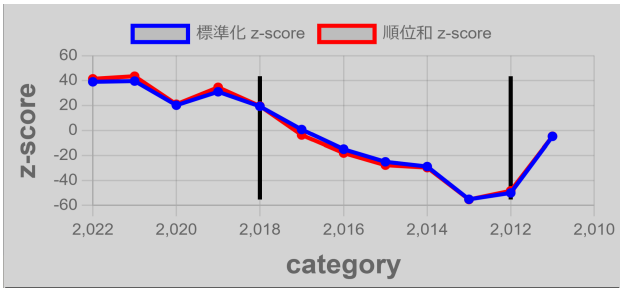
T-Score Estimation of Car Models and Years Based on Price Order of Used Cars
†Koki IGUCHI †Hiroki TAKAGI †Yuki YAMAGISHI ‡Ryuichi Hoda ‡Minoru Hoda
†Shizuoka Institute of Science and Technology
‡Hoda Oil Inc.

では，式 (8) の t_i を含む項，すなわち平均順位を扱うための補正値の計算は不要である．この多群 U 統計量は，基本的には 2 クラス分類器の SVM (Support Vector Machine) [3] を多クラス分類器に拡張するときに利用される one-against-all と類似した考え方となる．

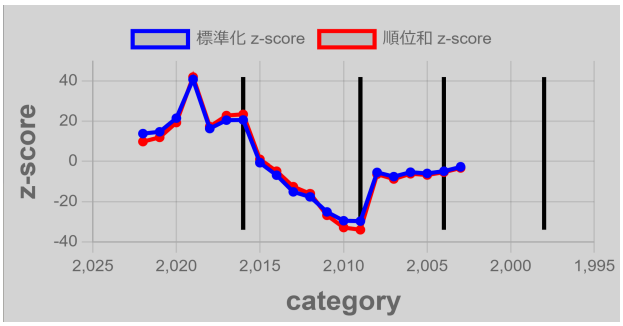
3 評価実験とまとめ

今回分析の対象とした中古車価格サイトは株式会社リクルート (https://www.recruit.co.jp/) が運営するカーセンサー (https://www.carsensor.net/) である．取得したデータセットは中古車 472,695 台，93 メーカー，2,011 車種を含む．以下では，社内システムに組み込んだことを想定し，実際にウェブブラウザでインタラクティブに描画している可視化結果を使用する．本実験では年式をカテゴリ j とし，各車種における年式ごとの標準化 z-score $z_{s,j}$ と順位和 z-score z_j を比較する．どちらの z-score も，$10z_* + 50$ とすれば一般的な偏差値（T-score）として利用が可能である．図 1a に軽自動車の中で最も台数が多かった「ホンダ N-BOX（16,791 台）」の結果を，図 1b に普通乗用車の中で最も台数が多かった「トヨタ プリウス（7,983 台）」の結果を，図 1c に小型乗用車の中で最も台数が多かった「日産 ノート（7,786 台）」の結果を，図 1d にスポーツカーの中で最も台数が多かった「トヨタ 86（1,800 台）」の結果をそれぞれ示す．なお，各図の黒の実線が示すのは各車種でフルモデルチェンジが行われた年月である．

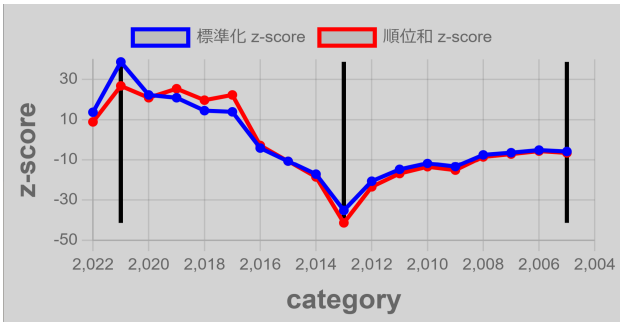
図 1 より，標準化 z-score と順位和 z-score はどのカテゴリにおいてもある程度類似した結果となった．基本的には年式が新しいほど z-score が高くなる傾向があるが，フルモデルチェンジの前後で値が大きく変動するなど，両 z-score は各車種に関する様々な要因を読み取るための重要な指標であることが見て取れる．例えば，「トヨタ プリウス（図 1b）」は車検が切れたと思われる 2019 年で両 z-score が最高となっており，高価格で大量に売りに出されていることが示唆される．同様な現象は「トヨタ 86（図 1d）」でも見られるが，2019 年は順位和 z-score のほうが僅かに値が高く，「日産 ノート（図 1c）」においても 2019 年は順位和 z-score の方が高い．それに対し，標準化 z-score は「日産 ノート（図 1c）」の 2021 年で極端に高くなるなど，外れ値の影響を受けやすい懸念があるため，順位和 z-score は比較的頑健な評価指標であることがうかがえる．



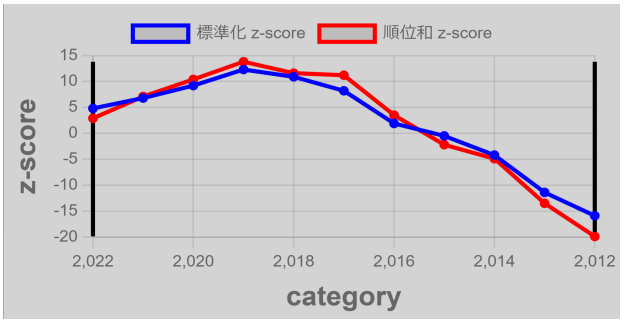
(a) ホンダ N-BOX (2022 年 – 2011 年)



(b) トヨタ プリウス (2022 年 – 2003 年)



(c) 日産 ノート (2022 年 – 2005 年)



(d) トヨタ 86 (2022 年 – 2012 年)

図 1: 年式をカテゴリとしたときの提案 z-score

参考文献

[1] Douglas G Altman and J Martin Bland. Standard deviations and standard errors. *BMJ*, Vol. 331, No. 7521, p. 903, 2005.

[2] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, Vol. 18, No. 1, pp. 50–60, 03 1947.

[3] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.