キャップレート推計の精度向上に関する一考察

2020 年 8 月 19 日

株式会社三井住友トラスト基礎研究所
私募投資顧問部　主任研究員　米倉勝弘

投資用不動産への投資判断においては、取引キャップレートを用いたキャップレート推計モデルが広く活用されているものと思われる。キャップレート推計モデルとしては線形回帰が用いられていることが多いが、目的変数であるキャップレートと説明変数（特徴量）である価格形成要因は必ずしも直線的な関係に従うとは限らない。築年数の経過による経年劣化や最寄り駅までの距離など、キャップレートと非線形の関係にあると推測される要因がいくつも存在する。

そこで、本稿ではキャップレートを目的変数とし、価格形成要因の特徴を抽出した合成変数を説明変数（特徴量）とする単回帰モデルを作成し、当該単回帰モデルを多項式に拡張することで目的変数と説明変数（特徴量）における非線形な関係を表現した。これに Ridge 回帰・Lasso 回帰を適用して過学習を抑制することで、フィッティングの精度向上が図れることを確認した。

本稿では、分析結果を二次元で分かりやすく表現することを優先し、単回帰モデルを前提モデルとして採用したが、本稿の結果を踏まえれば、説明変数が 2 次元以上の重回帰モデルを多項式に拡張する場合にもフィッティングの精度向上が期待できるものと考えている。

1. はじめに

日々行われている投資用不動産の取引においては、各取引当事者が様々な方法を用いて算出したフェアバリューに基づき投資判断が行われている。フェアバリューの把握にあたっては、マーケットで観測された取引キャップレートを用いて構築したキャップレート推計モデルを活用することが多い。通常、推計モデルとしては線形回帰が用いられるが、目的変数であるキャップレートと説明変数（特徴量）である価格形成要因は必ずしも直線的な関係に従うとは限らない。価格形成要因のなかには、キャップレートと非線形な関係にあると考えられる要因はいくつもある。たとえば、「築年数の経過による経年劣化」や「最寄り駅までの距離」などを挙げることができよう。「築年数の経過による経年劣化」は、新築からの 1 年、築 10 年後からの 1 年、築 20 年後からの 1 年、築 50 年後からの 1 年では、それぞれキャップレートの上昇に影響する傾きが一律ではないと思われる。また、「最寄り駅までの距離」についても駅直結・駅至近に対する効用は高い一方で、徒歩 15 分と徒歩 16 分では効用に有意な差がない可能性もある。そこで本稿では非線形な関係性に対応するためモデルを多項式回帰に拡張し、正則化により過学習を抑制する Ridge 回帰および Lasso 回帰を適用のうえ、モデルの精度向上に関して考察を行った。

2. 前提モデルの構築

本稿では、二次元による視覚的な比較を容易にするため、前提となるモデルとして単回帰モデルを用いる。目的変数としては 2018 年～2019 年上期において東京 23 区内で観測されたオフィスの取引キャップレートを、説明変数（特徴量）としては次に説明する価格形成要因を用いている。

＜前提となるモデル（単回帰）＞

$$y_i = w_0 + w_1 x_i + \varepsilon_i (i = 1,2,\cdots,n)$$

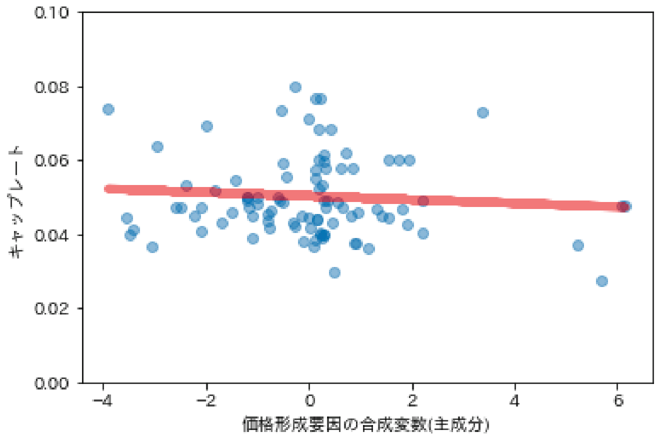
y ：キャップレート（目的変数）
x ：価格形成要因の第一主成分得点（説明変数：特徴量）
n ：サンプル数

ここで用いる価格形成要因は、その構成要素である「行政的要因（容積率、用途地域等）」、「ビルスペック要因（延床面積、築年数等）」、「立地要因（最寄り駅までの距離、中心部までの距離等）」、「市況要因（マーケット賃料水準等）」を主成分分析により1変数（第一主成分）に集約した合成変数である。ただし、第一主成分の寄与率は約30％（各要因の約30％を集約した合成変数となる）程度であり、これに基づくモデルの精度には限界があるが、前述のとおり二次元での議論を優先するため前提モデルの精度自体は劣後させるものとする。

また、本稿では多項式回帰に拡張した際の過学習（オーバーフィッティング）を把握するため、収集した取引キャップレートを訓練データ（8割）とテストデータ（2割）とに無作為で分割し、訓練データのみを用いて構築したモデルによってテストデータによる推計値の検証を行う。なお、過学習とは訓練データだけに最適化されてしまい、未知のデータに対して汎用性がない状態に陥ることである。過学習により、訓練データでは観測値と推計値の誤差が小さいものの、テストデータに当てはめると誤差が大きくなるという現象が発生する。

図表1は、訓練データにおけるキャップレートと価格形成要因との関係をプロットし、回帰直線を示したものである。キャップレートと価格形成要因の合成変数との間には僅かに負の相関関係が認められ、価格形成要因の主成分得点が高いほどキャップレートは低い傾向にあることが確認された。

図表1_キャップレートと価格形成要因の関係と回帰直線（赤線）



出所）三井住友トラスト基礎研究所

3. 多項式回帰への拡張

多項式回帰とは、目的変数を説明変数（特徴量）の n 次多項式でモデル化する手法である。高次の項を追加することにより目的変数と説明変数（特徴量）の非線形な関係を表現できる。モデルが複雑化することで訓練データにおけるフィッティングは向上するが、訓練データへの最適化を優先するあまり過学習が生じる懸念がある。

すなわち、多項式特徴量（説明変数）の導入によりリッチな表現が可能となる一方で、どの程度リッチにすれば良いのかが問題となるということである。回帰分析では観測値と推計値の誤差（RMSE：平均平方二乗誤差）を最小にすることをモチベーションとしている。多項式回帰では次数を増やすにつれて訓練データにおける誤差は小さくな

るが、テストデータに当てはめると誤差が大きくなるという現象がしばしば起きる。このようなケースにおいては過学習が生じているものと推察される。訓練データへのフィッティングを重視し過ぎるあまり、モデルとしての汎用性を欠いてしまうのである。

＜多項式回帰モデル＞

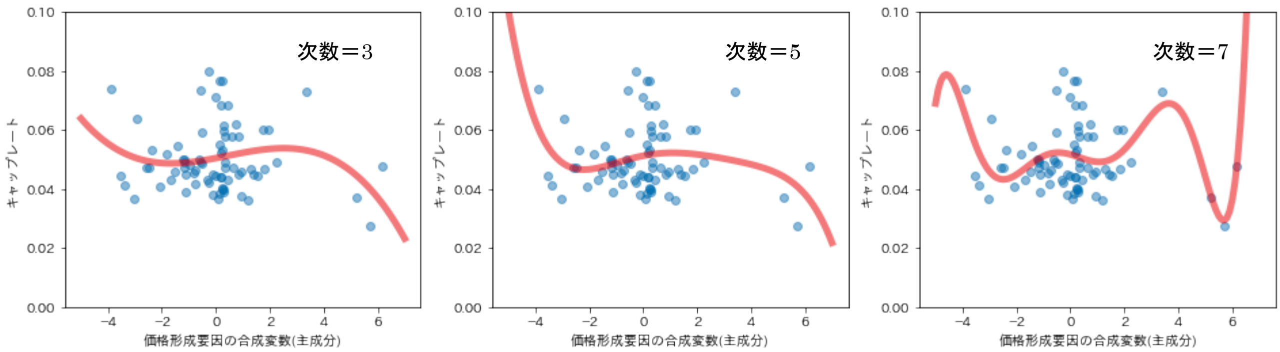
$$y_i = w_0 + w_1 x_i + w_2 x_i^2 \cdots + w_m x_i^m + \varepsilon_i (i = 1,2,\cdots,n)$$

y ：キャップレート（目的変数）
x ：価格形成要因の第一主成分得点（説明変数：特徴量）
n ：サンプル数
m ：次数

なお、近年注目されているニューラルネットワークを用いた深層学習では、シグモイド関数や ReLU 関数等の活性化関数を用いて非線形変換を行い複雑な構造を表現している。多項式回帰では、高次の項にて目的変数と説明変数（特徴量）の非線形な関係を表現しており、目指す方向性としては類似している部分があるといえる。

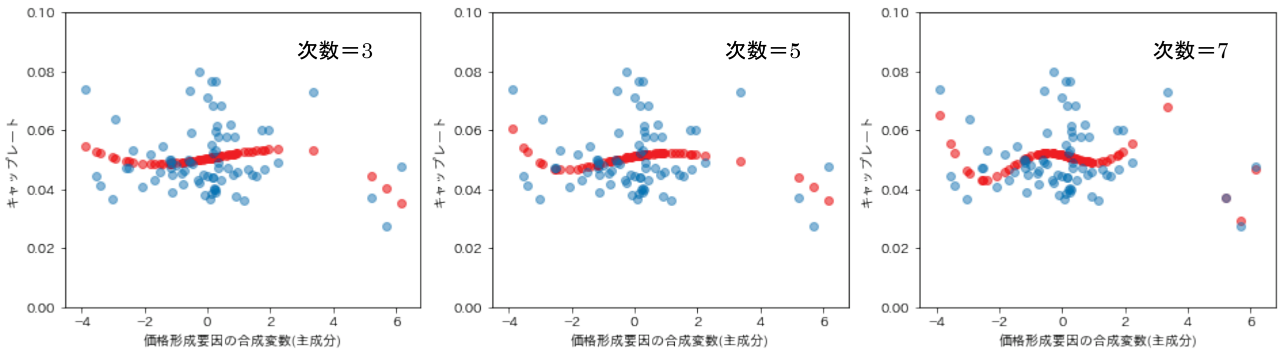
図表 2 および図表 3 は前提モデル（単回帰）を多項式回帰に拡張した結果を示したものである。多項式回帰の次数を左から 3 次、5 次、7 次と増やしているが、高次になるにつれて価格形成要因の主成分得点の変化に応じて回帰曲線の変曲点が多く発現しており、訓練データへのフィッティングが高くなっているのがわかる。モデルの自由度が高い多項式回帰の RMSE は、単回帰の RMSE よりも小さく、次数を増やすにつれて訓練データにおける RMSE は小さくなっている（図表 4）。一方で、多項式回帰（7次）では価格形成要因の主成分得点が上がるとキャップレートが高くなったり低くなったりと不安定な動きをしている（図表 2、図表 3）。

図表 2_多項式回帰の回帰曲線（赤線）



出所）三井住友トラスト基礎研究所

図表 3_多項式回帰による推計結果（赤点）



出所）三井住友トラスト基礎研究所

図表 4 で多項式回帰の RMSE を見ると、次数が増えるにつれて訓練データの RMSE は小さくなっている一方でテストデータの RMSE は大きくなっており、過学習（訓練データへのフィッティングを高め過ぎて、未知のデータに対応できていない）していることがわかる。

そこで、損失関数に正則化項（過剰適合に対するペナルティ）を追加して最適解を求めることで過学習を抑制する手法である Ridge 回帰および Lasso 回帰を適用して、過学習の抑制効果を確認する。

図表 4_単回帰と多項式回帰の RMSE 比較

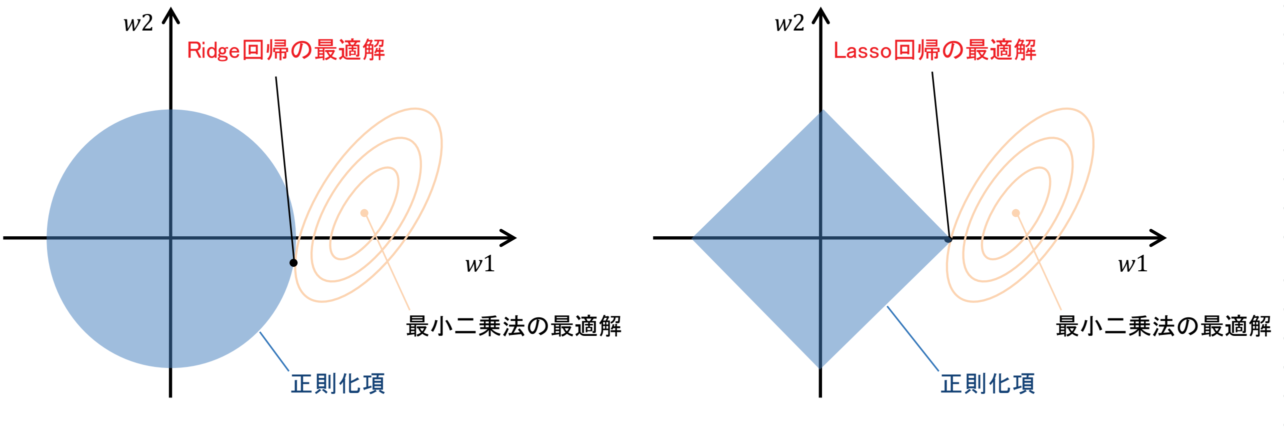
	単回帰	多項式回帰 次数＝3	多項式回帰 次数＝5	多項式回帰 次数＝7
RMSE(訓練データ)	1.107%	1.078%	1.071%	1.013%
RMSE(テストデータ)	1.400%	1.062%	1.074%	1.084%

出所）三井住友トラスト基礎研究所
注）RMSE は平均平方二乗誤差

4. Ridge 回帰および Lasso 回帰の適用

Ridge 回帰および Lasso 回帰は、最小二乗法の損失関数に正則化（過学習を抑制して汎化性を高める技術）を取り入れた分析手法である。過学習の状態にある場合には係数 *w* が大きくなることが知られており、*w* のベクトルの長さを抑えて最適解を導くこととなる（図表 5）。図表 5 は係数w_1と係数w_2の関係を例にとり二次元でイメージを図式化したものである。正則化を行うことによって Ridge 回帰・Lasso 回帰における最適解は、最小二乗法の最適解に比べて原点からの距離が短縮されていることが確認できる。なお、ベクトルの長さとして Ridge 回帰では *L2* ノルムを、Lasso 回帰では *L1* ノルムを用いる。*L2* ノルムは$\sqrt{w_1^2+\cdots+w_n^2}$と平方根をとるので二乗して$w^T w$とするのが一般的である。Ridge 回帰では係数の絶対値を小さくすることで過学習が抑制され、Lasso 回帰では変数選択により過学習が抑制される（図表 5（右）では *w2* がゼロのところで Lasso 回帰の最適解となっており、*w2* は変数から除外される）。

図表 5_Rigde 回帰と Lasso 回帰の損失関数（*L*）



出所）三井住友トラスト基礎研究所

＜Rigde 回帰の損失関数（*L*）＞

$$L = \sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \lambda\|\boldsymbol{w}\|_2^2$$

L ：損失関数　n ：サンプル数　y ：目的変数
w ：特徴量の係数（ウェイト）　x ：特徴量
λ ：正則化の強さ（ハイパーパラメーター）
$\|\boldsymbol{w}\|_2^2$：*L2* ノルムの二乗（ユークリッド距離）

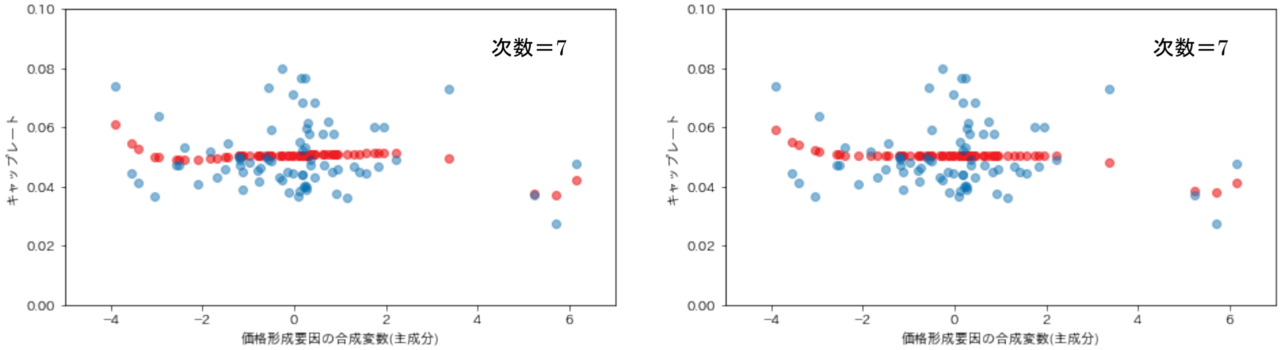
＜Lasso 回帰の損失関数（*L*）＞

$$L = \sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \lambda\|\boldsymbol{w}\|_1$$

L ：損失関数　n ：サンプル数　y ：目的変数
w ：特徴量の係数（ウェイト）　x ：特徴量
λ ：正則化の強さ（ハイパーパラメーター
$\|\boldsymbol{w}\|_1$：*L1* ノルム（マンハッタン距離）

図表 6 は多項式回帰（7 次）について Ridge 回帰および Lasso 回帰を適用した結果を示したものである。正則化により過学習が抑制されていることが見て取れる。

図表 6_Ridge 回帰（左）、Lasso 回帰（右）による推計結果（赤点）



出所）三井住友トラスト基礎研究所

図表 7 をみると訓練データの RMSE が大きくなっているものの、テストデータの RMSE が小さくなっており、過学習が抑制されていることが数値でも確認できる。

図表 7_多項式回帰（ノーマル）・Ridge 回帰・Lasso 回帰の RMSE 比較

	多項式回帰 次数＝7	Ridge回帰 次数＝7	Lasso回帰 次数＝7
RMSE(訓練データ)	1.013%	1.058%	1.074%
RMSE(テストデータ)	1.084%	1.012%	1.003%

出所）三井住友トラスト基礎研究所
注）RMSE は平均平方二乗誤差

また、図表 8 はそれぞれのモデルにおける係数 w を示したものである。Ridge 回帰は係数の絶対値を小さくして過学習を抑制するため、多項式回帰（7 次）に比べて Ridge 回帰（7 次）における係数の絶対値の方が小さくなっていることが確認できる。また、Lasso 回帰は変数選択により過学習を抑制するため、重要でないと判断された変数の係数をゼロとしており（図表 8 の青色部分）、変数選択が行われていることが確認できる。

図表 8_モデルの係数 w（ウェイト）比較

	多項式回帰 次数＝7	Ridge回帰 次数＝7	Lasso回帰 次数＝7
x^1	−0.0025440	0.0000369	0.0000000
x^2	−0.0018770	−0.0000093	0.0000000
x^3	0.0017130	0.0002129	0.0000543
x^4	0.0003510	−0.0000289	−0.0000087
x^5	−0.0001490	−0.0000236	−0.0000126
x^6	−0.0000110	0.0000029	0.0000012
x^7	0.0000030	0.0000001	0.0000001

出所）三井住友トラスト基礎研究所

4. キャップレート推計の精度向上に向けて

ここまで見てきたようにキャップレートと価格形成要因の間に非線形な関係が認められる場合には、多項式回帰への拡張や正則化による過学習の抑制によりモデルの推計精度が向上する（RMSE を小さくできる）ことが確認できた。本稿では視覚的な理解を優先するため二次元での例示を行ってきた。しかし実際は、価格形成要因には、先に挙げた「築年数の経過による経年劣化」や「最寄り駅までの距離」などのように、キャップレートと非線形な関係

にあると考えられる要因はいくつもあり、このような複数の説明変数（特徴量）を用いた重回帰モデルについて多項式による特徴量の拡張を行うこととなる。

重回帰モデルを多項式に拡張する場合には、説明変数（特徴量）が大きく増加するため、ある程度のサンプル数の確保が求められるが、本稿の結果と同様にフィッティングの精度向上が期待できるものと思われる。本稿では前提となるモデルとして合成変数による単回帰モデルを用いたため、マーケット環境が類似したサンプル（2018 年～2019 年上期において東京 23 区内で観測されたオフィスの取引キャップレート）を採用したが、取引時点・取引エリアなどを広げてサンプルを収集することにより重回帰モデルにおける多項式への拡張は可能である。

一方、モデルが複雑になることで、説明変数の変化による影響の解釈や説明が難しくなることも考えられるが、実務上は、線形回帰において大枠を把握しつつ、推計精度の高いモデルを合わせて判断材料とすることの意義は大きいと考えられる。なお、過学習を防止しつつ非線形に対応した機械学習の手法として、他にもランダムフォレストやサポートベクター回帰などが挙げられるが、機械学習や深層学習が比較的容易に行える環境が整ってきている現在においては、複数のモデルを比較考量したうえで意思決定をしていくことが合理的な判断になり得るのではないだろうか。

【本件のお問い合わせ先】

私募投資顧問部　担当：米倉
TEL：080-7207-5117(直)
https://www.smtri.jp/contact/form-private/index.php

株式会社三井住友トラスト基礎研究所
〒105-8574　東京都港区芝 3-33-1 三井住友信託銀行芝ビル 11 階
https://www.smtri.jp/