■■連載記事■

データをビジネスに活用する実践アナリティクス

<第 13 回>　住宅価格の予測モデル

梶山昌之
株式会社ワイハット

前回の連載では 10 名の成人男性の身長と体重のデータを用いて，身長と体重の関係を分析しました．

回帰分析をわかりやすく理解するため，身近な数値である身長と体重の例を取り上げたものと理解してください．

今回は住宅価格の予測の問題を例にして，実務で活用できる予測モデル構築の方法について解説します．

ただし，この方法を理解するためには，いくつかの前提知識も必要になります．そのため，前回までの連載内容と重複する部分もありますが，重要なポイントについては再度解説します．

1．関係性のモデル化

前回の連載で，身長と体重の関係を回帰分析の手法を用いて計算しましたが，回帰式が使用可能かどうかを示す数値（p 値）は 0.051 となり，回帰式は有意ではないという結論になりました．身長と体重は明らかに正の相関がある量ですので，このような結果になったのは，データ数が少なかったためです．

しかしながら，これが有意となった場合でも，得られた関係式が，ただちに身長と体重の関係を表すものとは考えられません．

なぜなら，回帰分析では，説明変数と目的変数は直線的関係があることを前提としているためです．

身長は高さのみが関係する量であるのに対し，体重は体の幅や奥行きも関係する量であるため，直線的な関係になるとは考えられません．

肥満度を示す指標に BMI がありますが，この値は 22 で最も病気になる確率が低いと言われています．身長から理想体重を求める式は，

[理想体重] = 22 * [身長]^2

となります．

ここで，身長の単位はメートル(m) です．例えば身長 170cm の方の理想体重は 63.6 kg (=22・1.7^2) になります．

この式では，身長と体重の関係は 2 次関数になっています．この様に身長と体重の関係は 2 次関数で表すのが適切なのかもしれません．

本来，自然は複雑な関係性を持つものかと思いますが，それをモデル化する（数式で表す）場合，最も基本的な近似の方法が直線的な関係として表すことです．

今回は住宅価格予測の問題を取り上げることにします．一生の内には住宅の売買を経験する可能性があると思いますが，そのときに是非身に着けておいて欲しい知識です．

2．母回帰式と母平均

連載の第 11 回では DM の発送数と成約数の関係について考えました．その時に，発送数 x とその成約数 y の母平均 η（イータ）には，次の関係

$$\eta = \alpha + \beta x$$

が存在すると考えました．この関係式が母回帰式です．

ここで，α および β は未知の定数であり確率変数ではありません．発送数 x がある値のときに成約数 y が観測されますが，y は発送数以外の要因にも影響されて変動します．

しかし，任意の発送数 x における y の平均は試行を無限に繰り返すことにより，一つの値に近づきます．その値を η で示しています．すなわち，任意の x に対する y の母平均が η です．

α，β は未知の母数ですので，私たちはこの値を，観測されたデータから推定することになりま

す．その推定値を a, b とします．従って，α，β を a, b で置き換えた値は η の推定値になります．

$$\hat{\eta} = a + bx$$

ここで，y ではなく η という文字を使いましたが，これは，y の母平均の推定値であり，y の推定値と区別するためです．

ここは，理解しにくい部分ですが，次のように考えてください．

「真の回帰式が存在することはわかるが，そのパラメータは観測された値から計算するしかない．この値は観測の都度変動する量であるため，母平均の推定値もバラツク量である．」

すなわち，予測値 y のバラツキを考えているのではなく，y の母平均の推定値のバラツキについて考えています．

後ほど，その具体的な内容の違いをグラフで示しますので，その時に，上記の記述の意味を，再度考えてみてください．

3. 回帰分析適用の条件

回帰分析では残差平方和を最小にするように，回帰式のパラメータ a, b を求めます．具体的には最小 2 乗法を用いますが，この方式では残差について以下の仮定を置いています．

残差の平均は 0 である（不偏性）
残差の分散はどの x においても一定（等分散性）
残差は正規分布に従う（正規性）
任意の残差は他の残差と互いに独立である（独立性）

これらの仮定が満たされている状況を図 1 で示します．
この図ではベル型の分布（正規分布）が残差の分布を表しています．また，この分布は x の任意の点で標準偏差が等しく，平均は母平均に一致していることを示しています．

残差については，回帰分析が適用可能かどうかを判断する条件として前述の制約がありますが，説明変数 x の分布については特に制約はありません．

身長と体重の分析の例では，身長は正規分布でしたが，それはランダムにサンプリングされた集団の身長の分布が正規分布に従っていたためです．

従って，身長を 10cm 単位で区分して各区分から 5 名ずつサンプリングするという方法でも，身長と体重の関係が得られるでしょう．この場合，x は一様分布になります．

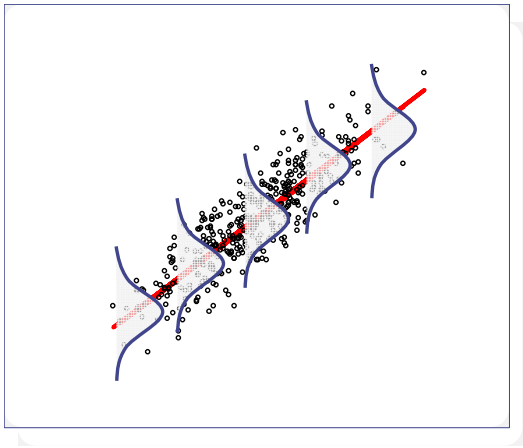

図 1　残差に対する仮定

ただし，x が任意の分布でもよいとしても，分析の対象とする x の範囲で極端にデータ数に偏りがある場合は，等分散性や正規性を確認できないという点で問題があります．

反対に，x の範囲にわたって均等にデータがある場合には，全体としてデータ数が少なく（10 程度）でも，意味のある関係式が得られる可能性が高いと考えられます．

4. 定数項と回帰係数の標準誤差

身長と体重の回帰分析では a+bx というモデルを考えました．これは x の一次関数ですので a は切片，b は傾きと呼ぶことが多いと思います．

単回帰分析では，説明変数が一つですが，一般には説明変数は複数あると考えられます．その場合を重回帰分析といいます．

例えば，説明変数が 2 つの場合は，x_1, x_2 を説明変数としたモデル

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

を考えることになります．

ここで，b_0 は切片ですが，本連載では，これを定数項と呼ぶことにします．b1, b2 は傾きに相当するものですが，単回帰分析では回帰係数，重回帰分析では偏回帰係数と呼ばれます．

定数項や回帰係数はどのような分布になるでしょうか．証明はやや複雑ですが，これらの分布は

正規分布になります.

今回は，説明変数が 1 つの場合

$$\hat{y} = b_0 + b_1 x$$

で考えます.

定数項 b_0 は確率変数ですので，その標準偏差 $s(b_0)$ を計算することができます．ここで，本来は標準偏差を表す s の添え字として b_0 を表示することになるのですが，非常に読みにくくなるため，ここでは，b_0 をカッコ内に表示しています.

また，推定値の標準偏差は標準誤差と呼ばれますので，以下，標準誤差の用語を用います.

$s(b_0)$ の導出はやや難しいので，結果のみを示すと

$$s(b_0) = \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \cdot se$$

となります．b_0 の標準誤差は x の変域が広く，データ数が多いほど小さくなり，x の平均が大きいほど大きくなることがわかります.

b_1 の標準誤差については，連載の第 11 回で解説しましたが，b_0 に倣い結果のみを示すと

$$s(b_1) = \sqrt{\frac{1}{S_{xx}}} \cdot se$$

となります. b_1 の標準誤差は x の変域が広いほど小さくなります.

上記の $s(b_0)$, $s(b_1)$ のいずれの式でも，残差標準偏差 se が使われていますが，これは 残差の母標準偏差 σ は未知なので，その推定値である残差標準偏差 se で置き換えた結果です.

その結果, b_0, b_1 を標準化した値は正規分布ではなく，t 分布に従うことになります．このことについては，連載の第 7 回で解説しました.

5. 住宅価格の予測

皆様は住宅の売買を行ったことがあるでしょうか．ここでは，東京の T 町にある築 10 年のマンションで 4LDK（92 平方メートル） の部屋にお住まいの方が，住み替えのためにマンションの売却を検討しているものとします.

まずは，幾つかの不動産屋に相談して，売買価格を査定してもらうと驚くことがあります．それは，どの不動産屋の査定額も異なることです. 4000 万～6000 万円まで大きな差があります. これでは，どの売出し価格が妥当なのかわかりません.

低い価格であれば，買い手がつくのは早いでしょうが，本来得られる利益を失うことになります．高すぎると，いつまでたっても売れない状況になってしまいます.

不動産屋は最初に高い価格での売り出しを提案するかもしれませんが，売れない状況がしばらく続くと，値下げを勧めてきます.

そのときに自分自身が価格に対する判断の基準を持っていないと，不本意な価格での売買に同意してしまうかもしれません.

そこで，インターネットからから T 町における中古マンションの物件価格の情報をダウンロードして，専有面積と価格の関係を調べることにしました.

住宅価格の変動要因には，専有面積以外に，築年数や駅からの徒歩分数などありますが，ここでは専有面積と価格の関係ついて分析することにします.

後日の連載で，専有面積以外の価格変動要因を反映したモデル化を検討する予定ですが，その場合でも，最も価格に最も影響がある要因と考えられる専有面積との関係を確認しておくことは，分析の手順として必須です.

収集した専有面積と価格のデータを表 1 に示します．この表では実際の売り出し価格と回帰分析で推定した価格（推定価格）との差（残差）も表示しています.

上記のデータを回帰分析した結果を表 2 および図 2 に示します．計算の手順については，連載の第 10 回から第 12 回までを，必要に応じて参照してください.

分析の結果，専有面積 x と価格 y には，以下の関係があることがわかりました.

$$\hat{y} = -741.4 + 63.3x$$

また，寄与率は 81.3% ですので，専有面積は，価格を説明する重要な要因であることがわかります.

表 1 専有面積と売出し価格，推定価格および残差

No	専有面積[m²]	価格[万円]	推定価格	残差
1	48.60	2490	2,335	155
2	56.52	2980	2,837	143
3	69.60	2980	3,665	−685
4	70.87	3590	3,745	−155
5	52.00	2480	2,551	−71
6	75.03	3580	4,009	−429
7	63.20	2980	3,260	−280
8	74.17	4750	3,954	796
9	52.00	2580	2,551	29
10	56.47	2380	2,834	−454
11	50.00	1980	2,424	−444
12	55.66	3080	2,782	298
13	75.56	3480	4,042	−562
14	106.24	7000	5,984	1,016
15	57.14	3500	2,876	624
16	57.14	3600	2,876	724
17	57.75	2630	2,915	−285
18	60.65	2080	3,098	−1,018
19	63.50	2980	3,279	−299
20	56.16	2780	2,814	−34
21	61.60	2780	3,158	−378
22	40.64	2250	1,831	419
23	40.64	2250	1,831	419
24	110.98	5740	6,285	−545
25	32.40	1380	1,310	70
26	75.43	4980	4,034	946
平均	62.31	3203.08	3203.08	0.00
標準偏差	17.54	1231.46	1110.37	532.51
歪み	1.240	1.569	1.240	0.314
尖り	2.400	2.861	2.400	−0.612

図 2 を観測すると，回帰式では説明できない変動があることが明らかになります．それが残差であり，残差は築年数や駅からの徒歩分，商業施設に近いか，その他多くの変動要因が重なりあった結果です．

ここで，連載の第 6 回で解説した「和の分布は正規分布に近づく」という原理を思い出してください．残差が多くの変動の和であるならば，その値は正規分布にとなることが想定できます．

そこで，表 1 の残差の列で歪み尖りの量を確認するといずれも±1.5 の範囲内であり，残差は正規分布に従っていると見做してよいという結果になります．

さて，売却対象の物件の専有面積は 92m² ですので，この値を予測式に入力すると，5083 万円という予測値が得られます．

モデルの作成に使用したデータは現時点で売り出されてる物件の価格ですので，予測された結果も，どの程度の価格で売り出すのが平均的な価格なのかを示しています．

実際には成約価格は売り出し価格よりも下がることが多いようですので，そのことも考慮して売出し価格を設定する必要があります．

しかし，あまりに高い売出し価格で開始すると買い手がつかなくなりますので，妥当な価格を設定する必要があります

表 2　専有面積と価格の回帰分析

【基本統計量】

特性値	面積	価格
データ数	26	26
平均	62.30576923	3203.076923
中央値	57.445	2980
標準偏差	17.53905155	1231.460171
歪み	1.240498242	1.569447645
尖り	2.399524631	2.860591084

【単回帰分析】　モデル: y=b0+b1*x

	b0	b1
回帰係数	−741.4083812	63.30850823
標準誤差	400.5825235	6.197526565
t 値	−1.85082558	10.21512495
p 値	0.076541351	3.23841E−10

【分散分析】

変動因	平方和 S	自由度 f	分散 V
回帰 R	30823104.47	1	30823104.47
残差 e	7089249.376	24	295385.3907
全体 T	37912353.85	25	
分散比 Fo	104.3487777		

【その他の統計量】

偏差平方和 Sxx	7690.458235
残差標準偏差 se	543.4936896
寄与率 R^2	81.3%

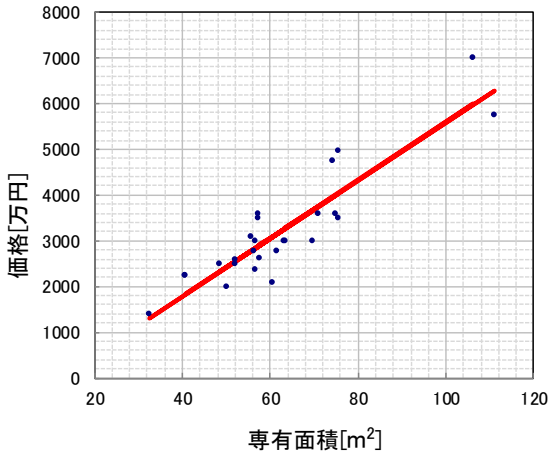

図 2　専有面積と売り出し価格

.

6. 回帰式の信頼区間

前章で回帰式における定数項と回帰係数の標準偏差が計算できましたので，x の任意の値における y の母平均 η がどの範囲に入るのかを計算することができます．

ここで，設定した範囲に 90% が入るならば，その範囲は信頼度 90% の信頼区間と呼ばれます．この信頼区間を求めるためには，η の推定値の標準偏差を計算する必要がありますが，これは b_0 と b_1 のバラツキを反映した量になります．これを導出する計算はやや複雑ですので，ここでは結果のみを示します．

$$s(\hat{\eta}) = \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \cdot se$$

この式から，母平均の標準誤差は x の平均から離れるほど大きく，x の変域が広く，データ数が多いほど小さくなることがわかります．残差の標準偏差は x の値とは無関係に一定という仮定がありますので，x の平均から離れるほど標準誤差が大きくなるというのは，一見矛盾するように感じるかもしれません．

これは x の平均を中心として得られたモデルからは，x の平均から大きく離れた値では正確な予測が難しくなる（誤差が大きくなる）と考えれば理解できるかと思います．

今，任意の x の値における η の推定値の分布を考えます．この分布で平均を中心として上下限の範囲に 90%が入るような区間はどのように求めたらよいでしょうか．

この場合，まずは η の推定値を標準化して，標準化された分布で確率の計算を行うのが常套手段です．

η の推定値は正規分布に従う量ですので，その母標準偏差が既知であれば，標準化した値も正規分布に従います．しかし，η の推定値の標準誤差の計算では，残差の母標準偏差の代わりに，その推定値である残差標準偏差を使用しています．その結果，η の推定値を標準化した値は，正規分布ではなく自由度 n-2 の t 分布に従うことになります．

文章による説明ですとわかりにくいと思いますので，具体的な計算方法を示します．

専有面積が 92m^2 における η の推定値と標準誤差は以下の通りです．

推定値 = 5083[万円]
標準誤差 = 212.7[万円]

90%の信頼区間とは上側の 5%,および下側の 5%はその区間に含まれないという意味ですので，自由度 n-2 の t 分布における下側確率 95%の標準化得点を求めれば上限の値が得られることになります．この値は，Excel の T.INV 関数（t 分布左側確率逆関数）を用いて，

T.INV(0.95,n-2)=T.INV(0.95,24)=1.711

と計算できます．この標準化得点から，90%信頼区間は

上限 = 5083 +1.711・212.7 = 5447[万円]
下限 = 5083 - 1.711・212.7 = 4719[万円]

となります．

上記と同様の計算を すべての x について計算した上限値を結ぶと曲線が現れます．下限についても同様の計算を行うと，信頼区間の上限，下限を示す関数は双曲線（破線）になります（図 3）．

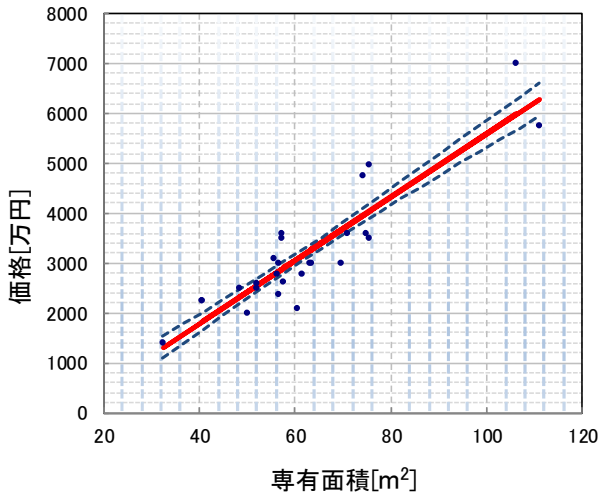

図 3　回帰式の信頼区間

図 3 から，予測対象の専有面積がデータの平均から離れるほど信頼区間の幅が広くなっているのが確認できます．

7. 予測区間

前章で求めた y の母平均の信頼区間は，その区間の中に母平均が入る確率が 90%あるという意味であり，この区間の中に y の 90%が入るという意味ではありません．

専有面積が 92m^2 のマンションは売出し価格の推定値は 5083 万円になることがわかりましたが，

私達が実務上の必要性から知りたいのは売出し価格の 90%が入る区間です.

残差は母平均の推定値（y の推定値）に対するバラツキですので，残差標準偏差から 90%信頼区間を求めればよいように思いますが，そうではありません.

何故なら母平均の推定値もバラツキを持つ量ですので，そのバラツキも加えたバラツキで信頼区間を求める必要があります.

2 つの要因でバラツク場合, 全体のバラツキは, 個々のバラツキを合計すれば得られます.

$$\{s(\hat{y})\}^2 = \{s(\hat{\eta})\}^2 + \{se\}^2$$

この関係から，y の推定値の標準誤差は

$$s(\hat{y}) = \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \cdot se$$

となります.

以上の様にして価格の推定値の標準誤差を求めることができますが，この値から推定値の 90%信頼区間を求めるための考え方は，回帰式の信頼区間を求めた場合と同様です.

専有面積が 92m^2 における y の推定値の 90%信頼区間は以下の用に求めることができます.

y の推定値 = 5083[万円]
y の推定値の標準誤差 = 583.6[万円]
上限 = 5083+1.711*583.6 = 6081[万円]
下限 = 5083-1.711*583.6 = 4084[万円]

図 3 に予測区間を追加すると図 4 が得られます. この図では外側の細い破線が予測区間の上限と下限を表しています.

統計学の教科書には，回帰式の信頼区間の記述はあるが, 予測区間の記述がない場合があります.

そのため, 回帰式の信頼区間を個々の y の値の信頼区間と勘違いして使用してしまうことがありますのでご注意ください．不動産会社から提示された売出し価格には, 4000 万円～6000 万円の幅がありましたが，駅からの徒歩分，築年数などを無視して予測すれば，この程度の幅があるのかもしれません.

住宅価格の予測では専有面積のみで 81%の寄与率がありましたので，5083 万円の推定値は概算として参考になる値です.

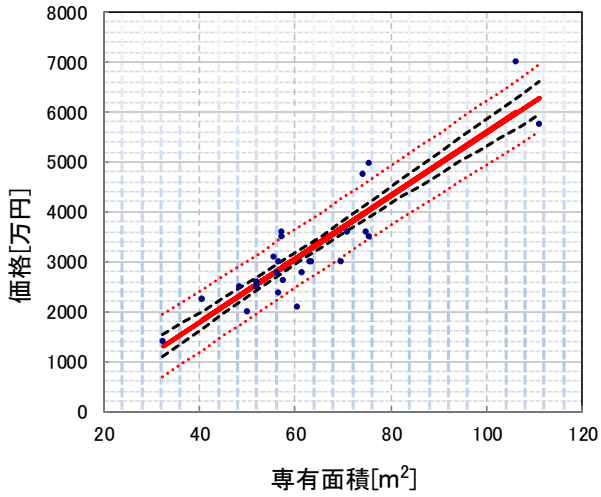

図 4　予測区間

また，今回は信頼度を 90%として設定しましたが，専有面積のみで判断する限り 90%の上限価格 6081 万円は，かなり高いと感じることになり，最初から検討の対象から外されてしまうかもしれません.

そこで，信頼度を 70%にすると，上限価格は 5701 万円になりますが，この価格であれば，やや高い程度ですので，購入を検討してもらえることが期待できます.

今回分析では，専有面積のみを説明変数として取り上げませんでしたので，推定値に対して，±1000 万円程度の変動があります.

これは，専有面積以外の要因も価格に影響していることを意味します．本連載では，それらの要因も取り入れて，より精度の高いモデルを作成する予定ですのでご期待ください.

参考文献

今泉忠, 田村義保, 中西寛子, 美添泰人 (2015). 日本統計学会公式認定 統計検定 2 級対応 統計学基礎. 東京図書.

奥野忠一, 久米均, 芳賀敏郎, 吉沢正 (1981). 多変量解析法（改訂版）. 日科技連.

末吉正成, 末吉美喜 (2017). EXCEL ビジネス統計分析 [ビジテク] 第 3 版. 翔泳社.