重回帰分析について

今回の豆知識ではエクセルによる重回帰分析の方法を紹介します。重回帰分析は、湿度や気温などのデータからビール販売量を予測するなど、複数の要因(**説明変数**)からあるデータ(**被説明変数**)を予測する分析手法です。この重回帰分析は日常生活でどのように役に立つのか、以下では中古マンションの販売価格の重回帰分析を例にあげて説明します。

A：「甲駅の周辺に 4,800 万円の中古物件が売りに出ていて、興味があるんだけど、これって買いなのかな。最近マンション価格が上がっているでしょ。多少高くても仕方ないとは思うけど、あまりにも高い値段で掴まされるのは避けたいんだよ。」

B：「どんなマンション？」

A：「甲駅まで徒歩 5 分、最寄りスーパーまで徒歩 5 分、築年 25 年、階層が 7 階、65 ㎡、大手マンションデベロッパーの〇〇シリーズだよ。」

B：「ひと昔前なら同じ値段で新築マンションが買えたね。でも、最近、中古マンション市場は過熱しているし、甲駅は△市のど真ん中にあるから、それぐらいするんだろうな。感覚で判断するのも心許ないので、エクセルの回帰分析機能を使って、収集した販売情報からお目当てのマンション価格・相場を予測するのも手だね。インターネットで甲駅周辺のファミリータイプの中古マンションの販売情報を集めてよ。」

A：「40 件見つかった。販売情報からは、①駅距離、②階層、③マンションのブランド、④最寄りスーパーまでの距離、⑤築年数が読み取れそうだ。」

B：「重回帰分析で採用する要因数の目安は、**10 事例に 1 要素程度**だから、今回の場合、分析できる要因は 4 要因程度だよ。重回帰分析をする場合、分析する要因が多いと分析結果の信頼性が低下してしまう危険性があるし、分析結果を検証するためにも要因を選別することが重要なんだけど、手始めに中古マンションの販売情報から読み取れる 5 要因すべてを使って重回帰分析してみようか。まず、エクセルで募集情報を整理してみて。」

A：「徒歩 7 分とか、築 5 年とかは数量で入力できるけど、大手ブランドマンションの有無は、どのように入力すればいい？」

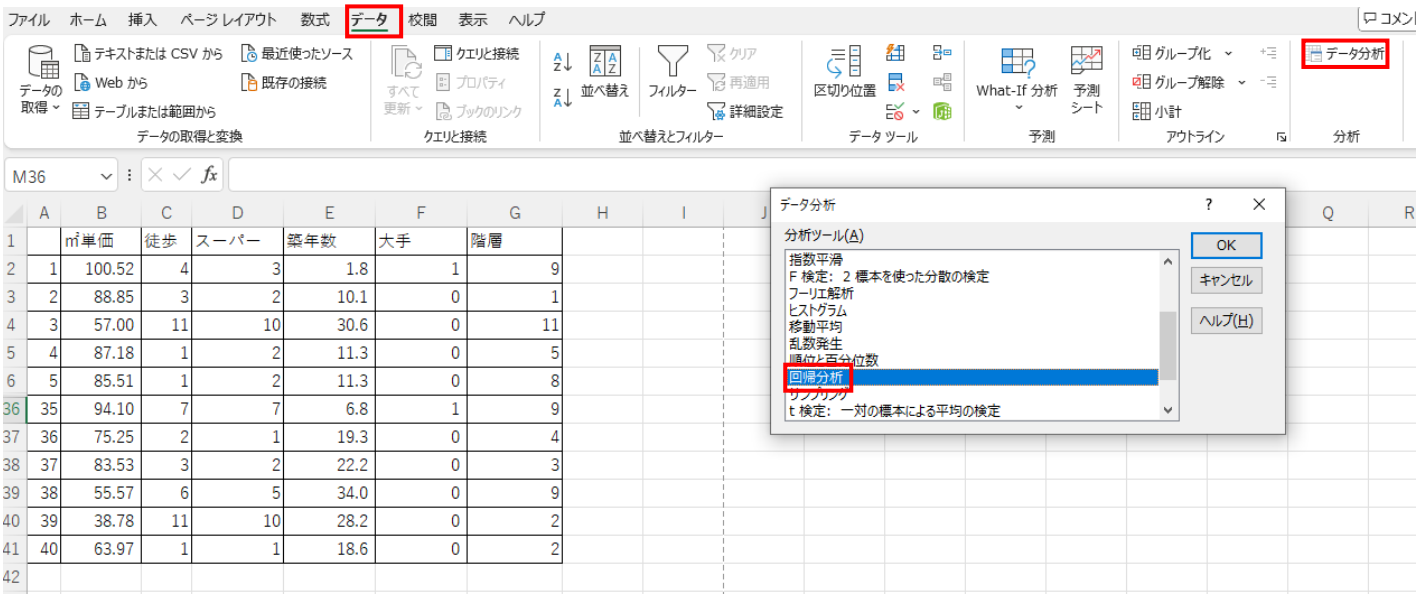
B：「大手ブランドか否かは、数量で表現できないよね。そのような場合には**ダミー変数**を使うよ。ここでは大手のブランドマンションの場合には『1』、大手のブランドマンションではない場合には『0』を入力してみて」

【図 1】

	㎡単価	駅	スーパー	築年数	大手	階層
1	100.52	4	3	1.8	1	9
2	88.85	3	2	10.1	0	1
3	57.00	11	10	30.6	0	11
4	87.18	1	2	11.3	0	5
5	85.51	1	2	11.3	0	8
			〜			
35	94.10	7	7	6.8	1	9
36	75.25	2	1	19.3	0	4
37	83.53	3	2	22.2	0	3
38	55.57	6	5	34.0	0	9
39	38.78	11	10	28.2	0	2
40	63.97	1	1	18.6	0	2

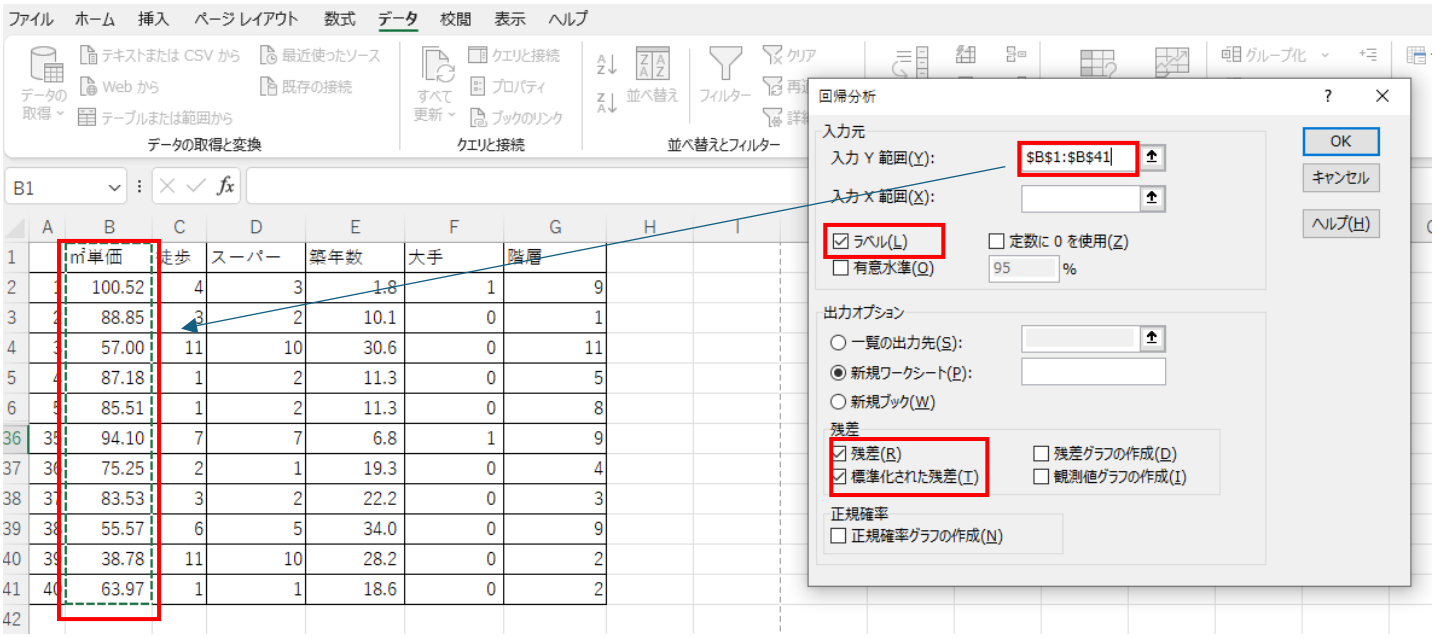
A：「できた。こんな感じでいい(図 1)？」

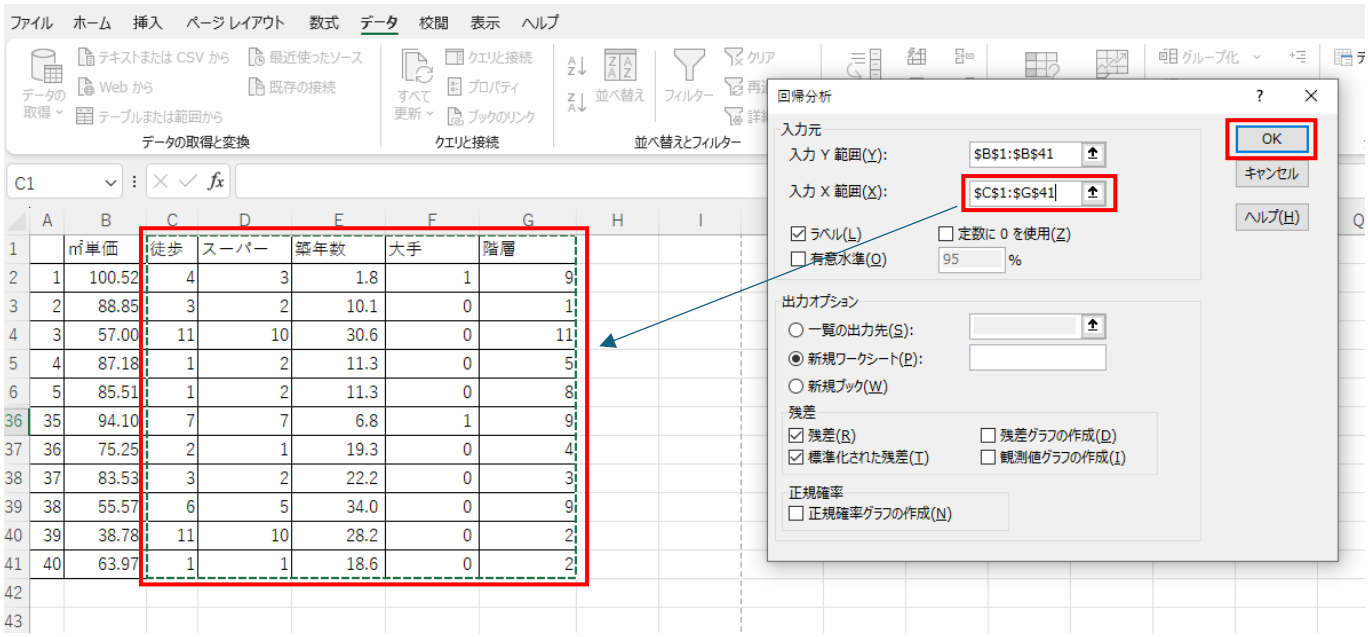
B：「ＯＫ。あとはエクセルの**アドイン機能**※を使って分析するだけだ。『データタブ』→『データ分析』→『回帰分析』をクリックして、『入力 Y 範囲(Y)』に**被説明変数**の㎡単価のデータ範囲を、『入力 X 範囲(X)』に**説明変数**の駅、スーパー、築年数、大手、階層のデータ範囲を指定。タイトル部分も含めて指定してもいい。ただし、『ラベル(L)』にチェックを忘れないで。それから『残差(R)』と『標準化された残差(T)』にもチェックだ。」



	A	B	C	D	E	F	G
1		㎡単価	徒歩	スーパー	築年数	大手	階層
2	1	100.52	4	3	1.8	1	9
3	2	88.85	3	2	10.1	0	1
4	3	57.00	11	10	30.6	0	11
5	4	87.18	1	2	11.3	0	5
6	5	85.51	1	2	11.3	0	8
36	35	94.10	7	7	6.8	1	9
37	36	75.25	2	1	19.3	0	4
38	37	83.53	3	2	22.2	0	3
39	38	55.57	6	5	34.0	0	9
40	39	38.78	11	10	28.2	0	2
41	40	63.97	1	1	18.6	0	2
42							

（注）便宜上、行 7～35 を非表示にしていますが、実際にはすべてを表示して操作します。





	A	B	C	D	E	F	G
1		㎡単価	徒歩	スーパー	築年数	大手	階層
2	1	100.52	4	3	1.8	1	9
3	2	88.85	3	2	10.1	0	1
4	3	57.00	11	10	30.6	0	11
5	4	87.18	1	2	11.3	0	5
6	5	85.51	1	2	11.3	0	8
36	35	94.10	7	7	6.8	1	9
37	36	75.25	2	1	19.3	0	4
38	37	83.53	3	2	22.2	0	3
39	38	55.57	6	5	34.0	0	9
40	39	38.78	11	10	28.2	0	2
41	40	63.97	1	1	18.6	0	2
42							
43							

A：「分析したら変な表(図 2)が出てきたけど、これってどういう意味？」

【図 2】

概要

回帰統計	
重相関 R	0.893031
重決定 R2	0.797504
補正 R2	0.767725
標準誤差	8.532216
観測数	40

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	5	9748.077	1949.615	26.78090205	6.9E-11
残差	34	2475.156	72.79872		
合計	39	12223.23			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	94.61047	4.20519	22.4985	5.34898E-22	86.06449	103.1564	86.06449	103.1564
駅	-1.55473	2.126752	-0.73103	0.469766845	-5.87681	2.767352	-5.87681	2.767352
スーパー	0.482167	2.311912	0.208557	0.836038125	-4.2162	5.180538	-4.2162	5.180538
築年数	-1.29636	0.156373	-8.29018	1.1246E-09	-1.61414	-0.97857	-1.61414	-0.97857
大手	12.66962	4.336843	2.921392	0.006151296	3.856093	21.48315	3.856093	21.48315
階層	0.717369	0.450332	1.59298	0.120420286	-0.19781	1.632553	-0.19781	1.632553

B：「注目すべきは**係数**だね。ここから㎡単価を求めるための計算式(回帰式)が分かるよ。

今回の分析では、

㎡単価＝（－1.55473）　×駅徒歩距離＋
（＋0.482167）×スーパー徒歩距離＋
（－1.29636）　×築年数＋
（＋12.66962）×大手ブランド＋
（＋0.717369）×階層＋
（＋94.61047）

という計算式が求められたことになる。

この式に君が注目しているマンションの基本情報(駅距離徒歩 5 分、スーパー徒歩距離 5 分、築年数 25 年、大手ブランド、階層 7 階)を入力すると、マンションの㎡単価を求めることができるんだ。

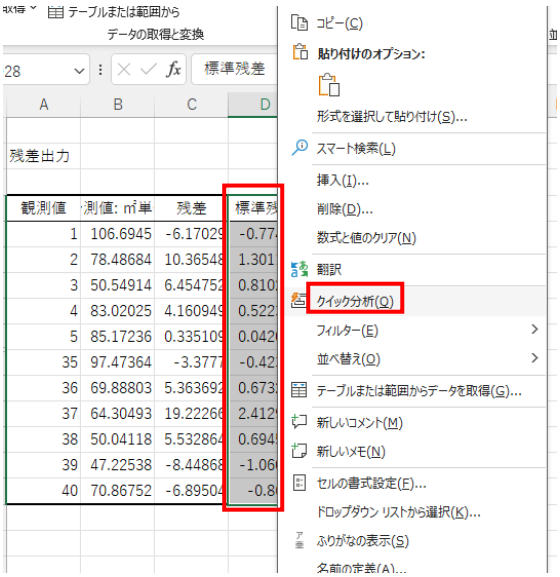
実際に入力して計算すると

㎡単価＝（－1.55473）　×駅徒歩距離(5 分)＋
（＋0.482167）×スーパー徒歩距離(5 分)＋
（－1.29636）　×築年数(25 年)＋
（＋12.66962）×大手ブランド(1)＋
（＋0.717369）×階層（7 階）＋
（＋94.61047）　　≒74.5 万円/㎡
総額：74.5 万円/㎡×65 ㎡≒4,843 万円

ということになる。」

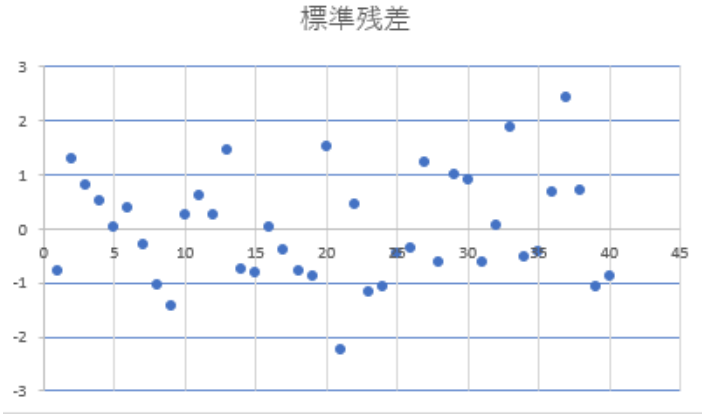
A：「狙っている物件の販売価格が 4,800 万円だから、概ね妥当ってことだね。」

B：「結論を出すのはまだ早い。**外れ値**があると結論が大きく変わってしまうから、念のため採用した事例に外れ値がないのかチェックしよう。先ほどの表の下には標準残差が表示されている。これを選択して右クリック。『クイック分析』→『グラフ』→『散布図』で標準残差の散布図を作成するとこうなる(図 4)。



(注)便宜上、観測値 6～34 は非表示となっているが、実際にはすべてを表示した上で操作を行う。

【図 4】



一般に**標準残差が±2.5 を超えるもの**は、外れ値である可能性があるんだ。今回の分析結果は概ね±2.5 の範囲内に収まっているから、外れ値が含まれている可能性は低そうだね。それに、標準残差の分布に一定の傾向があると、重回帰分析が有効でない可能性がある。だけど、今回は**ゼロを挟んでプラスとマイナスに満遍なく分布**していて一定の傾向が見当たらないから、重回帰分析しても問題なさそうだ。」

B：「次に、今回求めた回帰式の精度を確認するよ。精度について**決定係数**の**補正 R２**を見て。決定係数としては重相関 R、重決定 R２もあるけど、**重回帰分析の場合は補正 R２を使う**んだ。この数字が 1 に近いほど精度が高いといえる。今回は 0.767725 となっている。

回帰統計	
重相関 R	0.893031
重決定 R2	0.797504
補正 R2	0.767725
標準誤差	8.532216
観測数	40

これは事例の約 77%がこの回帰式で説明できたことを意味しているんだ。一般に**補正 R2 が 0.6 以上**であれば回帰式を利用してもよいといえるから、なかなかいい結果だと思うよ。」

A：「他に確認することは？」

B：「t 値と p 値の確認も必要だ。 **t 値**は、求めようとしている値(被説明変数)、今回はマンションの㎡単価に対する各要因(説明変数)の**影響力**を示していて、数字が大きいほど影響力が大きい。**p 値**は要因の**有意性**を意味していて、値が小さいほど有意性が高い。統計の慣習では**「t 値は±2 以上、p 値は 0.05 以下」が目安**となっているよ。」

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	94.61047	4.20519	22.4985	5.34898E-22	86.06449	103.1564	86.06449	103.1564
駅	-1.55473	2.126752	-0.73103	0.469766845	-5.87681	2.767352	-5.87681	2.767352
スーパー	0.482167	2.311912	0.208557	0.836038125	-4.2162	5.180538	-4.2162	5.180538
築年数	-1.29636	0.156373	-8.29018	1.1246E-09	-1.61414	-0.97857	-1.61414	-0.97857
大手	12.66962	4.336843	2.921392	0.006151296	3.856093	21.48315	3.856093	21.48315
階層	0.717369	0.450332	1.59298	0.120420286	-0.19781	1.632553	-0.19781	1.632553

A：「あれ。**駅距離**と**スーパーの距離**それから**階層**の t 値が 2 以下、 p 値は 0.05 を超えている。」

B：「そうだね。マンション価格と駅距離は相関関係があるはずなのに、t 値が基準を下回るのは変だね。それにスーパーの係数はプラスになっている。スーパーまでの距離が遠いほどマンションの値段があがるという意味だけど、これは非常識な分析結果だ。おそらく**マルチコ**が発生して、重回帰分析の精度が落ちているはずだ。」

A：「マルチコって何？」

B：「**多重共線性**ともいうんだけど、要因間に強い相関関係がある場合に重回帰分析の精度が落ちてしまう現象のことだ。」

A：「マルチコが発生しているかはどうやって確認するの？」

B：「最初に作成した表（図 1）を開いて、『データタブ』→『データ分析』→『相関』を選択して、『入力範囲(I)』にデータの入力部分を選択。この時、先頭のタイトルも選択してもいいけど、『先頭行をラベルとして使用』にチェックを忘れないで。そうするとこんな図(図 5)が出てくる。この図は、各要因間の相関関係を示していて、絶対値が 1 に近いほど要因間の相関関係が強いことを意味している。要因間の相関関係をチェックしているわけだから、各要因と㎡単価との相関関係は無視してね。」

（注）便宜上、行 7～35 を非表示にしていますが、実際にはすべてを表示して操作します。

【図 5】

	㎡単価	駅	スーパー	築年数	大手	階層
㎡単価	1					
駅	-0.27104	1				
スーパー	-0.22985	0.974224	1			
築年数	-0.82368	0.163054	0.125163	1		
大手	0.539902	0.039144	0.057413	-0.34234	1	
階層	0.119645	0.146224	0.153174	0.096671	0.313377	1

A：「駅とスーパーの数字が 0.974224 と異常に高い相関関係になっている。そういえば駅のそばにスーパーがあるから、駅までの距離とスーパーまでの距離との相関関係が強くなったんだね。」

B：「互いに相関関係の強い要因がある場合、**目安として 0.9 以上の場合**には、さっき説明したようにマルチコが発生してしまうから、相関関係の強い**二つの要因を合成した新たな要因を作る**か、**いずれかの要因を分析から外す**必要がある。今回は、一般的にマンション価格と相関関係があると考えられている駅距離を残して、スーパーまでの距離を外そう。要因を 4 つにして再度、重回帰分析をやってみて」

A：「できた(図 6)。補正 R2 は 0.774073 でなかなかいい結果だ。さっきの分析では、駅距離の t 値の絶対値が 2 を下回っていたけど、今回は 2 を上回っているし、p 値も 0.05 を下回っている。要因間の相関関係を確認しても強い相関関係も見受けられないし、各係数のプラスマイナスも矛盾がなさそうだ。マルチコが解消されたみたいだ。」

【図 6】

概要

回帰統計	
重相関 R	0.892886
重決定 R2	0.797245
補正 R2	0.774073
標準誤差	8.414822
観測数	40

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	4	9744.91	2436.228	34.40551	1.11E-11
残差	35	2478.323	70.80922		
合計	39	12223.23			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	94.68318	4.133055	22.90876	1.18E-22	86.29263	103.0737	86.29263	103.0737
駅	-1.12276	0.476232	-2.3576	0.024114	-2.08956	-0.15596	-2.08956	-0.15596
築年数	-1.30085	0.152752	-8.51608	4.75E-10	-1.61095	-0.99074	-1.61095	-0.99074
大手	12.68207	4.276768	2.965339	0.005415	3.999765	21.36437	3.999765	21.36437
階層	0.722174	0.443554	1.628155	0.112462	-0.17829	1.622637	-0.17829	1.622637

	㎡単価	駅	築年数	大手	階層
㎡単価	1				
駅	-0.27104	1			
築年数	-0.82368	0.163054	1		
大手	0.539902	0.039144	-0.34234	1	
階層	0.119645	0.146224	0.096671	0.313377	1

B：「マルチコは解消されたみたいだけど、もう少し重回帰分析の精度を上げていこう。重回帰分析の精度を上げる方法としては、**ステップワイズ法**(変数増減法)、**変数増加法**、**変数減少法**などの手順(選択の手順)で、**補正 R2**、**AIC**(赤池情報量規準)、**BIC**、**説明変数の p 値**、**Ru**(上田の説明変数選択基準)などの数値を改善(選択の基準)していく方法がある。採用すべき選択の手順や基準については、統計を扱っている書籍でもまちまちで正解はない。検証レベルの分析だと手間のかからない変数減少法を採用して、補正 R 2 や説明変数の p 値を改善していく方法がおすすめだ。p 値の大きい要因を 1 つずつ減らしながら、補正 R 2 や説明変数 p の値を改善していく方法だ。ただし、**p 値の大きさだけに着目して要因を減らすのはＮＧ**。p 値が多少大きくても、分析の中で重要だと考える要因は残すなど p 値だけに依存しないように注意しよう。スーパーの距離を外すことで駅距離の t 値と p 値は改善したけど、階層の p 値が基準値に満たないから、変数減少法の考え方を採用して、階層を説明変数から外し、3 要因で重回帰分析しようか。」

A：「できたよ(図 7)。」

【図 7】

概要

回帰統計	
重相関 R	0.884244
重決定 R2	0.781888
補正 R2	0.763712
標準誤差	8.605603
観測数	40

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	9557.203	3185.734	43.01767997	5.43155E-12
残差	36	2666.03	74.05639		
合計	39	12223.23			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	97.54821	3.824528	25.50594	1.2166E-24	89.79170728	105.3047	89.79171	105.3047
駅	-1.04312	0.484453	-2.15319	0.038077434	-2.025637274	-0.06061	-2.02564	-0.06061
築年数	-1.24917	0.152805	-8.1749	1.00866E-09	-1.559070414	-0.93926	-1.55907	-0.93926
大手	15.20258	4.077139	3.728738	0.000659912	6.933761569	23.4714	6.933762	23.4714

B：「残った駅距離、築年数、大手ブランドの要因のうち、p 値が一番大きい駅距離を外してさらに重回帰分析を

してもいいけど、マンション価格分析で駅距離は外したくないし、いずれの要因の p 値、t 値は基準値を満たしている。さっきの(図 6)の補正 R 2 に比べるとやや劣るけど補正 R2 も 0.763712 となかなかいい数字だから、この重回帰分析の結果を採用しようか。」

A：「係数に着目すると、回帰式が求められるんだったね。この回帰式にお目当てのマンションの基本情報を入力して計算すると、

㎡単価＝（－1.04312）　×駅徒歩距離(5 分)＋ （－1.24917）　×築年数(25 年)＋ （＋15.20258）×大手ブランド(1)＋ （＋97.54821）　≒76.3 万円/㎡ 総額：76.3 万円/㎡×65 ㎡≒4,960 万円

この重回帰分析から予測されるマンションの価格は、4,960 万円ということだね。それじゃ。4,800 万円で買っても悪くないってことかな。」

B：「今回集めた情報によるとそういえそうだね。でも、これは販売価格であって、成約価格ではない。やや高めの価格設定をしておいて様子見している売り手も多いはずだから、実際には一定のディスカウントを見込めるかもしれない。どのくらいのディスカウントが見込めるかは中古マンションの価格乖離率などの資料を参考にするといいよ。成約価格ベースのより精度の高い価格を知りたいなら、不動産の専門家等にセカンドオピニオンを依頼するのもいいかもしれないね。」

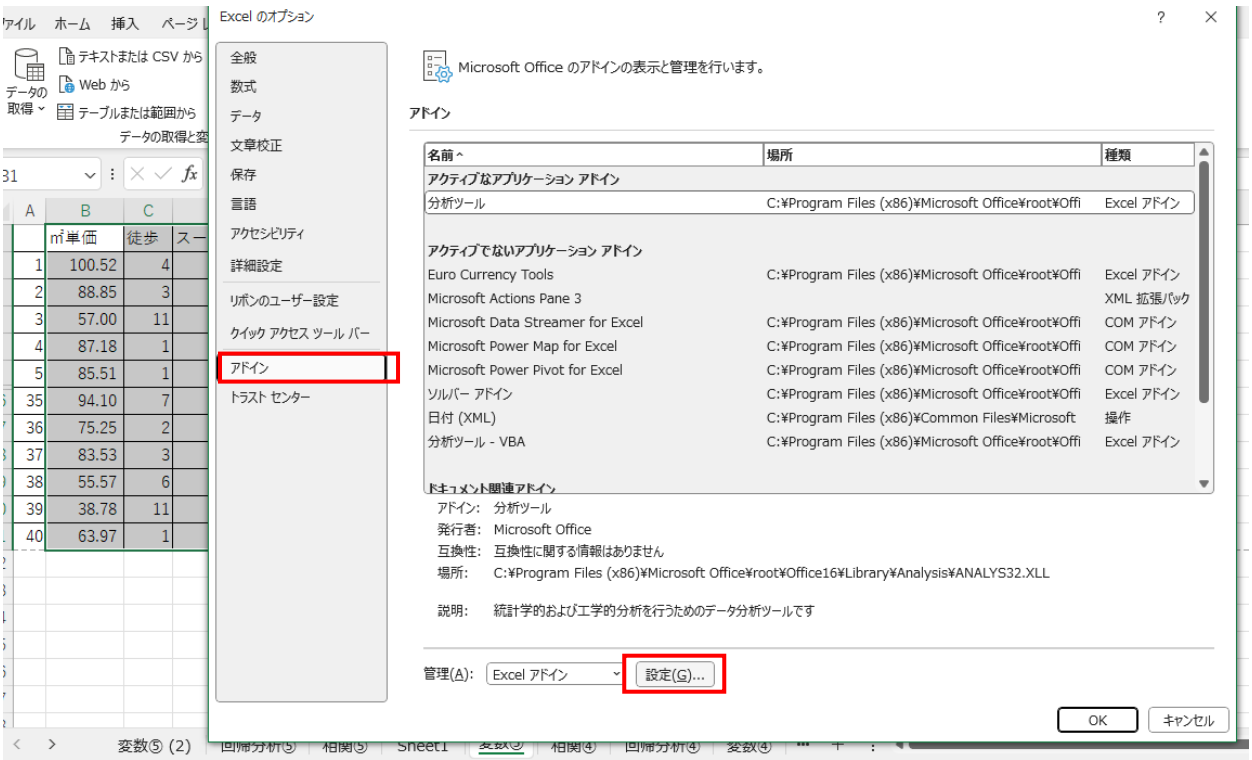
エクセルに搭載されている回帰分析機能は、当社の不動産鑑定評価の現場においても、試算結果の検証手段等で活用されており、使い勝手のいい機能となっています。今回の豆知識をきっかけに、日常生活やビジネスなど様々なシーンにおいてエクセルの回帰分析機能を活用いただければと思います。

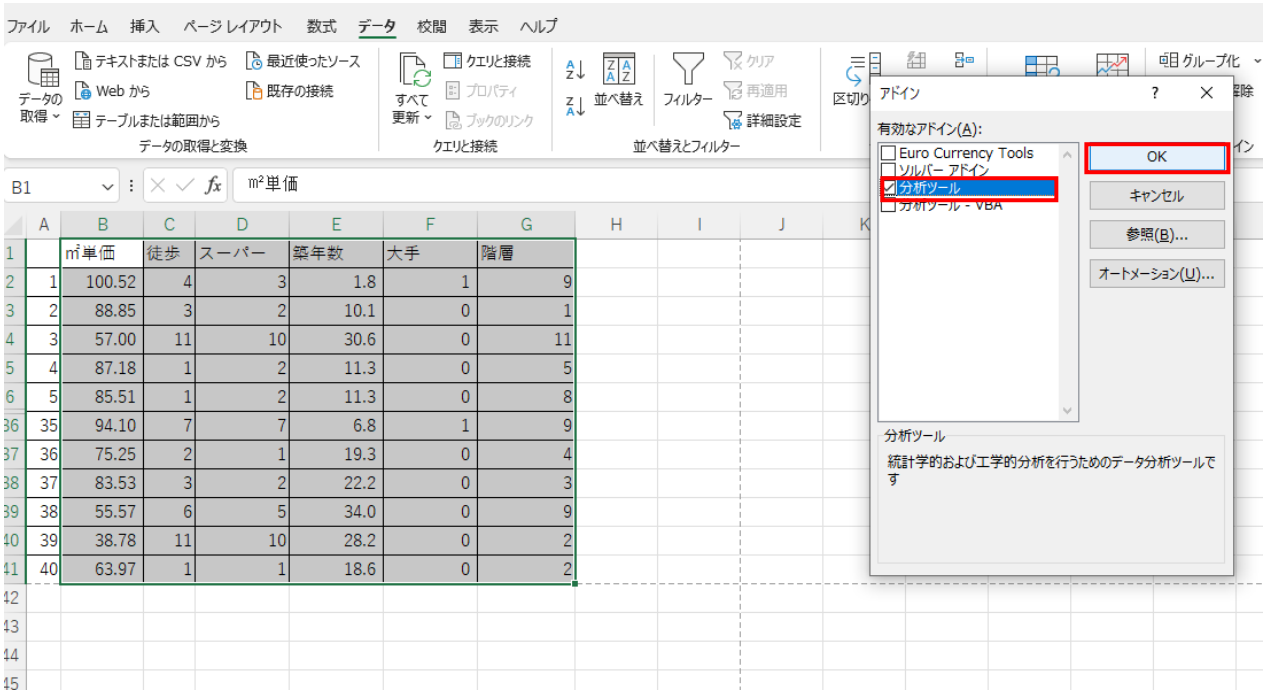
※エクセルのアドオン機能

データタブの中にデータ分析が表示されていない場合には、以下の操作を行ってください。

「ファイルタブ」→「その他」→「オプション」

「アドイン」→「設定」→「分析ツール」にチェック→「OK」





	㎡単価	徒歩	スーパー	築年数	大手	階層
1	100.52	4	3	1.8	1	9
2	88.85	3	2	10.1	0	1
3	57.00	11	10	30.6	0	11
4	87.18	1	2	11.3	0	5
5	85.51	1	2	11.3	0	8
35	94.10	7	7	6.8	1	9
36	75.25	2	1	19.3	0	4
37	83.53	3	2	22.2	0	3
38	55.57	6	5	34.0	0	9
39	38.78	11	10	28.2	0	2
40	63.97	1	1	18.6	0	2

【参考文献】

・西内啓「1 億人のための統計解析」（日経ＢＰ）
・日花弘子「Excel で学ぶデータ分析本格入門」（SB Creative）
・小川正樹「データ分析と統計」（ナツメ社）
・日本統計学会編「統計学Ⅲ・多変量データ解析法」（日本統計協会）