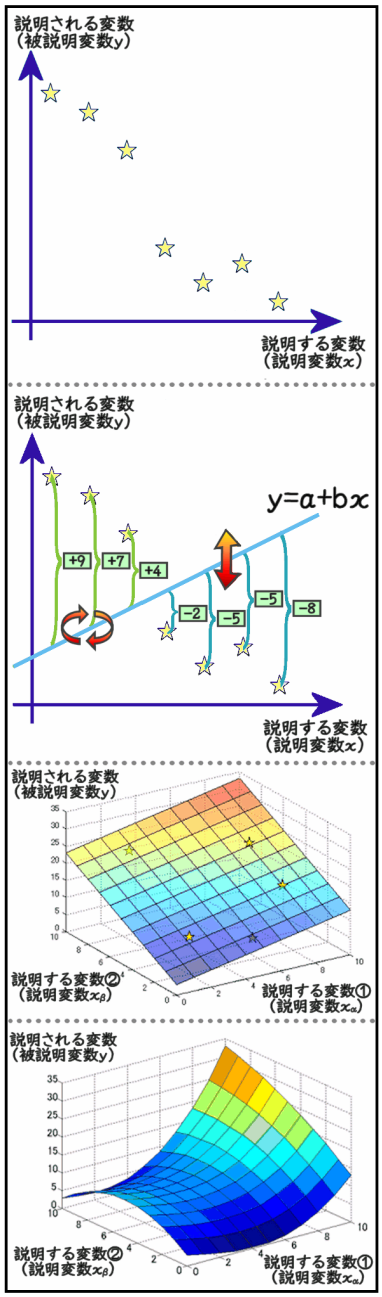
Excelにおける回帰分析（最小二乗法）の手順と出力

齋藤経史†

このレポートでは、Microsoft（マイクロソフト）の Excel（エクセル）における回帰分析の手順と出力の意味を説明します。予備知識がない方であっても、Excel の分析ツールを使って回帰分析の結果を出せるようになります。また、回帰分析の性質、統計量の意味、実証分析を行う上での注意点が書いてあります。

1　回帰分析（最小二乗法）の発想



右図 1 段目のような説明する変数（説明変数）x と説明される変数（被説明変数）y があるとします。説明変数 x ⇒被説明変数 y の効果の大きさを調べたり、予測をしたりするために関係を示す直線を引くことを考えます。$y = a + bx$ として、切片の高さ a や傾き b を動かして、当てはまりの良い直線を考えます。

単純に考えれば、星と線の差分の距離が小さくなるように線を引くと、当てはまりが良さそうです。しかし、差分の総和を最小化すると、2 段目の図のようにプラスとマイナスが相殺することがあります。また、星のずっと上の方に線を引けば、差分の総和は絶対値の大きなマイナスの値をとることができます。よって、差分の総和の最小化では当てはまりの良い線は引けません。当てはまりの良い線を引く一つの方法は『**差分の二乗**』の総和を最小化する最小二乗法です。回帰分析とも言われる最小二乗法は、**二乗することで全ての差分をプラスにしてから総和を最小化する**という発想です。

2 段目の図は、切片の高さの a の部分を除いて説明変数 x が 1 種類なので単回帰と呼ばれます。一方、3 段目の図は 2 種類の説明変数による回帰分析です。2 種類以上の説明変数による回帰分析を多重回帰と言います。単回帰では最小化の対象は星と線の差分の二乗和ですが、多重回帰ではあてはめる線が板になります。それでも二乗和を最小化するという発想自体は同じです。四次元以上になると想像できませんが、イメージは二次元が三次元になる時と同じです。

回帰分析（最小二乗法）によって引かれた線を回帰線と言います。単に最小二乗法という場合、回帰線は直線となり、線形と呼ばれます。線形の場合は x が『$0 \to 1$』に上昇する時と『$30 \to 31$』に上昇する時とで y に与える効果は、同じ b で等しくなります。ただ、線形の中でもあらかじめ log をとって対数変換をしたり、二乗項 x_α^2 や交差項 $x_\alpha \cdot x_\beta$ を説明変数に入れたりすることで、説明変数を元の値に戻すと 4 段目の図のような曲がった回帰線を描くことができます。

---

version 2.01（2008 年 3 月 30 日改訂）

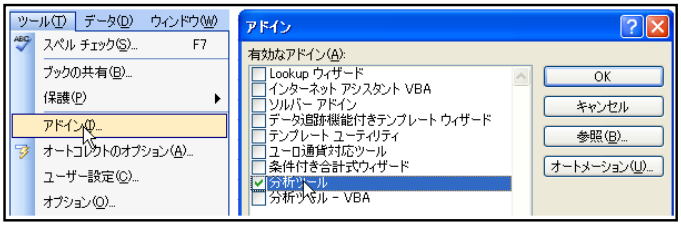
† 東京大学 経済学研究科 博士課程（Website:http://keijisaito.info, E-mail:master@keijisaito.info）

2　Excel における回帰分析の手順

この章では、Excel の分析ツールのアドインと回帰分析の実行方法を紹介します。

2.1　アドインから分析ツールを追加する

Excel での回帰分析は分析ツールを使います。これまで分析ツールを使ったことのない場合は、分析ツールをアドインする必要があります。Excel 2003 以前の場合はメニューにある [**ツール**] ⇒ [**アドイン**] をクリックします。[1] 表示された候補の中の [**分析ツール**] にチェックを入れて [**OK**] をクリックします。Microsoft Office や Excel の CD が必要になるかもしれません。



[Excel のアドイン]

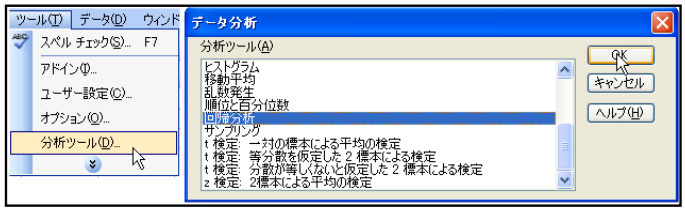
2.2　回帰分析用データを縦に並べる

Excel で回帰分析を行うためには、被説明変数 y と説明変数 x のデータを縦に並べる必要があります。ここでは例示として、**Excel での回帰分析（最小二乗法）用乱数データ**(http://keijisaito.info/arc/excel/ols_data.xls) を用います。

この**乱数データ**では、被説明変数 y に対して x_α と x_β の 2 つの説明変数で回帰分析を行います。1 枚目の [**乱数シート**] では、y と x の関係に対応する係数、その関係の間に入るノイズに対応するエラーの値を設定することができます。再計算の F9 を押すことで、乱数からエラーが再生成されます。[**乱数シート**] の中には、分析ツールの回帰分析に似せた出力があり、推定値の挙動を見ることができます。また、変化する [**乱数シート**] を値で貼り付けて固定したものが [**固定シート**] です。以降のページでは、回帰分析結果の例に [**固定シート**] を用いて説明します。

2.3　分析ツールから回帰分析を実行する

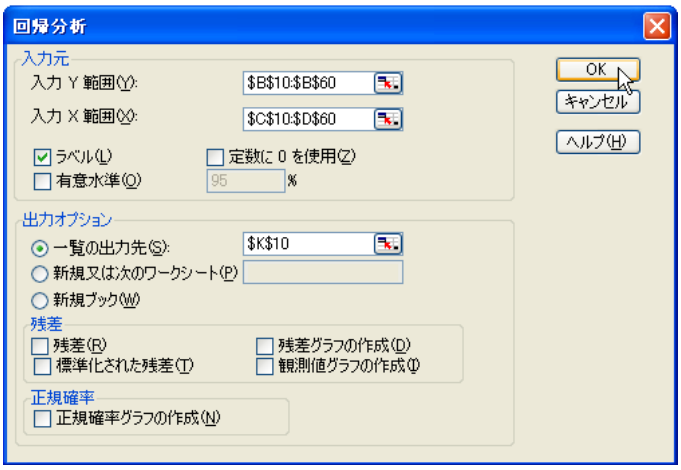
縦に並べたデータを使って回帰分析を実行します。[**ツール**] ⇒ [**分析ツール**] をクリックし、[**回帰分析**] を選択して、[**OK**] をクリックします。[2]



[分析ツールから回帰分析を選択]

[1] Excel 2007 の場合は画面左上の Microsoft Office のボタンをクリックして、表示されたメニューの中から [**Excel のオプション**] をクリックします。Excel のオプションの中から [**アドイン**] を選択し、[**設定**] をクリックします。

[2] Excel 2007 の場合は [**データ**] のタブの中にある [**データ分析**] をクリックします。



[Excel の回帰分析のダイアログボックス]

表示された回帰分析のダイアログボックスには、入力 Y 範囲に被説明変数の 1 列を、入力 X 範囲に説明変数の列を指定してください。[3] のボタンを押せば、マウスで範囲を指定することができます。

例示の**乱数データ**では、一行目に変数名を入れて、入力 Y 範囲を〈$B$10:$B$60〉、入力 X 範囲を〈$C$10:$D$60〉、出力先を〈$K$10〉に指定しています。[4] [5] [6] 回帰分析の指定が終われば、[**OK**] をクリックします。指定した出力先に下の画像のような回帰分析の結果が表示されます。

概要								
回帰統計								
重相関 R	0.951759							
重決定 R2	0.905845							
補正 R2	0.901838							
標準誤差	8.612262							
観測数	50							
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	2	33538.42	16769.21	226.0882953	7.68E−25			
残差	47	3486.04	74.17106					
合計	49	37024.46						
	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	21.96217	9.353082	2.348121	0.023127974	3.146194	40.77814	3.146194	40.77814
x_α	3.202912	0.230515	13.89456	2.92143E−18	2.739174	3.666649	2.739174	3.666649
x_β	4.609189	0.505876	9.111303	5.9088E−12	3.591498	5.62688	3.591498	5.62688

[Excel の回帰分析の出力]

[3] Excel の回帰分析において、説明変数に指定できるのは最大 16 種類です。

[4] 推定結果に変数名を表示する場合は [**ラベル**] にチェックを入れて、変数名を入力範囲の第一行目に指定してください。

[5] 入力範囲の**$**は Excel で絶対参照を示す記号です。マウスでセルを指定した場合は、自動的に**$**が付きます。打ち込む場合は、**$**不要で『左上セル:右下セル』と記入できます。

[6] 出力オプションに〈$K$10〉のようにセルを指定すれば、同じシートの指定したセルから右下に回帰分析結果が出力されます。新規ワークシートにすれば他のシートに出力します。

3　回帰分析（最小二乗法）の性質

この章では、Excel での表示と対応させながら回帰分析（最小二乗法）の性質を紹介します。推定や統計量の理解にも役立ちます。

[記号と表記]

ここでは、このレポートで用いる記号と表記を説明します。まず、回帰分析の被説明変数は y で表します。合計 n 個の標本を用いて推定するとし、i 個目の標本の被説明変数は y_i と下添え字をつけて表記します。また、回帰分析の説明変数は x で表します。ところで $y = a + bx$ という単回帰においても $y = a \cdot 1 + bx$ として、切片にあたる a には常に 1 の説明変数がついていると見なせます。全ての標本に対して 1 をとる切片用の説明変数を 1 種類目に数え、データで与える説明変数は 2 種類目から数えます。標本 i の j 種類目の説明変数は x_{ij} と下添え字を並べて表記します。

回帰分析（最小二乗法）から得られた推定係数は b で表します。切片の推定係数は b_1 とし、j 種類目の説明変数の推定係数を b_j と表記します。また、推定係数と説明変数があれば、当てはめ値（y の予測値）を算出することができます。当てはめ値を $\widehat{y}$（y ハット）と表記すると、合計 k 種類の説明変数による標本 i に対する $\widehat{y}_i$ は、以下の (1) 式のように表すことができます。

$$\widehat{y}_i = b_1 + b_2 \cdot x_{i2} + b_3 \cdot x_{i3} + \cdots + b_k \cdot x_{ik} \tag{1}$$

なお (1) 式で表される $\widehat{y}_i$ が被説明変数 y_i に一致するケースはほとんどなく、両者の間には推定エラー e_i が存在します。[7] 推定エラー e_i は、標本 i に関して回帰分析で説明できない部分に相当します。逆に言えば、結果的に算出された推定エラー e_i を用いて、当てはめ値 $\widehat{y}_i$ を調整すると、(2) 式のように被説明変数 y_i になります。

$$y_i = \widehat{y}_i + e_i \tag{2}$$

3.1　説明変数と推定エラーの積の総和は 0 になる

回帰分析（最小二乗法）では、j 種類目の説明変数 x_j と推定エラー e の積を n 個の標本で合計すると

$$\sum_{i=1}^{n} x_{ij} \cdot e_i = 0 \tag{3}$$

と必ず 0 になります。[8] 0 になっていることを確認したい場合は、**乱数データ**の [**乱数シート**] で V〜Y 列の 62 行目にある平均値が F9 を押しても、0 から動かないことで確かめてください。平均値が 0 なので、標本数 n をかけた合計値も 0 になることが分かります。

---

[7] 回帰分析（最小二乗法）よって得られた推定エラーは e（イー）で表すのに対し、説明変数と被説明変数の間にある実際のノイズは ϵ（イプシロン）で表します。最小二乗法の問題点をふまえて高度な分析を行う場合には、e と ϵ をきちんと区別することが重要です。

[8] 推定エラーの二乗和 $\sum e_i^2 = \sum (y_i - \widehat{y}_i)^2 = \sum \{y_i - (b_1 + b_2 \cdot x_{i2} + \cdots + b_k \cdot x_{ik})\}^2$ の最小化の必要条件は一階微分 $\frac{\partial \sum e_i^2}{\partial b_j} = 0$ です。$\frac{\partial \sum e_i^2}{\partial b_j} = -2 \sum \{x_{ij}(y_i - b_1 + b_2 \cdot x_{i2} + \cdots + b_k \cdot x_{ik})\} = -2 \sum \{x_{ij} \cdot e_i\}$ なので、一階微分=0 は $\sum \{x_{ij} \cdot e_i\} = 0$ に書きかえることができます。

前述したように、(1) 式の切片である b_1 には常に 1 の説明変数が付いていると考えることができます。すると (3) 式の性質は、

$$\sum_{i=1}^{n} 1 \cdot e_i = \sum_{i=1}^{n} e_i = 0 \tag{4}$$

と推定エラー e の総和が 0 と書きかえることができます。e の平均を $\bar{e}$（e バー）とすると、切片を含めた回帰分析では、推定エラーの総和 $\sum e_i$ や平均 $\bar{e}$ は必ず 0 になります。

また、(1) 式の両辺に e_i をかけ、総和をとると

$$\begin{aligned}\sum_{i=1}^{n} \widehat{y_i} \cdot e_i &= \sum_{i=1}^{n} [b_1 + b_2 \cdot x_{i2} + \cdots + b_k \cdot x_{ik}] e_i = \sum_{i=1}^{n} [b_1 \cdot e_i + b_2 \cdot x_{i2} \cdot e_i + \cdots + b_k \cdot x_{ik} \cdot e_i] \\ &= b_1 \sum_{i=1}^{n} e_i + b_2 \sum_{i=1}^{n} x_{i2} \cdot e_i + \cdots + b_k \sum_{i=1}^{n} x_{ik} \cdot e_i = b_1 \cdot 0 + b_2 \cdot 0 + \cdots + b_k \cdot 0 = 0 \end{aligned} \tag{5}$$

と当てはめ値 $\hat{y}$ と推定エラー e の積の総和も 0 になることが分かります。総和が 0 なので、標本数 n で割った平均値も 0 になります。

	U	V	W	X	Y
62	平均値	0.0000	0.0000	0.0000	0.0000
63	分散の不偏推定値	71.14	14782.66	27975.10	1746726.69
64					
65		推定エラーe	x_αとeの積	x_βとeの積	yhatとeの積

[説明変数 x と推定エラー e の積の平均値が 0]

3.2 回帰線は説明変数と被説明変数の標本平均を通る

(2) 式の $y_i = \widehat{y_i} + e_i$ を標本 n で総和をとります。

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y_i} + \sum_{i=1}^{n} e_i \tag{6}$$

また、(4) 式で示されるように切片のある回帰分析では (6) 式の右辺第二項の $\sum e_i$ は 0 になります。よって (6) 式は

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y_i} \tag{7}$$

と、被説明変数 y と当てはめ値 $\hat{y}$ の総和同士が等しい条件に書きかえることができます。総和が等しいので標本数 n で割った平均も等しくなります。よって y の標本平均を $\overline{y}$（y バー）で表すと $\overline{y} = \overline{\widehat{y}}$ です。また、当てはめ値の平均 $\overline{\widehat{y}}$ を説明変数を用いて書くと

$$\overline{y} = \overline{\widehat{y}} = b_1 + b_2 \cdot \overline{x_2} + b_3 \cdot \overline{x_3} + \cdots + b_k \cdot \overline{x_k} \tag{8}$$

となります。(8) 式は [被説明変数 y の平均] = [当てはめ値 $\hat{y}$ の平均] = [回帰式に説明変数 x の平均を代入した値] であることを示しています。つまり、切片のある回帰分析において回帰線は必ず被説明変数 y、説明変数 x の標本平均を通ります。このため、**乱数データ**の [**乱数シート**] では、y の平均値〈B62〉と x の平均値での当てはめ値〈B67〉は常に一致します。

62	平均値	150.7405	11.6796	19.8234
63	分散の不偏推定値	755.60	34.04	7.07
64				
65		y	x_α	x_β
66	説明変数の平均値から当てはめ値を作成			
67	ybar=	150.7405	=21.96+3.2$x_\alpha$bar+4.61$x_\beta$bar	

[説明変数 x の平均値に対する当てはめ値は被説明変数 y の平均]

ところで、切片を含めた回帰分析の特殊形の『切片のみでの回帰分析』を考えます。すると (8) 式は

$$\overline{y} = \overline{\widehat{y}} = b_1 \tag{9}$$

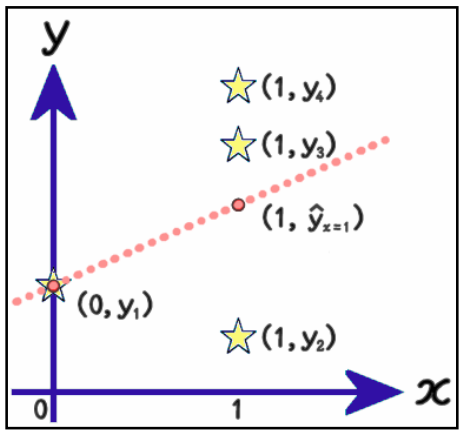
となります。つまり、切片のみの回帰分析において切片の高さ b_1 は被説明変数の標本平均 $\bar{y}$ になります。また、回帰分析（最小二乗法）の発想に戻ると (9) 式は

$$b_1\text{を動かして}\sum_{i=1}^{n}(y_i - b_1)^2 \text{ を最小化} \quad \Rightarrow \quad b_1 = \overline{y} \tag{10}$$

となります。[9] (10) 式から切片のみの回帰分析において $\sum e_i$ を最小化する切片 b_1 は、被説明変数の標本平均 $\bar{y}$ であることが分かります。

3.3 推定係数は平均的な効果になる

右図のような $(x, y) = (0, y_1)\ (1, y_2)\ (1, y_3)\ (1, y_4)$ の 4 つの標本に対して単回帰を行う場合を考えます。回帰分析から得られる $x = 0$ の当てはめ値を $\widehat{y}_{x=0}$、$x = 1$ の当てはめ値を $\widehat{y}_{x=1}$ で表します。



この例では $x = 0$ の標本は $(x, y) = (0, y_1)$ の 1 つしかありません。また、$\widehat{y}_{x=1}$ の値に依存せず、必ず $(0, y_1)$ と $(1, \widehat{y}_{x=1})$ を結ぶ直線を書くことができます。このため、$\widehat{y}_{x=0}$ が y_1 になることは明らかです。一方、$\widehat{y}_{x=1}$ は (10) 式の b_1 を $\widehat{y}_{x=1}$ として、同種の問題を解くことになります。すると $\widehat{y}_{x=1}$ は $x = 1$ の標本における y の平均値から、$\widehat{y}_{x=1} = \frac{y_2+y_3+y_4}{3}$ となることが分かります。

よって回帰線は $(0, y_1)$ と $(1, \frac{y_2+y_3+y_4}{3})$ を結ぶ直線として

$$\widehat{y} = y_1 + (\frac{y_2 + y_3 + y_4}{3} - y_1)x \tag{11}$$

になります。(11) 式では、x が 1 増えた場合の y への効果は $(\frac{y_2+y_3+y_4}{3} - y_1)$ です。この推定係数は $(y_2 - y_1), (y_3 - y_1), (y_4 - y_1)$ の平均値になっています。(11) 式から類推されるように、回帰分析（最小二乗法）から得られる推定係数は、平均的な効果を算出しています。

---

[9] $\frac{\partial \sum (y_i - b_1)^2}{\partial b_1} = -2(\sum y_i - nb_1)$ から、最小化の必要条件の一階微分=0 は $b_1 = \overline{y}$ と書きかえることができます。

4　回帰分析全体に関する Excel の出力

この章では、回帰分析全体に関する Excel の出力について説明します。[10]

4.1　決定係数〈重決定 R2〉：回帰分析の当てはまりの指標

回帰分析から得られた当てはめ値 $\widehat{y}$ が、どれだけ被説明変数 y を説明できているかを考えます。前章の (2) 式の $y_i = \widehat{y}_i + e_i$ の右辺第一項は『回帰分析によって説明できる部分』、右辺第二項は『説明できない推定エラーの部分』に相当します。第一項で説明できる部分が大きい方が、回帰分析の当てはまりが良いという印象があります。しかし、(2) 式では平均的な y の水準が高ければ、$\widehat{y}$ が高くなります。そこで、(2) 式の両辺から平均値を引いた偏差を用いて当てはまりの指標を作ります。

また、偏差を用いても単純に総和をとって当てはまりの指標を作ることはできません。(4) 式にあるように e の総和 $\sum e_i$ と平均値 $\bar{e}$ は 0 になるからです。そこで、最小二乗法の発想のように二乗してから総和をとることを考えます。切片のある回帰分析では (7) 式から $\bar{y} = \overline{\widehat{y}}$、(4) 式から $\bar{e} = 0$ なので、(2) 式の両辺から平均値 $\bar{y}$ を引いた $y_i - \bar{y} = \widehat{y}_i - \bar{y} + e_i$ に対して二乗和をとります。

$$\sum_{i=1}^{n}[y_i - \bar{y}]^2 = \sum_{i=1}^{n}[\widehat{y}_i - \bar{y} + e_i]^2 = \sum_{i=1}^{n}[\widehat{y}_i - \bar{y}]^2 + 2\sum_{i=1}^{n}[\widehat{y}_i - \bar{y}]e_i + \sum_{i=1}^{n}e_i^2 \tag{12}$$

(12) 式の第二項の $2\sum[\widehat{y}_i - \bar{y}]e_i$ は (4) 式と (5) 式より 0 になります。よって、(12) 式は

$$\sum_{i=1}^{n}[y_i - \bar{y}]^2 = \sum_{i=1}^{n}[\widehat{y}_i - \bar{y}]^2 + \sum_{i=1}^{n}e_i^2 \tag{13}$$

と書けます。(13) 式の形は [y の偏差平方和]=[$\widehat{y}$ の偏差平方和]+[e の二乗和] です。[11] つまり、『被説明変数 y の偏差平方和』は『説明できる分の $\widehat{y}$ の偏差平方和』と『説明できない分の e の二乗和』に分解することができます。[12] (13) 式の右辺第一項が第二項に比べて大きければ、回帰分析の当てはまりが良いことになります。割合の指標にするために (13) 式の両辺を左辺で割ります。

$$1 = \frac{\sum[y_i - \bar{y}]^2}{\sum[y_i - \bar{y}]^2} = \frac{\sum[\widehat{y}_i - \bar{y}]^2}{\sum[y_i - \bar{y}]^2} + \frac{\sum e_i^2}{\sum[y_i - \overline{y_i}]^2} \tag{14}$$

(14) 式の黄色に塗った部分が被説明変数 y の偏差平方和に占める $\widehat{y}$ の偏差平方和で説明できる割合に相当し、決定係数と呼ばれます。[13] (14) 式の右辺の二つの項は、分子分母ともに二乗の総和の形で必ずプラスです。その和が 1 になるので、決定係数は 0 から 1 の間の値になります。[14] **決定係数〈重決定 R2〉は回帰分析の当てはまりの指標であり、y の偏差平方和のうち $\widehat{y}$ の偏差平方和によって説明できる割合**を表します。

---

[10] 実証分析のレポートにおいて〈重決定 R2〉〈重相関 R〉〈補正 R2〉という用語は、ほとんど使われません。

[11] $\bar{e} = 0$ より、$\sum e_i^2$ は e の偏差平方和とも言えます。また、推定エラーが残差と呼ばれることに対応して、$\sum e_i^2$ は残差平方和とも言われます。

[12] 分散分析表の変動と書かれている列には、上から『$\widehat{y}$ の偏差平方和』『e の二乗和』『y の偏差平方和』が表示されます。

[13] 回帰分析の結果を示す際、決定係数は R^2 と表記されることがあります。

[14] 切片のみによる回帰分析では、(9) 式より $\sum[\widehat{y}_i - \bar{y}]^2 = \sum[\bar{y} - \bar{y}]^2 = 0$ となるので、決定係数は 0 になります。

$$\text{決定係数〈重決定 R2〉} = \frac{\sum[\widehat{y}_i - \overline{y}]^2}{\sum[y_i - \overline{y}]^2} = 1 - \frac{\sum e_i^2}{\sum[y_i - \overline{y}]^2}$$
$$= \frac{\text{当てはめ値の偏差平方和}}{\text{被説明変数の偏差平方和}} = 1 - \frac{\text{推定エラーの二乗和}}{\text{被説明変数の偏差平方和}} \tag{15}$$

4.2 〈重相関 R〉：決定係数の正の平方根

決定係数〈重決定 R2〉は計算過程で二乗をとっているので、尺度を元に戻すために**決定係数の正の平方根をとった値が〈重相関 R〉**です。[15] 決定係数は 0 から 1 の範囲にあるので、正の平方根をとった〈重相関 R〉は必ず決定係数以上、1 以下の値になります。

$$\text{〈重相関 R〉} = \sqrt{1 - \frac{\sum e_i^2}{\sum[y_i - \overline{y}]^2}} = \sqrt{1 - \frac{\text{推定エラーの二乗和}}{\text{被説明変数の偏差平方和}}} = \sqrt{\text{決定係数}} \tag{16}$$

4.3 自由度修正決定係数〈補正 R2〉：説明変数の数を考慮した当てはまりの指標

いったん行った回帰分析に説明変数を追加して、再び回帰分析をする場合を考えます。もし、追加した説明変数が全く回帰分析の役に立たない場合は、回帰分析の結果として追加した説明変数の推定係数は 0 になります。この場合、推定エラーの二乗和 $\sum e_i^2$ も追加前の値と変わりません。一方で、追加した説明変数が少しでも回帰分析の役に立つ場合は、0 以外の推定係数が得られます。この場合、$\sum e_i^2$ は減少します。 実際には、無関係な説明変数であっても推定係数がちょうど 0 となることはありません。説明変数の追加によって、多少なりとも $\sum e_i^2$ は減少します。

(15) 式で示したように、決定係数は説明変数の数に関係なく、$\sum e_i^2$ の $\sum[y_i - \overline{y}]^2$ に占める割合から算出されます。このため、説明変数の追加は決定係数にプラスの影響しかありません。よって、無関係な説明変数であっても追加すればするほど、決定係数は増加します。[16]

そこで、説明変数を増やすことにマイナスの影響もある当てはまりの指標として、自由度修正決定係数を示すことがあります。[17] 切片を含めて説明変数が k 種類あるとすると、自由度修正決定係数は以下の (17) 式で定義されます。

$$\text{自由度修正決定係数〈補正 R2〉} = 1 - \frac{\sum e_i^2}{\sum[y_i - \overline{y_i}]^2} \cdot \frac{(n-1)}{(n-k)}$$
$$= 1 - \text{(1 − 決定係数)} \cdot \frac{\text{(標本数 − 1)}}{\text{(標本数 − 説明変数の数)}} \tag{17}$$

自由度修正決定係数は、決定係数よりも小さく 1 以下の値となり、マイナスもあり得ます。[18] 説明変数を追加すれば決定係数を高め、(17) 式の黄色に塗った値が減少します。一方で水色に塗った値は分母が減少することで増加します。説明変数を追加した場合、自由度修正決定係数の変化の方向

[15] Excel では ^ が累乗を表します。正の平方根は 0.5 乗なので、Excel のセルに『=数値^0.5』と入力して計算できます。また、Excel の関数の『=sqrt(数値)』を使うこともできます。

[16] 標本数と切片を含めた説明変数の数が等しければ、回帰分析が実行できる限り決定係数は必ず 1 になります。

[17] 自由度修正決定係数は、Adjusted-R^2 や Adj-R^2 と表記されることがあります。

[18] 標本数 n が説明変数 k の数に比べてはるかに多い場合、(17) 式の水色の部分が 1 に近くなります。この場合、自由度修正決定係数と決定係数の差は極めて小さくなります。

は『黄色の減少分』と『水色の増加分』の逆方向の作用のどちらが大きいかの綱引きによって決まります。**自由度修正決定係数〈補正 R2〉は、説明変数の数を考慮した当てはまりの指標**です。

4.4 〔エラーの〕標準誤差：エラーの平均的なばらつきの推定値

(4) 式に示されるように、切片のある回帰分析では推定エラーの平均 $\bar{e} = 0$ です。しかし同じ平均 0 でも ±1 と ±2 のそれぞれ 2 つの標本では、±2 の方がばらつきが大きいと言えます。このばらつきの指標を計算するために、まず $\sum e_i^2$ と推定エラー e を二乗をしてから総和をとります。その後で標本 1 つあたりの指標に変換します。

単純に考えれば、$\sum e_i^2$ を標本数 n で割れば、標本 1 つあたりの指標になりそうです。しかし、$\sum e_i^2$ を標本数 n で割ると、真のばらつきに比べて小さめの推定値になる傾向があります。[19] 例えば、極端なケースとして標本が 2 つなら切片を含めた単回帰で完全に説明できて当たり前です。[20] この場合、$\sum e_i^2$ は 0 になります。しかし、エラーが 0 でばらつかないのではなく、切片を含めた説明変数の数と標本数が同じなので $\sum e_i^2 = 0$ となっていると考えられます。標本数を n、切片を含めた説明変数の数を k とすると、説明できて当たり前でないのは $(n-k)$ の値に依存し、この値を自由度と呼びます。[21]

エラーの二乗和 $\sum e_i^2$ を自由度 $(n-k)$ で割ると、不偏分散と呼ばれる偏りのない標本 1 つあたりのエラーのばらつきが導出できます。[22] また、不偏分散は計算過程で二乗をとっているので、正の平方根をとることで元の尺度の標準誤差という指標にします。[23] [24] **〔回帰統計の〕標準誤差は、エラーの平均的なばらつきの推定値**です。

$$\text{エラーの不偏分散} = \frac{\sum e_i^2}{(n-k)} = \frac{\text{推定エラーの二乗和}}{(\text{標本数} - \text{説明変数の数})} \tag{18}$$

$$\text{エラーの標準誤差} = \sqrt{\frac{\sum e_i^2}{(n-k)}} = \sqrt{\frac{\text{推定エラーの二乗和}}{(\text{標本数} - \text{説明変数の数})}} \tag{19}$$

4.5 〈分散分析表〉と〈有意 F〉：切片以外の説明変数は全て無効の検定と確率の上限

統計学や計量経済学では『異なっていない（同質）』や『効果がない（無効）』を主張する仮説を帰無仮説と言います。一方、帰無仮説の反対側の『異なっている（異質）』や『効果がある（有効）』を主張する仮説を対立仮説と言います。直接、対立仮説を肯定することが難しい場合、帰無仮説を否定することで間接的に対立仮説を肯定するという手続きをとります。

〈分散分析表〉は『切片以外の全ての説明変数は無効 ⇒ 切片以外の説明変数の真の係数は全て 0

---

[19] 回帰分析のみならず、一般的な不偏分散の推定も偏差平方和を標本数 n ではなく、$(n-1)$ で割ります。**分散の不偏推定量**(http://keijisaito.info/arc/excel/variance.xls) で F9 を押せば、標本数で割ることのズレや不偏推定を視覚的に確認できます。

[20] 2 つの点が散布図のどこにあったとしても、その 2 点を通る直線は必ず書けます。

[21] (17) 式は、水色の部分の分母が自由度になっていることから、自由度修正決定係数という名称になっています。

[22] 分散分析表の [残差の行] と [分散の列] が交差するセルには、エラーの不偏分散が表示されます。

[23] 分布から直接計算される分散の平方根を標準**偏差**(standard deviation) と言い、データから作られた統計量（例えば標本平均、残差平方和、推定値）の標準偏差を標準**誤差**(standard error) と言います。

[24] エラーの不偏分散は s^2、エラーの標準誤差は s で表されることがあります。

である』という帰無仮説の検定を行っています。[25] この帰無仮説が正しい場合、切片だけで回帰分析をしても、説明変数を入れて回帰分析をしても、推定エラーの二乗和 $\sum e_i^2$ に大きな差がないと考えるのが自然です。切片だけで回帰分析をした場合、(10) 式から $b_1 = \overline{y}$ となり、推定エラーの二乗和は $\sum[y_i - \overline{y}]^2$ と被説明変数 y の偏差平方和そのものになります。よって、切片以外の説明変数を追加することで減少した推定エラーの二乗和は $\sum[y_i - \overline{y}]^2 - \sum e_i^2$ となり、(13) 式から $\sum[\widehat{y_i} - \overline{y}]^2$ であることが分かります。

〈分散分析表〉の [分散の列] には、左隣のセルの $\sum[\widehat{y_i} - \overline{y}]^2$ を $(k-1)$ で割ることで、『説明変数あたりの推定エラーの二乗和の減少』を表しています。その 1 つ下のセルには (18) 式で算出される『説明変数を入れても残ったエラーのばらつき』の不偏分散が表示されます。[26]〈観測された分散比〉は以下の (20) 式の形で、この 2 つの数の比率をとっています。

$$\text{〈観測された分散比〉} = \frac{\frac{\sum(\widehat{y_i}-\overline{y})^2}{(k-1)}}{\frac{\sum e_i^2}{(n-k)}} = \frac{\text{説明変数あたりの推定エラーの二乗和の減少}}{\text{不偏分散}} \tag{20}$$

また〈観測された分散比〉は、以下のように変形して決定係数で表すことができます。

$$\begin{aligned}\text{〈観測された分散比〉} &= \frac{\frac{\sum(\widehat{y_i}-\overline{y})^2}{(k-1)}}{\frac{\sum e_i^2}{(n-k)}} = \frac{\sum(\widehat{y_i}-\overline{y})^2}{\sum(y_i-\overline{y})^2} \cdot \frac{\sum(y_i-\overline{y})^2}{\sum e_i^2} \cdot \frac{(n-k)}{(k-1)} \\ &= \frac{\text{決定係数}}{(1-\text{決定係数})} \cdot \frac{(\text{標本数} - \text{説明変数の数})}{(\text{説明変数の数} - 1)}\end{aligned} \tag{21}$$

当てはまりの指標である決定係数が高ければ、(21) 式の黄色に塗られている分数の分子が大きく、分母が小さいことで〈観測された分散比〉も高くなることが分かります。

(20) 式から『説明変数あたりの推定エラーの二乗和の減少』が大きいほど、(21) 式から『決定係数』が大きいほど〈観測された分散比〉が大きくなることが分かります。このため、〈観測された分散比〉が大きい場合は『切片以外の説明変数の真の係数は全て 0 である』という帰無仮説は不自然になります。この不自然という感覚を統計で表したのが、〈分散分析表〉の〈有意 F〉です。[27] **〈有意 F〉は『切片以外の全ての説明変数の効果が 0 である』という帰無仮説のもとで、偶然によって標本が観測されてしまう確率の上限**を示しています。

乱数データの [**固定シート**]〈P21〉表示されている〈有意 F〉の [7.68E-25] は $7.68 \times (0.1)^{25}$ を表し、0 が 25 個ならぶほど小さい数です。[28] この場合、「『全ての変数が無効』という帰無仮説が正しければ、$7.68 \times (0.1)^{25}$ 以下の確率でしか起こらないことが起こった。」⇒ 「帰無仮説は不自然で、ほぼ確実に効果のある説明変数がある。」という解釈になります。

---

[25] Excel には〈分散分析表〉と表示されますが、回帰分析の F 検定を分散分析と呼ぶことは稀です。統計学で一般に分散分析と呼ばれる ANOVA(analysis of variance) は、他の分析ツールとして存在します。これらは同じ F 分布を使った検定ですが、回帰分析では『F 検定』や『複数制約』と呼ぶのが一般的です。また、〈有意 F〉という用語を使うことはなく『複数制約の P 値』が一般的です。

[26] **乱数データ**の [**固定シート**] の〈N21〉は説明変数あたり推定エラーの二乗和を [16769.21] 減らしたこと、〈N22〉は切片以外の説明変数を入れても、不偏分散として [74.17106] 残ったことを示しています。

[27] 〈有意 F〉の F は、F 分布を指しています。帰無仮説が正しければ〈観測された分散比〉は自由度 $(k-1, n-k)$ の F 分布に従うことが知られています。

[28] Excel 関数から直接 F 分布による統計量を計算することもできます。[=FDIST(観測された分散比, 回帰自由度 $(k-1)$, 残差自由度 $(n-k)$]=[FDIST(226.0883,2,47)] と入力すると、『7.68E-25』が出力されます。

5　説明変数に関する Excel の出力

この章では、Excel 回帰分析における説明変数の出力について説明します。**乱数データ**の [**固定シート**] では、以下のような出力が表示されています。

	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	21.96217	9.353082	2.348121	0.023127974	3.146194	40.77814	3.146194	40.77814
x_α	3.202912	0.230515	13.89456	2.92143E-18	2.739174	3.666649	2.739174	3.666649
x_β	4.609189	0.505876	9.111303	5.9088E-12	3.591498	5.62688	3.591498	5.62688

[説明変数に関する Excel の出力]

5.1　〔推定〕係数：説明変数 1 単位の増加⇒被説明変数への効果

〔推定〕係数は、説明変数の 1 単位の増加⇒被説明変数への効果の推定値を示しています。[29] 回帰分析の推定係数は、[3.3 推定係数は平均的な効果になる] に示したように平均的な効果に対応しています。また、切片は他の説明変数が全て 0 の時の被説明変数の当てはめ値を示しています。

5.2　〔推定係数の〕標準誤差：推定係数の不確かさ

推定係数は、あくまで推定しているわけですから、真の係数からの乖離が予想されます。この乖離の指標が推定係数の標準誤差です。**〔推定係数の〕標準誤差は、推定係数の不確かさ**を示しています。 推定係数の標準誤差が小さいと、推定精度が高いことになります。標準誤差は、以下のような要因によって変化します。

[**1**] 標本数が多ければ推定精度は高まり、推定係数の標準誤差は小さくなります。
[**2**] 説明変数のばらつきが大きく、よく動いてくれた方が推定係数を測りやすく、推定係数の標準誤差は小さくなります。[30]
[**3**] エラーのばらつきが大きいと推定係数を測りづらく、推定係数の標準誤差は大きくなります。
[**4**] 説明変数同士が相関を持つ場合は、どの説明変数の効果かを判別しづらく、推定係数の標準誤差は大きくなります。[31]

[**2**]〜[**4**] に関して、**乱数データ**の [**乱数シート**] の設定を変えて F9 を押すことで確認できます。[**2**] に関しては、〈F4, F5〉の値を大きくすると説明変数のばらつきが大きくなり、推定係数の標準誤差が減少することが分かります。[**3**] に関しては、〈B7〉の値を大きくするとエラーのばらつきが大きくなり、推定係数の標準誤差が増加することが分かります。[**4**] に関しては、〈B6〉の大きさを [-1〜1] の間で変更して、説明変数同士の相関を設定できます。説明変数の x_α と x_β が独立となる 0 を入力すると、どちらの説明変数の効果かが判別しやすく、標準誤差が小さくなることが分かります。一方で ± 1 に近い値を入力すると、推定係数の標準誤差は大きくなることが分かります。

---

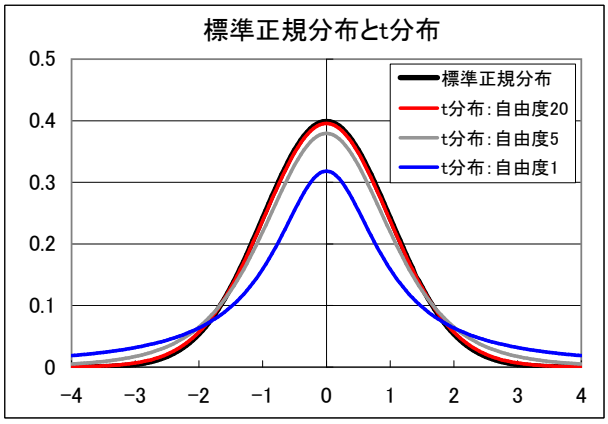
[29] 説明変数の単位を変え（万円を円にする等）10000 倍にした場合、推定係数は $\frac{1}{10000}$ 倍され、調整されます。
[30] 説明変数の単位を変え（万円を円にする等）10000 倍にした場合、標準誤差は $\frac{1}{10000}$ 倍され、調整されます。
[31] 説明変数の相関係数が ± 1 で、完全な多重共線性がある場合はどちらの説明変数の効果かを判別できません。Excel の回帰分析では、自動で完全な共線関係にある説明変数を省き、省いた説明変数の標準誤差は 0 になります。なお、Excel が自動で説明変数を省いた場合、〈分散分析表〉の結果は不正確になります。

5.3 t〔値〕：基準精度で評価した推定係数

推定係数の標準誤差は、推定精度と表裏一体です。よって、標準誤差で調整すれば、基準の推定精度で推定係数を評価できます。以下の (22) 式のように、t **値は標準誤差で割ることで基準精度で評価した推定係数**です。

$$t\text{ 値} = \frac{\text{推定係数}}{\text{推定係数の標準誤差}} \tag{22}$$

推定係数の絶対値が大きければ、t 値の絶対値も大きくなります。また、推定係数の標準誤差が小さければ (22) 式の分母が小さくなり、t 値が大きくなります。[32]



t 値を用いて『説明変数の真の効果が 0 である』という帰無仮説を検定することができます。帰無仮説が正しければ、t 値は t 分布に従うことが知られています。t 分布は、右図のように 0 を中心に左右対称にばらつく山形の分布です。t 分布は自由度が上がると標準正規分布に近づきます。自由度が 20 以上なら、実務上は標準正規分布と考えて差し支えありません。t 分布や標準正規分布において、絶対値からはみ出す面積の割合は、プラス側とマイナス側のそれぞれ約 2.5% で両側約 5% です。帰無仮説のもとでは約 95% の確率で、t 値は絶対値 2 以下に収まります。[33]

t 値が絶対値 2 以下に収まれば、『前提としていた帰無仮説のもと、約 95% の確率の範囲で発生するありふれた t 値だった』という解釈になります。[34] 一方で、t 値が絶対値 2 を超えた時の解釈は、二通りあります。一つの解釈は『帰無仮説のもとで、5% 以下でしか発生しない珍しい t 値だった』です。もう一つの解釈は『5% 以下でしか発生しない t 値が出るのは不自然だ。前提とした “説明変数の真の効果が 0 である” という帰無仮説が間違っていた。』です。どちらの解釈をすべきかは分析者や読者の判断にも依存します。ただ t 値が約 1.7 であれば『10% で発生する珍しい t 値だった』と言っても苦しくありませんが、t 値が約 2.5 であれば『1% で発生する珍しい t 値だった』と言うのは苦しくなります。[35] t 値の絶対値が大きいほど『珍しいこと』という解釈より『不自然なこと ⇒ 前提とした帰無仮説が間違っていた。』という解釈の方がもっともらしくなります。

キリの良さや慣例から、判定基準を両側で 5%（片側で 2.5%）とすることが一般的です。[36] この場合、前述したように t 値が絶対値で 2 を超えているかを目安にできます。t 値が絶対値で 2 を切っ

---

[32] 説明変数の単位を変え（万円を円にする等）10000 倍にした場合、分子の推定係数も分母の標準誤差も $\frac{1}{10000}$ 倍され、t 値は変わりません。

[33] **中心極限定理と擬似正規乱数**(http://keijisaito.info/arc/excel/clt_normal.xls) で F9 を押せば、標準正規分布において約 95% が ±2 の範囲に入ることが視覚的に確認できます。

[34] 『ありふれた t 値』と解釈する範囲に入っても、説明変数の真の効果が 0 でないケースは十分考えられます。この範囲に入っても『帰無仮説を否定しにくい』と言えるだけで『帰無仮説が正しい』とは言えません。

[35] 代表的な有意水準や自由度における t 値は、統計の教科書の巻末などにある t 分布表で照合できます。また Excel 関数で [=TINV(確率, 自由度)] を使うこともできます。両側 5%、自由度 47 では [=TINV(0.05, 47)] と入力すると、2.012 と出力されます。

[36] 『両側 5% の有意水準』といった形で、有意水準 (significance level) という用語も使われます。

ていれば『帰無仮説の前提のもと、約 95% の確率の範囲で発生するありふれた t 値 ⇒ 効果のない説明変数かもしれない。』という解釈になります。一方で、t 値が絶対値で 2 を超えていれば『帰無仮説の前提のもとでは、5% 以下でしか発生しない t 値 ⇒ 効果のある説明変数と考えるのが自然である。』となります。

また、t 値を用いた検定は『推定精度をふまえて、効果が 0 なのか？』を判定しています。このため『推定係数の絶対値が小さくとも、推定精度が高い変数』を効果のある説明変数と判定します。[37] 一方で『推定係数の絶対値が大きくとも、推定精度が低い変数』を効果のない説明変数かもしれないと判定します。t 値は、良くも悪くも不確かさを考慮しているので推定係数と使い分ける必要があります。

5.4 P 値：t 値の検定の境目となる確率

自由度が低い場合や両側 5% 以外の判定基準（有意水準）を用いる場合など、絶対値 2 を目安にする t 値の判定ができない場合もあります。統計の本にある t 分布表を見れば、各自由度と t 値で、帰無仮説を前提とした場合に収まる判定基準を調べることができます。しかし、t 分布表は紙面の都合から、区切りの良い自由度や判定基準しか掲載されていません。また、検定のたびに自由度を照合して t 分布表を見るのは面倒です。そうした面倒なことは、コンピューターに任せて判定基準の境目を出力するのが P 値です。[38] **P 値は『説明変数の効果が 0 である』という帰無仮説のもとで、分析結果の t 値が出る境目の確率**を示しています。

例えば、自由度が 100 で t 値が 2.123 の場合、両側 5% の判定基準では『帰無仮説のもとで珍しい、もしくは不自然な t 値』ですが、両側 1% の判定基準では『帰無仮説のもとでも、ありふれた t 値』になります。この 5% と 1% の間に境目があったはずです。ここで t 値の右隣のセルに P 値が 0.0362 と表示されることから、境目となる判定基準が両側 3.62% であることが分かります。[39] また、両側 5% の判定基準に関しても、『t 値が絶対値で 2 を超えているか？』の目安よりも『P 値が 0.05 を切っているか？』の方が、自由度を考慮していて正確です。

5.5 〔信頼区間の〕下限，上限：真の効果がありそうな範囲

推定係数によって一点で示されている値が、真の係数に一致すると考えるのは楽観的ですが、真の係数は推定係数の近くにあると考えるのが自然です。また、推定係数の標準誤差が小さく推定精度が高ければ、推定係数と真の係数はより近いと考えられます。**〔信頼区間の〕下限，上限 95% は、信頼係数 95% で真の係数がありそうな範囲**を示しています。[40] [41]

---

[37] 万単位の標本があるデータでは、[5.2 〔推定係数の〕標準誤差] の [1] から推定係数の標準誤差が極端に小さくなります。この場合、0 に近い推定係数であっても、t 値は極めて大きくなります。

[38] p 値と小文字で表記される場合もあります。また、P 値には有意確率という訳語も使われます。

[39] Excel 関数でも [=TDIST(t 値, 自由度, 両側=2)] と入力すれば、境目である P 値を表示します。[=TDIST(2.123, 100, 2)] で、0.0362 が出力されます。

[40] ここで確率と言わず、信頼係数という言葉を用いるのは、真の係数は分析者にはっきりと分からないだけで、固有の値があるという発想から来ています。例えば、はっきりと分からなくても『西暦 100 年に大地震が起こったか？』に対して『確率〇 % で起こった』とは言いません。しかし、信頼係数という言葉に馴染めなければ、主観的な確率と読みかえてもかまいません。

[41] 回帰分析のダイアログボックスで 95% 以外の信頼係数を選択することもできます。また、**乱数データ**の [**乱数シート**] では、〈R23〉で指定できます。

6　実証分析を行う際の注意点

この章では、実証分析を行う際の注意点を説明します。例として、被説明変数を傘屋での【傘の販売本数】、説明変数を【降水量、風速、傘の価格】とした以下の (23) 式を用います。1 節〜3 節は水色の部分の回帰分析の式、4 節は黄色の部分のエラー、5 節と 6 節はデータに関する注意点です。

$$【傘の販売本数_i】= b_1 + b_2【降水量_i】+ b_3【風速_i】+ b_4【傘の価格_i】 + e_i \tag{23}$$

6.1　回帰分析の式の形

(23) 式の形では【降水量】が『0mm → 1mm』でも『100mm → 101mm』でも一定の効果が【傘の販売本数】にあることを前提にしています。また、【降水量】と【風速】は独立して【傘の販売本数】に影響を与える形になっており、雨と風の相乗効果は考えていません。(23) 式の形では、説明変数の効果が水準に依存せず一定であること、説明変数同士の相乗効果がないことをあらかじめ決めつけています。一般的に、あらかじめ決めつけた式の形が複雑な現実の関係を正しく表していると考える方が不自然です。[42] 実証分析において、回帰分析の式の形（関数形）はどうやっても近似にすぎませんが、無理のない近似になっている必要があります。『**回帰分析の式の形は、現実の関係の無理のない近似になっているか？**』という注意点があります。[43]

[変数の変換]

式の形が現実の良い近似となるように、被説明変数や説明変数をあらかじめ変換しておく場合があります。例えば、被説明変数 y に自然対数をとることで推定係数 b_j は x_j が 1 増えた場合の y の変化率として解釈できます。[44] [45] また、説明変数 x_j に自然対数をとることで推定係数 b_j は x_j が 1% 増加した場合の y への効果と解釈できます。[46] (23) 式の b_4 の解釈は、『【傘の価格】が 1 円上昇した場合の【傘の販売本数】に与える効果』ですが、以下の (24) 式の b_{4*} は『【傘の価格】が 1% 上昇した場合の【傘の販売本数】に与える効果』と解釈できます。

$$b_1 + b_2【降水量_i】+ b_3【風速_i】+ b_{4*} \log(【傘の価格_i】) \tag{24}$$

[ダミー変数]

標本として得られているデータの中に地域、時点などのグループがあれば、ダミー変数を用いてグループの違いを回帰分析に入れることができます。グループ A とグループ B で切片が異なる可能性のある場合は、以下の (25) 式のように定数項ダミーを説明変数に加えます。

$$b_1 + b_2【降水量_i】+ b_3【風速_i】+ b_4【傘の価格_i】+ b_5【定数項ダミー_i】 \tag{25}$$

---

[42] 式の形を決めつけないノンパラメトリック回帰もありますが、推定結果の表示や解釈が難しくなります。

[43] 式の形は標本の範囲を含めて考えます。日常的な【降水量】のみの範囲では、水準に依存せず一定の効果と近似しても無理がないかもしれません。しかし、その回帰分析での推定係数は、大雨において【降水量】が【傘の販売本数】に与える効果とは異なると考えられます。

[44] 自然対数の微分が $\frac{d\log(y)}{dy}=\frac{1}{y}$ と分数になることから、被説明変数 y に対数をとった場合は $b_j \simeq \frac{\frac{\triangle y}{y}}{\triangle x}$ となります。

[45] 自然対数の Excel 関数は natural logarithm の ln(数値) です。Excel 関数で log(数値) として対数の底を省略すると、10 を底とする常用対数になるので注意してください。

[46] x_j と y の双方に自然対数をとれば、推定係数は変化率と変化率の比である弾力性として解釈できます。

定数項ダミーは基準のグループ A なら 0、グループ B なら 1 を入力した列です。Excel の回帰分析にこの列を説明変数として加えることで、定数項ダミーを加えた回帰分析となります。(25) 式において、グループ A の切片の推定値は b_1 ですが、グループ B の切片の推定値は $b_1 + b_5$ となります。

また、グループによって説明変数の効果が異なる可能性を考え、以下の (26) 式のように係数ダミーを設定することもできます。

$$b_1 + \{b_2 + b_6【係数ダミー_i】\}【降水量_i】 + b_3【風速_i】 + b_4【傘の価格_i】 \tag{26}$$

(26) 式の係数ダミーは (25) の定数項ダミーと同じ値が入り、降水量$_i$ の推定係数は基準のグループ A には b_2、グループ B には $b_2 + b_6$ となります。しかし、(26) 式の形では、Excel で回帰分析を実行できません。Excel で回帰分析が実行できるようにするには、以下の (27) 式のように降水量をグループ毎に分け、該当グループ以外を 0 とした新しい説明変数の列を作ります。

$$b_1 + b_{2*}【降水量_{i\mathrm{A}}】 + b_{6*}【降水量_{i\mathrm{B}}】 + b_3【風速_i】 + b_4【傘の価格_i】 \tag{27}$$

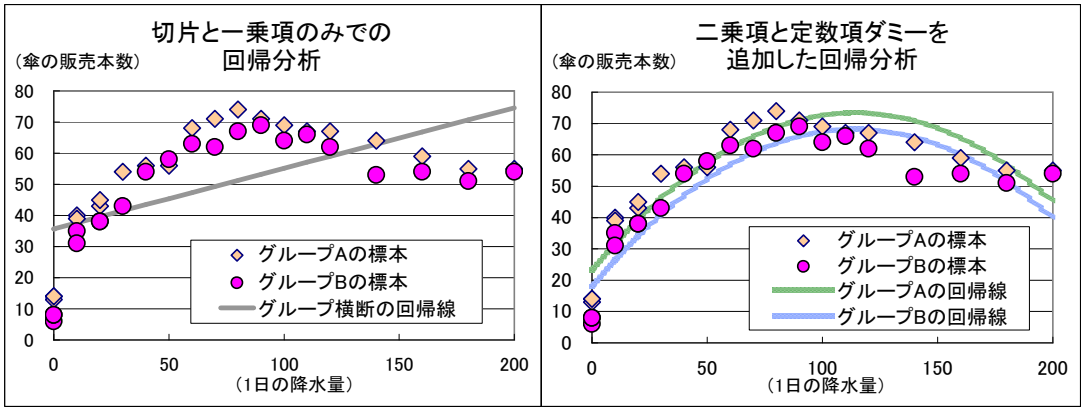
(27) 式において、【降水量】の推定係数はグループ A には b_{2*}、グループ B には b_{6*} となります。$b_{6*} = b_2 + b_6$ と変換すると、(26) 式と (27) 式の回帰分析の結果は同じになります。

グループ A、グループ B、グループ C、・・・ と 3 グループ以上ある場合にも同様の手順で定数項ダミー、係数ダミーを指定ができます。また、地域と時点といった形で複数の種類のグループに別々のダミー変数を設定することもできます。[47]

[近似の確認と方法]

多重回帰であっても、j 種類目の説明変数 x_j と被説明変数 y の散布図を見ることで、両者の関係に見当をつけることができます。また、候補の式の形で回帰分析を行い、得られた当てはめ値 $\hat{y}$ と推定エラー e の散布図を用いて確認できます。例えば、左下の図では切片と一乗項だけの式では良い近似とはならないことが分かります。また、結果的に得られた $\hat{y}$ と e の散布図を描けば、$\hat{y}$ の大きさによって e が偏っていることが分かります。

近似の度合いを高める簡単な方法は、(降水量$_i$)^2 の二乗項や (降水量$_i$ · 風速$_i$) の交差項を説明変数に追加することです。右下の図では、左下の図に (降水量$_i$)^2 の二乗項と定数項ダミーを追加することで左下の図より現実の関係に近くなっています。しかし、複雑な関係も許容する式の形が望ましい反面、単純化して分かりやすく表すことも重要なのでバランスをとる必要があります。



[47] パネルデータに対する固定効果モデルになります。**学校は人的資本を形成するのか？の賃金格差の実証分析**(http://keijisaito.info/econ/jp/gjk/j2_wage_data.htm) では、年齢層と生年層での固定効果モデルから賃金格差の変化をグラフに表しています。

6.2 必要な説明変数

複数の説明変数を用いる多重回帰では、他の説明変数の効果を調整した上での各説明変数の効果を推定しています。(23) 式は【降水量、風速、傘の価格】を用いた多重回帰ですが、感覚的に一番効果が大きそうな説明変数は【降水量】です。しかし、(23) 式における【降水量】のデータがなかったり、説明変数であることに気がつかない場合があります。[48] また、雨が降る時には風が強いことが多く、(23) 式から【降水量】を抜いて回帰分析を行うと【風速】が大きな効果を持っているという推定結果が得られそうです。しかし、仮に『風が強いだけの日』があっても【傘の販売本数】は伸びそうにありません。必要な説明変数を省くと、他の説明変数の推定値にも悪影響があります。**『必要な説明変数が全て回帰分析の式に入っているか？』**という注意点があります。

この注意点に対しては、必要な説明変数をよく考える必要があります。また、説明変数を入れるか省くかを迷ったら、入れる方が無難です。[49] 一般的に関係のある説明変数を省くと、他の説明変数の推定係数に悪影響があります。一方、無関係の説明変数を入れると推定精度は落ちますが、現実に対応しない推定係数になる傾向はありません。

6.3 効果の方向

(23) 式では【降水量、風速、傘の価格】が【傘の販売本数】に与える効果を調べています。しかし、『傘がよく売れる時期や場所では、傘屋が傘の値上げをして利益を出す』という逆方向の効果があるかもしれません。傘の価格は安い方がよく売れるなら (23) 式の b_4 はマイナスになるはずですが、逆方向の効果があれば b_4 は 0 に近づき、プラスになることも考えられます。また、逆方向の効果の極端な例として【降水量】を被説明変数にして【傘の販売本数】を説明変数にすることが考えられます。理屈の上では【傘の販売本数】が伸びても、雨は降らないことが分かります。しかし、機械的な回帰分析では『傘が売れれば、雨が降る』という結果になります。これらの結果を単純に解釈すると『傘を多く売るために値上げをしよう』や『雨を降らせるために傘を買おう』という話になりかねません。**『説明変数⇒被説明変数の効果の方向は正しいか？』**という注意点があります。

この注意点に対しては、効果の方向や経路をよく考える必要があります。また、【降水量、風速】は、『傘屋にはどうにもならないもの』です。こうした説明変数なら『逆方向の効果があるかもしれない』という問題はありません。しかし、理系の実験と違い、経済などのデータは『景気と失業の関係』『企業の業績と広告費の関係』のように因果が分からないこと、両方向が考えられるケースが多々あります。[50] こうしたケースでは、時間的な前後関係を利用することで『雨が降ってから、傘が売れる』のか『傘が売れてから、雨が降る』のかを識別する場合があります。この場合は、過去の説明変数が将来の被説明変数に影響を与えるという枠組みで回帰分析を行います。[51] また、操作変数法によって対処できる場合もあるので、関心のある方は計量経済学の本を参照してください。[52]

---

[48] 【降水量】のデータがないので【湿度】を説明変数に用いるなど、代理変数を使う場合があります。しかし、代理変数の推定値の解釈には注意が必要です。また、[6.6 データの正確さ] の問題も生じます。

[49] 効果や他の推定値への影響がないことを確認してから説明変数を減らしたり、複数の推定結果を示して安定性を論じたりすることがあります。

[50] 逆方向の効果が考えられない対象を取り出すことで、擬似的な実験環境を分析することもあります。

[51] 『雨が降りそうだから、傘を買う』『景気の見通しが暗いから、失業が増える』といった形で、時間的な前後関係が効果の方向に対応しない場合もあります。

[52] 操作変数法の一つである二段階最小二乗法は、推定係数を導出するだけなら Excel の分析ツールでも実行できます。

6.4　エラーの形状と分散

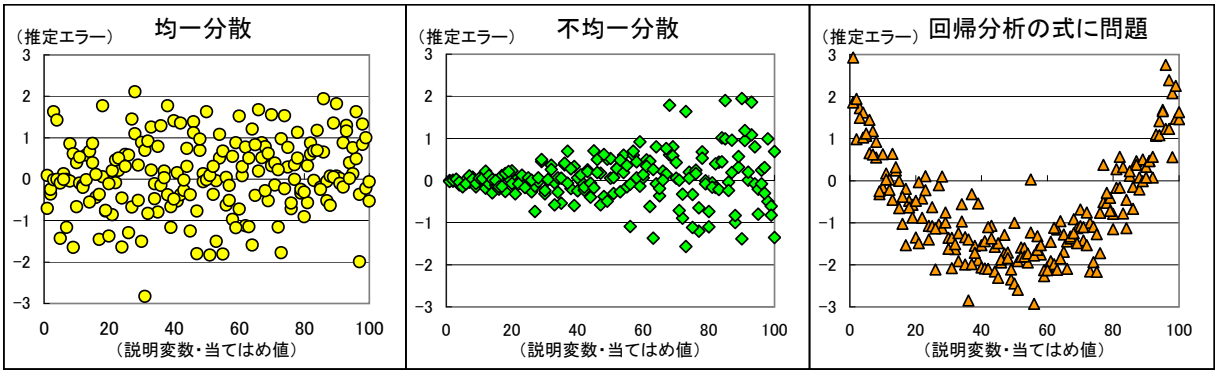
Excel の回帰分析の結果は、推定エラー e の背後にある真のエラー ϵ は、以下の (28) 式の平均 0、分散 σ^2 の正規分布から発生することを仮定して導出しています。[53] [54]

$$\epsilon \sim N(0, \sigma^2) \tag{28}$$

(28) 式の仮定では、どの標本に関してもエラーは一定の分散の正規分布から発生します。しかし、【降水量】の多い日はエラーのばらつきが大きいなど、説明変数の大きさによってエラーの分散が異なる場合があります。こうした場合は、最小二乗法よりも精度が高い推定方法が存在します。また、最小二乗法の [推定係数の標準誤差] や [帰無仮説の検定結果] は不正確になります。『**エラーは一定の分散の正規分布から発生するか？**』という注意点があります。

実証分析の場合、真のエラー ϵ は回帰分析の後でも分かりません。[55] このため、推定エラー e から真のエラー ϵ を推測するしかありません。この確認として、推定エラー e と主な説明変数 x_j や当てはめ値 $\hat{y}$ との散布図を描きます。[56] 散布図が左下図のようになっていれば、エラーのばらつきは説明変数から影響を受けない均一分散と考えられます。一方、中央下図のようであれば、説明変数が増加するとエラーのばらつきが増える不均一分散であると考えられます。もし、右下図のようになっていれば、[6.1 回帰分析の式の形] や [6.2 必要な説明変数] の問題であると考えられます。

この注意点に対する対処として、エラーの構造を指定する一般化最小二乗法やグループによってエラーの分散が異なるというランダム効果モデルを使うこともあります。[57] しかし、中央下図のような不均一分散がある場合に、通常の回帰分析（最小二乗法）を行っても深刻な問題は起きません。通常の回帰分析でも、現実に対応した推定係数を得ることができます。



6.5　データの抜け落ち

快晴の日は傘屋が臨時休業してデータが抜けている可能性があります。また、傘の売れ行きが悪い場合、傘屋は回答を拒否するかもしれません。傘屋の状態や判断を通じた休業や回答拒否がある

---

[53] (19) 式のエラーの標準誤差は σ（シグマ）の推定値です。また、(19) 式の推定には正規分布の仮定は不要です。

[54] 中心極限定理の一つとして『同じ分布に従っていなくても、独立な確率変数の和は正規分布に近づく』があります。このため、特に理由がない限り種々のノイズの合計値である回帰分析のエラーは正規分布に従うと考えます。

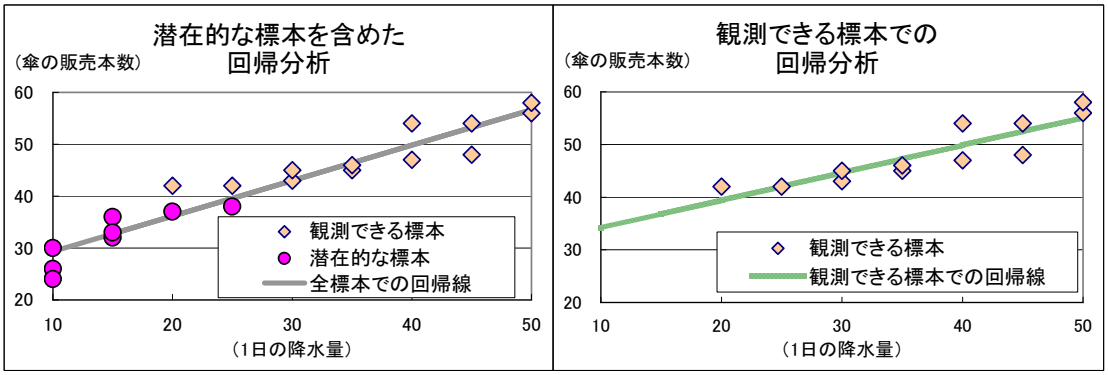
[55] **乱数データ**のようなシミュレーションでは、ϵ のばらつきの大きさを設定でき、値を確認できます。

[56] 時系列データなら、ダービン・ワトソン統計量 $\frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2}$ を使って検定することもあります。

[57] Excel の分析ツールでは、不均一分散に対応した推定はできません。

場合、回帰分析の結果は現実に対応しなくなる可能性があります。潜在データを含めた左下図の回帰線は、観測できるデータのみの右下図の回帰線とはっきりと異なっています。**『回答やデータ作成に偏りのある抜け落ちがないか？』**という注意点があります。

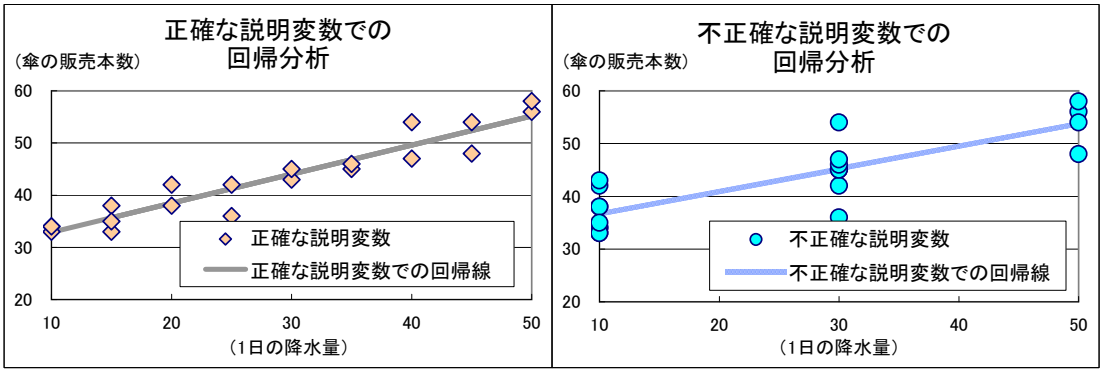
多くのケースでは、既に存在するデータを使うしかありません。回答やデータの作成プロセスで標本が偏る可能性がないかを考え、必要に応じて記述します。また、Heckit（ヘックマンの二段階推定）などのサンプルセレクションに対応した推定方法があります。ただ、サンプルセレクションに対応した推定方法は休業、無回答といった『この標本のデータが得られていない』という情報が必要です。[58]



6.6 データの正確さ

【傘の販売本数】が 10 本単位であったり、【降水量】が県単位の調査であったりとデータが不正確になっている場合があります。観測段階での測定誤差、データ公表段階での四捨五入や代表値、回帰分析の段階での代理変数など、複数の要因から回帰分析に用いるデータは不正確になります。**『データは正確か？』**という注意点があります。

被説明変数 y が不正確であれば、推定エラーが大きくなり、推定精度が落ちます。しかし、現実と異なる推定係数になる傾向はありません。一方で説明変数 x が不正確であれば、右下図のように現実の関係を離れて推定係数が小さくなる傾向があります。[59] 説明変数が不正確である方がより深刻ですが、いずれの変数もできるだけ正確なデータを使ってください。



[58] サンプルセレクションに対応した推定では、データから抜け落ちる確率の変換値を説明変数として加えます。近似的な方法であれば Excel の分析ツールでも実行できます。

[59] 説明変数の不正確さの大きさが分かれば、真の推定係数を見積もることができます。また、操作変数法を使うことでも対処できます。