11. 回帰分析

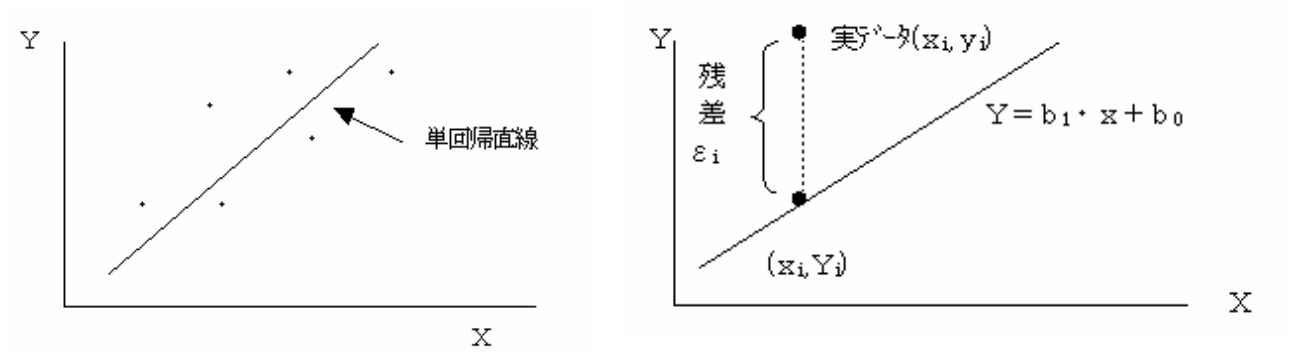
どんな現象にもそれを引き起こす原因がある．多くの原因が考えられる中で，どの原因が主であるか．また，注目した原因が具体的にどの程度その現象に関わっているのか．このような問題を扱う分析が**回帰分析** (regression analysis)である．つまり，回帰分析とは，原因系の変数(x)と結果系の変数(y)に関してデータを何組か採り，そのデータを使って x と y の関係をあらわす公式 y=f(x)を作成する作業であると言える．

11.1 単回帰分析の考え方

月刊誌に連載記事を執筆することになった．要求された原稿は刷り上がりでおよそ 5〜6 ページである．要求された紙幅で原稿を書くことが求められているのは当然だが，なにせ図表を含む原稿なので，一概に原稿のページ数だけでページ数を決めることができない．また，図表は出版社で縮小される場合があるので，自分で図表も配置した原稿を作ったとしても，それが正確に刷り上がりページ数を反映されるわけでもない．

こうした場合に，自分なりのフォーマットで作成した原稿のページ数から，雑誌に掲載される際の刷り上がりページ数を予測する式を作成することができれば，原稿を執筆する際に量的なめどが立てやすいだろう．

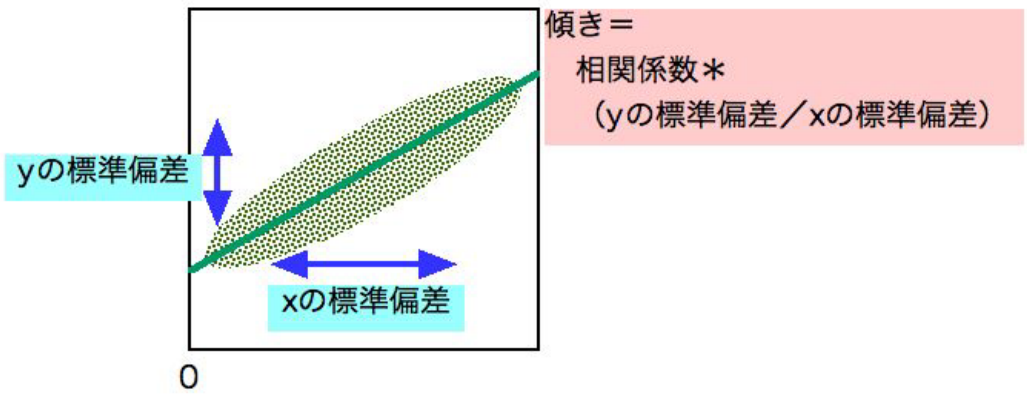
この「予測式の作成」に用いるのが，刷り上がりページ数を目的 (従属) 変数，原稿のページ数を説明 (独立) 変数とした単回帰分析である．具体的には，目的変数と説明変数の間の相関係数を求め，それにもとづいて予測式を立てる．



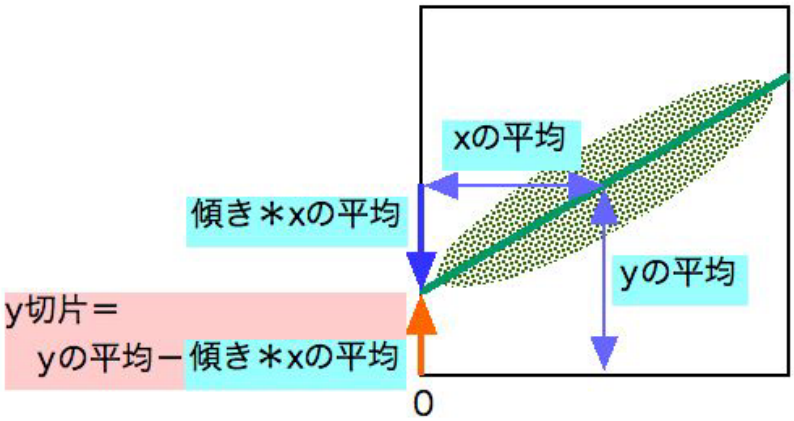
(x, y)=(説明変数のデータ，目的変数のデータ)の散布図を描き，x と y の関係をあらわす適当な直線を考える．この直線を単回帰直線といい，単回帰直線は，右のグラフで示す残差 (つまり実データとのずれ) が最小となるように定める必要がある．

単回帰直線を求めるためには，残差を最小にするような直線の y 切片と傾きを求める必要がある．回帰分析では，単回帰直線の残差を最小にするための方法として「最小 2 乗法 (残差の 2 乗和を計算して，それをもっとも小さくする直線を求める手法)」を用いる．この傾きのことを回帰係数と呼ぶ．最小 2 乗法によって傾きと y 切片を求めるためには，次のような計算式を用いる．

● 傾き＝x と y の相関係数×(y の標準偏差／x の標準偏差)



● y 切片＝y の平均－(傾き×x の平均)



では，この公式にもとづいて，原稿のページ数から刷り上がりのページ数を予測するための単回帰直線を求めてみよう．

データ(14 回連載した実績)

連載	刷上ページ数	原稿ページ数
1	4.3	6.2
2	4.4	8.1
3	6.3	10.0
4	6.8	13.2
5	6.0	9.7
6	6.5	11.3
7	6.5	11.3
8	7.0	12.2
9	5.7	8.8
10	5.5	8.3
11	7.0	9.5
12	7.0	10.5
13	6.1	8.7
14	6.5	8.8

```
data regsample:
input rensai shuppan genko;
cards:
1 4.3 6.2
2 4.4 8.1
3 6.3 10.0
4 6.8 13.2
5 6.0 9.7
6 6.5 11.3
7 6.5 11.3
8 7.0 12.2
9 5.7 8.8
10 5.5 8.3
11 7.0 9.5
12 7.0 10.5
13 6.1 8.7
14 6.5 8.8
;
proc means;
  var shuppan genko;
proc corr;
  var shuppan genko;
run;
```

SAS の means，corr プロシジャによって，原稿のページ数と刷り上がりページ数の平均，標準偏差，および両者の相関係数を求めると，

x：原稿のページ数(genko)の平均 9.757，標準偏差 1.834

y：刷り上がりページ数(shuppan)の平均 6.114，標準偏差：0.881

相関係数：0.782

となる．よって，単回帰直線は，

傾き＝0.782×(0.881／1.834)＝0.376

y 切片＝6.114－(0.376×9.757)＝2.445

となり，刷り上がりページ数は，

刷り上がりページ数 y＝0.376×原稿のページ数 x＋2.445

という予測式によって大体のめどをつけられるということになる．

11.2 SAS による単回帰分析

では次に，回帰分析をおこなう PROC REG プロシージャを用いて単回帰分析をおこなってみよう．

PROC REG <options>;
 model 目的変数＝説明変数;
 print 出力指定
  stb：標準化した回帰係数の出力
  r：残差＝回帰式による予測値とのずれ
(stb オプションは model の後に / に続けて記述してもよい)

GLM プロシージャと同様，プログラムの最後に quit; を記述する必要がある．このプログラムを実行すると，次のようなアウトプットが出力される．

```
data regsample;
input rensai shuppan genko;
cards;
1 4.3 6.2
2 4.4 8.1
3 6.3 10.0
4 6.8 13.2
5 6.0 9.7
6 6.5 11.3
7 6.5 11.3
8 7.0 12.2
9 5.7 8.8
10 5.5 8.3
11 7.0 9.5
12 7.0 10.5
13 6.1 8.7
14 6.5 8.8
;
proc reg;
 model shuppan=genko ;
 print stb r;
run; quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: shuppan

モデルの分散分析

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.17131	6.17131	18.86	0.0010
Error	12	3.92583	0.32715		
Corrected Total	13	10.09714			

モデルは有意

Root MSE	0.57197	R-Square	0.6112 → 相関係数 0.782の2乗
Dependent Mean	6.11429	Adj R-Sq	0.5788 → (自由度調整済) 決定係数
Coeff Var	9.35469		

Parameter Estimates

Variable	DF	Parameter Estimate		Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	2.44906	y切片	0.85763	2.86	0.0145	0
genko	1	0.37565	傾き	0.08649	4.34	0.0010	0.78179

Output Statistics

Obs	Dep Var shuppan 実データ	Predicted Value 予測値	Std Error Mean Predict	Residual 残差	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	4.3000	4.7781	0.3435	-0.4781	0.457	-1.045	**	0.308
2	4.4000	5.4918	0.2095	-1.0918	0.532	-2.051	****	0.326
3	6.3000	6.2055	0.1543	0.0945	0.551	0.172		0.001
4	6.8000	7.4076	0.3347	-0.6076	0.464	-1.310	**	0.447
5	6.0000	6.0928	0.1529	-0.0928	0.551	-0.168		0.001
6	6.5000	6.6939	0.2029	-0.1939	0.535	-0.362		0.009
7	6.5000	6.6939	0.2029	-0.1939	0.535	-0.362		0.009
8	7.0000	7.0319	0.2608	-0.0319	0.509	-0.0627		0.001
9	5.7000	5.7547	0.1738	-0.0547	0.545	-0.100		0.001
10	5.5000	5.5669	0.1981	-0.0669	0.537	-0.125		0.001
11	7.0000	6.0177	0.1545	0.9823	0.551	1.784	***	0.125
12	7.0000	6.3933	0.1658	0.6067	0.547	1.108	**	0.056
13	6.1000	5.7172	0.1781	0.3828	0.544	0.704	*	0.027
14	6.5000	5.7547	0.1738	0.7453	0.545	1.368	**	0.095

予測値と実測値との残差

Sum of Residuals	0
Sum of Squared Residuals	3.92583
Predicted Residual SS (PRESS)	5.54847

この結果から，回帰式は，

y (刷り上がりページ数) ＝0.37565×x (原稿のページ数) ＋2.44906
傾き　　y 切片

という直線であらわすことができることがわかり，先ほどの計算結果とほぼ一致している．
また，残差の出力から，第 2 回目が予測値からもっとも大きく外れたケースであることも分かる．

また，Standardized Estimate に出力されているのは，傾き＝回帰係数を標準化 (平均 0，分散 1 に変換する→Z 変換と呼ぶ) した値であり，単回帰分析の場合は目的変数と説明変数の相関係数と一致する．有意性の検定 (帰無仮説：標準化された回帰係数＝0) は *t* 検定によっておこなわれる．この統計量は，単位やちらばり具合の異なる複数の変数を説明変数として投入する場合 (重回帰分析) に，それぞれの変数の影響力の「強さ」を相対的に判定する際によく用いられる．

課題 1&2

以下のデータから，「入試の成績」から「入学後の成績」を予測する回帰方程式を求め，および回帰係数の有意性を検討せよ

学生番号	入試時	入学後
1	60	65
2	50	45
3	70	75
4	80	60
5	90	80
6	85	90
7	65	55
8	60	75
9	45	40
10	45	60

課題 1
means, corr プロシージャを用いて得られた結果から，傾き (回帰係数) と y 切片を計算せよ

課題 2
reg プロシージャを用いて回帰分析をおこない，上記の結果と一致することを確認し，もっとも残差の大きいケースを特定せよ．また，得られた回帰係数の有意性を検討せよ．