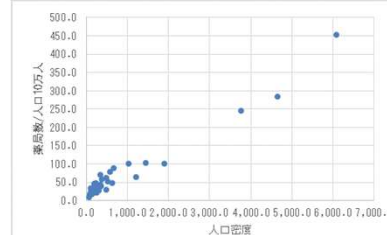


回帰分析

佐賀大学 新井康平

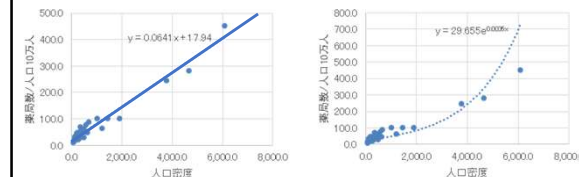
1

- ・回帰とは、目的変数について説明変数を使った式で表すことをいいます。この式のことを「回帰方程式」、あるいは簡単に「回帰式」といいます。また、回帰式を求めることを「回帰分析」といいます。
- ・次の散布図は都道府県の人口密度と人口10万人あたりの薬局の数を示したものです。薬局の数を目的変数、人口密度を説明変数とするとき、回帰式を求めることができるでしょうか。



2

次の2つの図は散布図上に回帰式を描いたものです。このように、データに対しては様々な回帰式を求めることができます：**単回帰**



3

重回帰分析

県	A	B	C	D	E	F
1		年日照時間 (時間)	年平均気温 (°C)	人口密度 (人/km2)	持ち家比率 (%)	降水量 (mm)
2	北海道	1,648	9.2	69.3	57.2	1,347
3	青森県	1,516	10.5	138.4	71.7	1,664
4	岩手県	1,671	10.6	84.8	71.9	1,643
5	宮城県	1,880	12.7	319.5	60.8	1,112
6	秋田県	1,469	11.9	90.2	78.4	2,373
43	長崎県	2,018	17.5	340.2	65.7	1,684
44	熊本県	2,197	17.2	243.2	64.3	1,975
45	大分県	2,186	16.9	185.8	62.6	1,506
46	宮崎県	2,411	17.9	144.8	66.0	2,080
47	鹿児島県	2,183	18.9	182.8	65.8	1,778
48	沖縄県	1,809	23.3	621.5	50.2	2,071

- ・このデータでは年日照時間を目的変数とし、残りの4変数を説明変数とします。年平均気温を、人口密度を、持ち家比率を、降水量をとすると、次のような重回帰式を求める

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\hat{y} = 538.7 + 77.2x_1 + 0.04x_2 + 11.2x_3 - 0.26x_4$$

4

重回帰分析の偏回帰係数も、単回帰分析と同様に最小二乗法で求めます。このデータを用いてエクセル統計で重回帰分析を行うと、偏回帰係数について次のような結果が得られます

県	A	B	C	D	E	F	G	H	I	J
1	回帰式に含まれる変数 (偏回帰係数・信頼区間等)									
2	変数	偏回帰係数	標準誤差	標準偏回帰係数	下限値	上限値	F値	t値	P値	*: P<0.05 **: P<0.01
3	年平均気温(°C)	77.2442	12.8295	0.7195	51.3532	103.1351	36.2503	6.0208	p < 0.001	**
4	人口密度 (人/km2)	0.0420	0.0302	0.1923	-0.0189	0.1029	1.9352	1.3911	0.1715	
5	持ち家比率 (%)	11.2115	5.5100	0.3145	0.0919	22.3311	4.1403	2.0348	0.0482	*
6	降水量 (mm)	-0.2594	0.0582	-0.4897	-0.3769	-0.1420	19.8804	-4.4587	p < 0.001	**
7	定数項	538.6966	480.8432		-431.6842	1509.0774	1.2551	1.1203	0.2689	

5

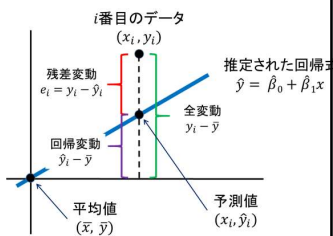
決定係数：R²値

- ・決定係数はデータに対する、推定された回帰式の当てはまりの良さ（度合い）を表します。
- ・決定係数は一般にσで示され、0から1までの値をとります。
- ・1に近いほど、回帰式が実際のデータに当てはまっていることを表しており、説明変数が目的変数をよく説明している

6

変動は二乗和として算出

- ・「全変動」：実際のデータとデータ全体の平均値との差を表します（上の図の緑の部分）
- ・「回帰変動」：推定された回帰式から得られた予測値とデータ全体の平均値の差を表します（上の図の紫の部分）
- ・「残差変動」：実際のデータと推定された回帰式から得られた予測値との差を表します（上の図の赤の部分）



7

「全変動」＝「回帰変動」＋「残差変動」

- ・決定係数は、説明変数が目的変数をどれくらい説明しているか、つまり「回帰変動が全変動に対してどれだけ多いか＝残差変動が全変動に対してどれだけ少ないか」を表すものです。したがって決定係数は、次に示すように回帰変動を全変動で割ることで求められます。

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

	変動	計算式
全変動の平方和		$\sum_{i=1}^n (y_i - \bar{y})^2$
回帰変動の平方和		$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
残差変動の平方和		$\sum_{i=1}^n (y_i - \hat{y}_i)^2$

8

- ・次のデータは2015年プロ野球12球団のチーム成績を表したものです。このデータを使い、「打率」、「得点」、「失点」、「防衛率」の4つの説明変数から「勝利数」を予測するための重回帰式を求めます。

県	A	B	C	D	E	F
1	球団	勝利数	打率	得点	失点	防衛率
2	福岡ソフトバンクホークス	90	0.267	651	491	3.16
3	北海道日本ハムファイターズ	79	0.258	615	581	3.62
4	千葉ロッテマリーンズ	73	0.257	561	563	3.69
5	西武ライオンズ	69	0.263	631	573	3.69
6	オリックス・バファローズ	61	0.249	519	548	3.59
7	東北楽天ゴールデンイーグルス	57	0.241	463	612	3.82
8	東京ヤクルトスワローズ	76	0.257	574	518	3.31
9	読売ジャイアンツ	75	0.243	489	443	2.78
10	阪神タイガース	70	0.247	465	550	3.47
11	広島東洋カープ	69	0.246	506	474	2.92
12	中日ドラゴンズ	62	0.253	473	504	3.19
13	横浜DeNAベイスターズ	62	0.249	508	598	3.8

9

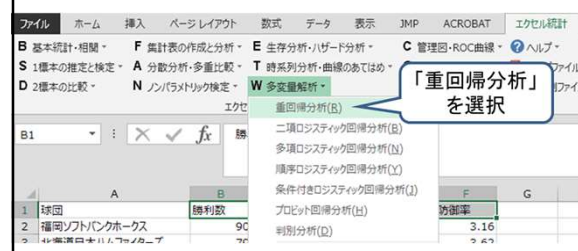
データ選択

	A	B	C	D	E	F
1	対戦	勝利数	打率	得点	失点	防御率
2	福岡ソフトバンクホークス	90	0.267	651	491	3.16
3	北海道日本ハムファイターズ	79	0.258	615	581	3.62
4	千葉ロッテマリーンズ	73	0.257	561	563	3.69
5	西武ライオンズ	69	0.263	631	573	3.69
6	オリックス・バファローズ	61	0.249	519	548	3.59
7	東北楽天ゴールデンイーグルス	57	0.241	463	612	3.82
8	東京ヤクルトスワローズ	76	0.257	574	518	3.31
9	読売ジャイアンツ	75	0.243	489	443	2.78
10	阪神タイガース	70	0.247	465	550	3.47
11	広島東洋カープ	69	0.246	506	474	2.92
12	中日ドラゴンズ	62	0.253	473	504	3.19
13	横浜DeNAベイスターズ	62	0.249	508	598	3.8

B1からF1セルの範囲を選択

10

手法選択

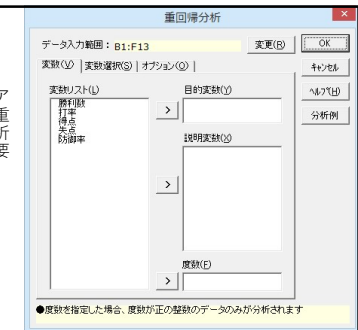


「重回帰分析」を選択

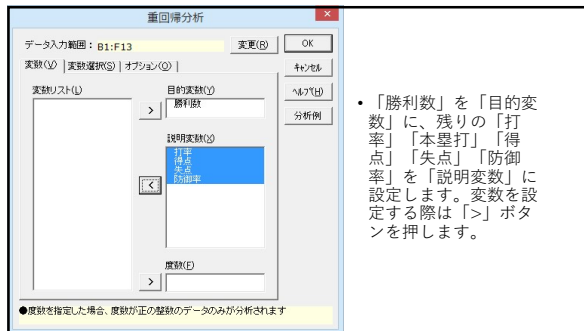
11

変数選択

- 手法を選択するとダイアログが表示されます。重回帰分析ではまず、分析する変数を選択する必要があります。

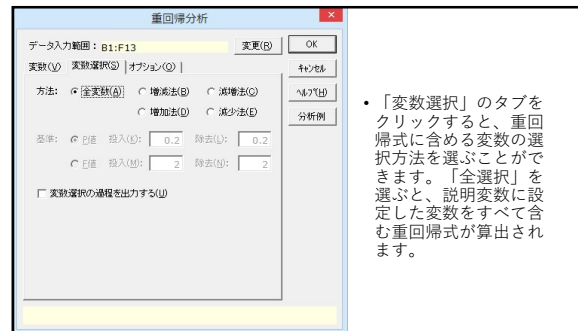


12



- 「勝利数」を「目的変数」に、残りの「打率」「本塁打」「得点」「失点」「防御率」を「説明変数」に設定します。変数を設定する際は「>」ボタンを押します。

13



- 「変数選択」のタブをクリックすると、重回帰式に含める変数の選択方法を選ぶことができます。「全選択」を選ぶと、説明変数に設定した変数をすべて含む重回帰式が算出されます。

14

回帰式の精度

- 回帰式の精度は「自由度調整済み決定係数」から確認できることは27-4章で既に学びました。エクセル統計では決定係数の「修正R2乗」に出力されます。このデータでは0.6121となっており、目的変数を説明変数によって6割程度説明できるということを示しています。

59	回帰式の精度			
60	重相関係数		決定係数	
61	R	修正R	R2乗	修正R2乗
62		0.8679	0.7824	0.7532

15

重相関係数

- 実際のデータと回帰式から求められた予測値との相関係数を「重相関係数R」といいます。また、説明変数の数の割合が多くなるほど重相関係数の値は過大評価されやすいため、この値を調整したものを「自由度調整済み重相関係数（修正R）」といます。決定係数は重相関係数の2乗に等しくなります。

16

LINEST関数による重回帰分析 **最尤推定**

- LINEST 関数は、「最小二乗法」を使って指定したデータに最もよく適合する直線を算出し、この直線を記述する配列を返すことによって直線の補正項を計算します。LINEST 関数を他の関数と共に使用して、多項式近似、対数近似、指数近似、べき級数をはじめとする、不明なパラメーター内で線形近似を示す他の種類のモデルの統計を計算することもできます。この関数は値の配列を返すため、配列数式として入力する必要があります。直線は次の方程式で表されます。

$$y = mx + b: \text{単回帰}$$

または

$$y = m_1x_1 + m_2x_2 + \dots + b: \text{重回帰}$$

- これは、x の値が複数の範囲にある場合に適用されます（ここで、従属変数 y は独立変数 x の関数です）。

17

LINEST関数

- m の値はそれぞれの x の値に対応する係数であり、b は定数です。y、x、および m がベクトル（1 次元配列）であり得ることに注意してください。LINEST 関数が返す配列は、 $\{m_1, m_2, \dots, m_n, b\}$ となります。また、回帰直線の補正項も追加情報として返されます。

$$y = mx + b \rightarrow \text{LINEST(目的変数, 説明変数)} \rightarrow (m, b)$$

$$\text{傾き: 比例係数項} = \text{INDEX}(\text{LINEST}(\text{既知の } y, \text{既知の } x), 1)$$

- y 切片: 定数項

$$= \text{INDEX}(\text{LINEST}(\text{既知の } y, \text{既知の } x), 2) \quad b = \bar{y} - m\bar{x}$$

18

=INDEX(LINEST(H2:H18,\$B\$2:\$F\$18),1)

	Total(A)	Total(B)	Total
BT	-0.00738	-0.00197	-0.00934
BP(H)	-0.04382	-0.57586	-0.61969
BP(L)	0.735262	-0.84447	-0.10921
HB	-0.09252	-0.08335	-0.17587
NS	-19.1292	8.250936	-10.8783
Const.	738.9954	-125.069	613.9267