

データ解析 第13回 回帰分析

北九州市立大学経済学部

齋藤 朗宏

今日の内容

- 単回帰分析
- 重回帰分析

2010年セ・リーグ順位表

	勝率	打率	得点	本塁打	盗塁	失点	被本塁打
中日	0.560	0.259	539	119	53	521	121
阪神	0.553	0.290	740	173	71	640	138
巨人	0.552	0.266	711	226	96	617	140
ヤクルト	0.514	0.268	617	124	66	621	146
広島	0.408	0.263	596	104	119	737	171
横浜	0.336	0.255	521	117	59	743	176

プロ野球では、勝率が最も高いチームが優勝となる。つまり、最終目標は、高い勝率を獲得する事である。そのためには、どんな要因が必要となるだろうか？

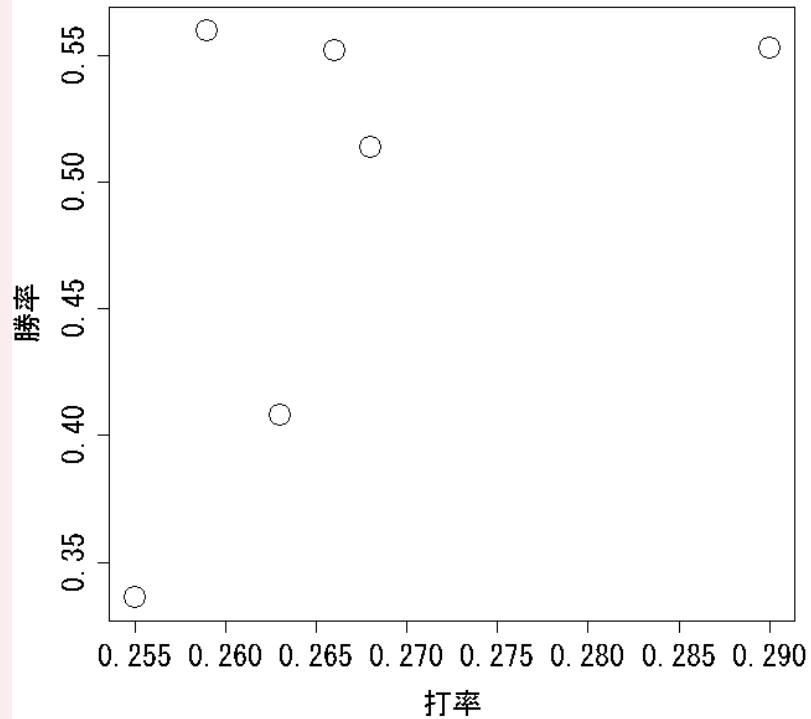
<http://bis.npb.or.jp/2010/stats/> より作成。

2010年セ・リーグ順位表

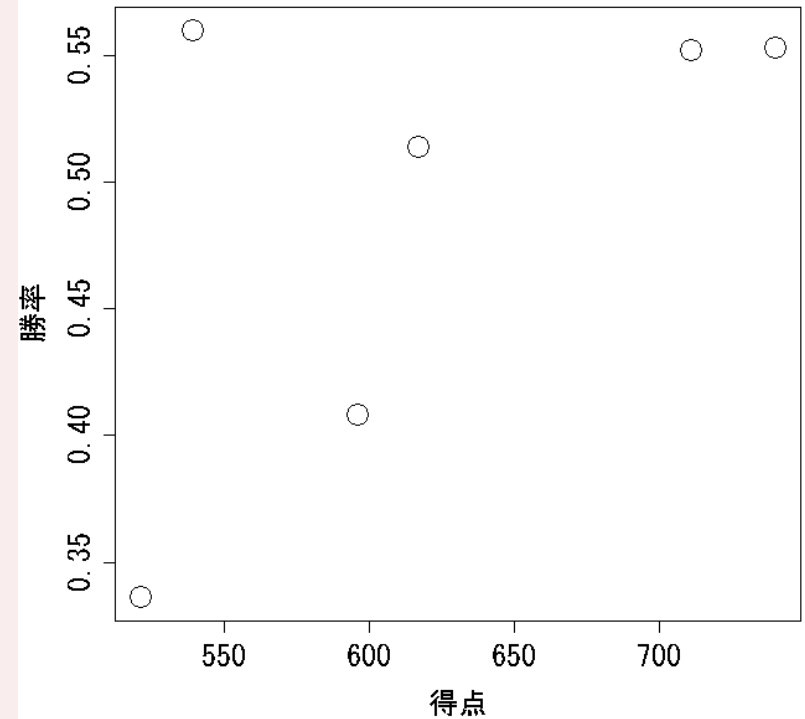
	勝率	打率	得点	本塁打	盗塁	失点	被本塁打
中日	0.560	0.259	539	119	53	521	121
阪神	0.553	0.290	740	173	71	640	138
巨人	0.552	0.266	711	226	96	617	140
ヤクルト	0.514	0.268	617	124	66	621	146
広島	0.408	0.263	596	104	119	737	171
横浜	0.336	0.255	521	117	59	743	176

- 優勝の中日は、失点・被本塁打が最少だが盗塁が最少であり、打率、得点は2番目に少ない。
- 最下位の横浜は、本塁打、盗塁以外のすべての指標で最悪。
- 阪神は打率、得点などで最高だが2位である。

勝率との散布図

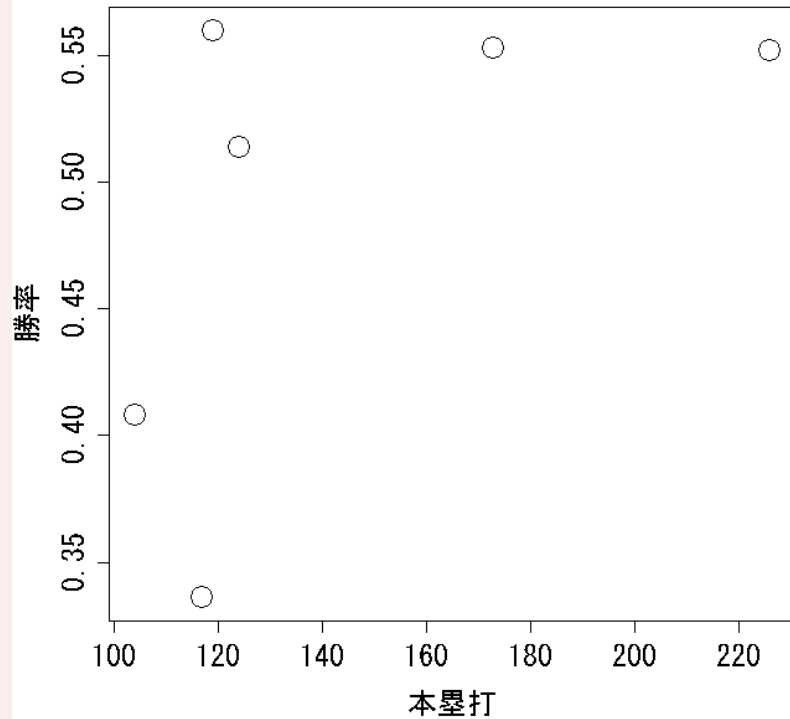


打率, $r = 0.53$

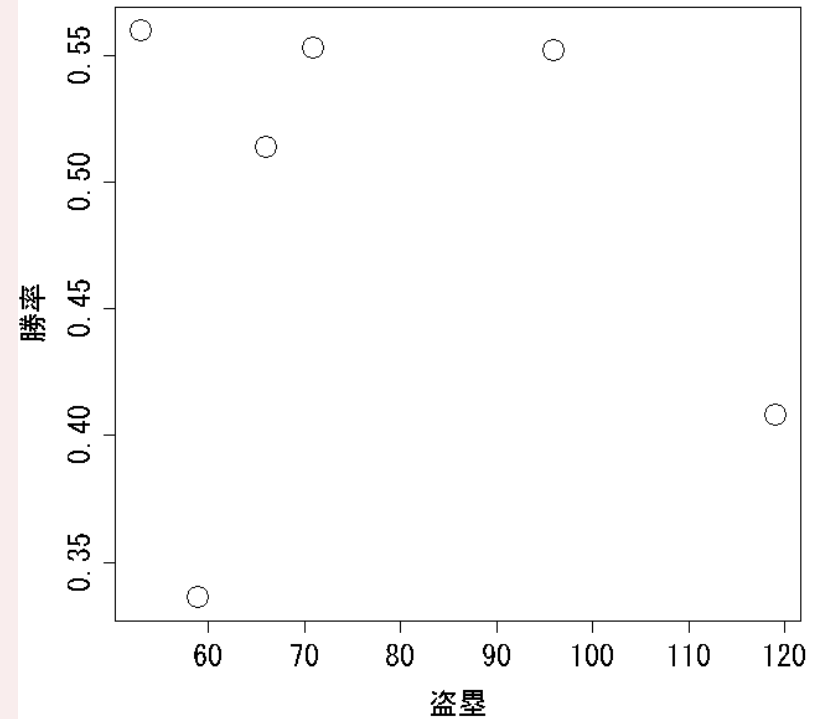


得点, $r = 0.59$

勝率との散布図

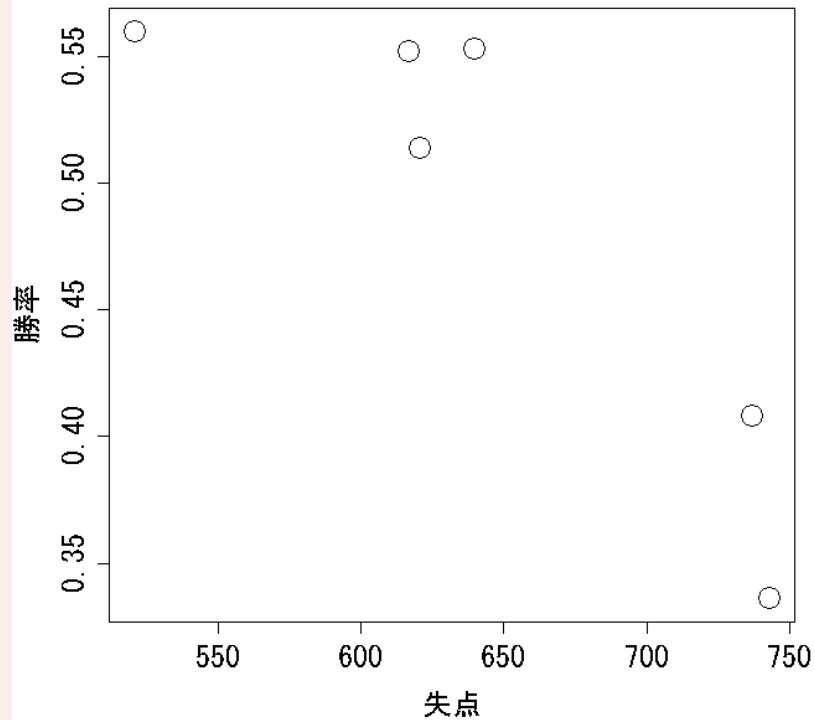


本塁打, $r = 0.55$

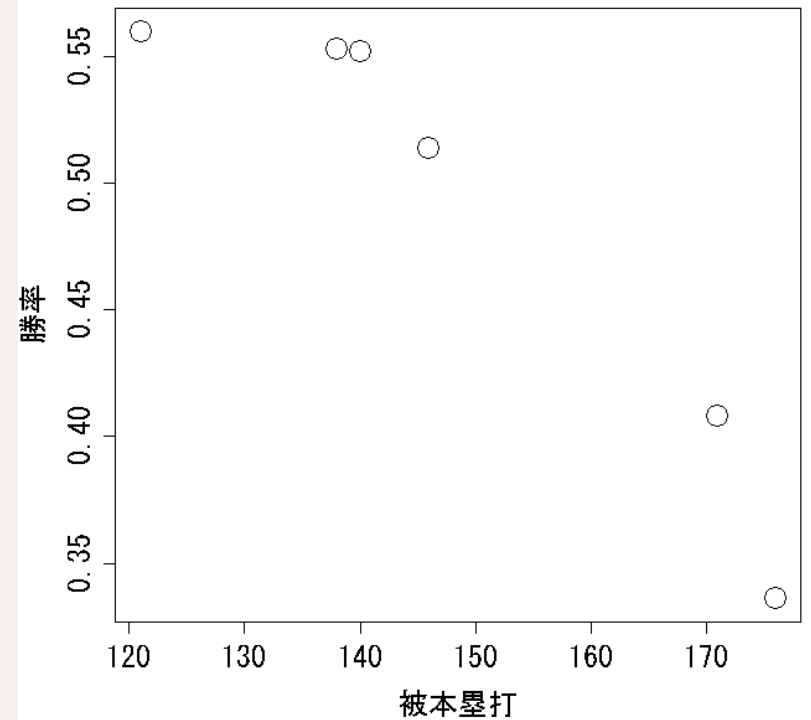


盗塁, $r = -0.15$

勝率との散布図



失点, $r = -0.87$



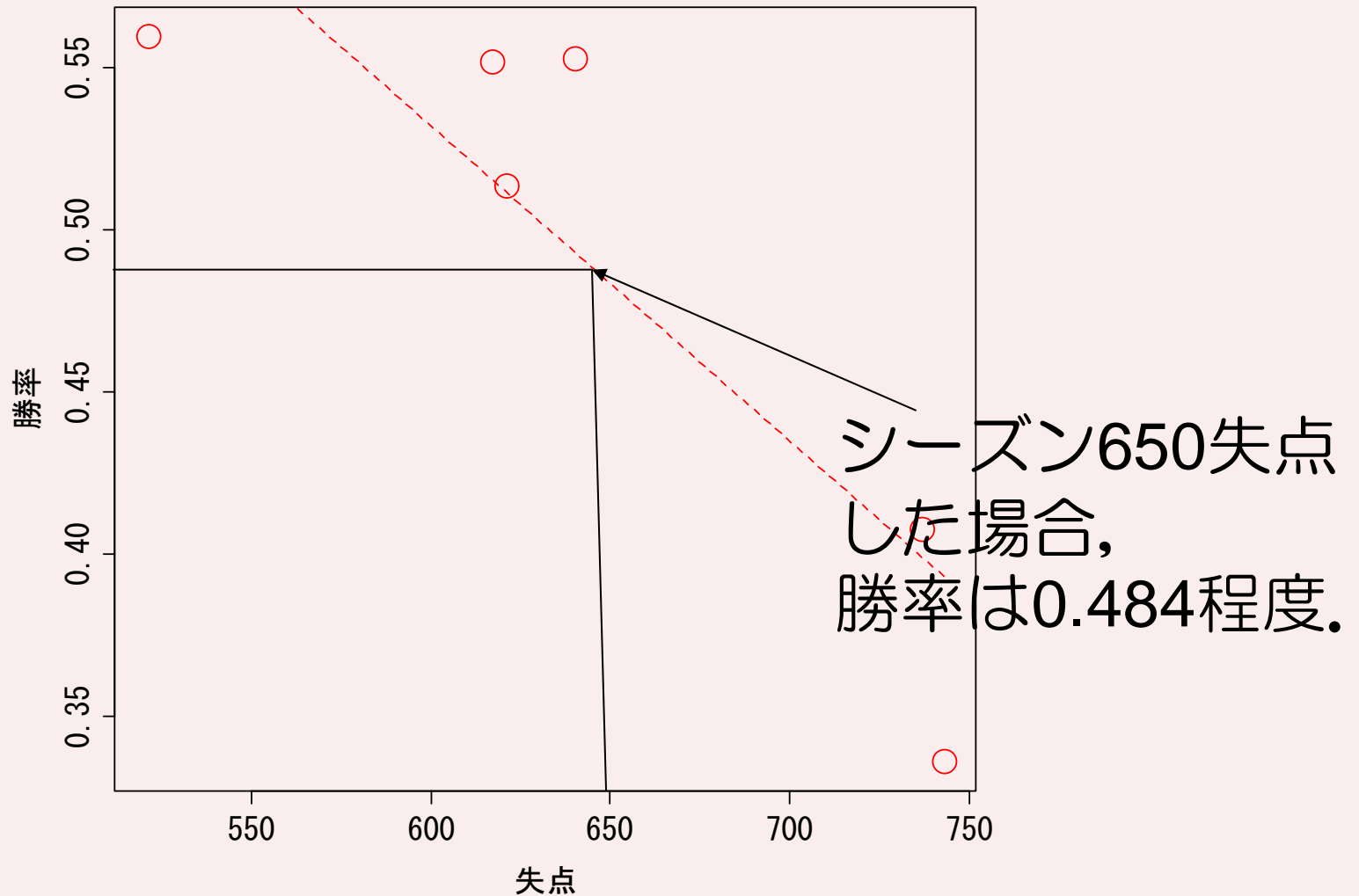
被本塁打, $r = -0.94$

勝率との相関

	勝率	打率	得点	本塁打	盗塁	失点	被本塁打
勝率	1						
打率	0.526665	1					
得点	0.593345	0.840853	1				
本塁打	0.554794	0.435002	0.800973	1			
盗塁	-0.15357	0.046678	0.334767	0.171424	1		
失点	-0.86975	-0.12904	-0.13322	-0.26596	0.459415	1	
被本塁打	-0.94209	-0.33726	-0.32737	-0.38831	0.391956	0.972351	1

- 打率，得点，本塁打とは中程度の正の相関。
- 盗塁とはほぼ無相関。
- 失点，被本塁打とは高い負の相関。
- こういったデータから，勝率を予想，説明することはできるだろうか？
- 勝率と関係性の高いデータ（失点，得点など）から，予測，説明が可能となるのではないか。

勝率の予測



单回归分析

回帰分析

- ある変数（**予測変数**・**説明変数**）の値から他の変数（**目的変数**・**基準変数**）の値を予測・説明する分析を**回帰分析**（regression analysis）と呼ぶ。特に，一次関数を用いて予測・説明を行う場合，線形回帰と呼ぶ。また，予測変数が1つの場合を**単回帰分析**，複数ある場合を**重回帰分析**と呼ぶ。
 - 「**入学時の成績**」から「**卒業時の成績**」を予測・説明。
 - 「**失点**」から「**勝率**」を予測・説明。
→単回帰分析
 - 「**得点**」と「**失点**」から，「**勝率**」を予測・説明。
→重回帰分析

単回帰分析

- 単回帰分析では、以下のようなモデル式をまず考える。

$$y_i = \alpha x_i + \beta + e_i$$

- この式は、前回説明した一元配置の分散分析と記号は少々異なるが、ほぼ同じ形である。分散分析と回帰分析は、線形モデルとして、数学的性質を同じくする。
- ここで、 y_i は**目的変数**（勝率など）、 x_i は**予測変数**（得点など）であり、 α, β はそれぞれ一次関数における**傾き**、**切片**であり、 e_i は**誤差**である。

単回帰分析

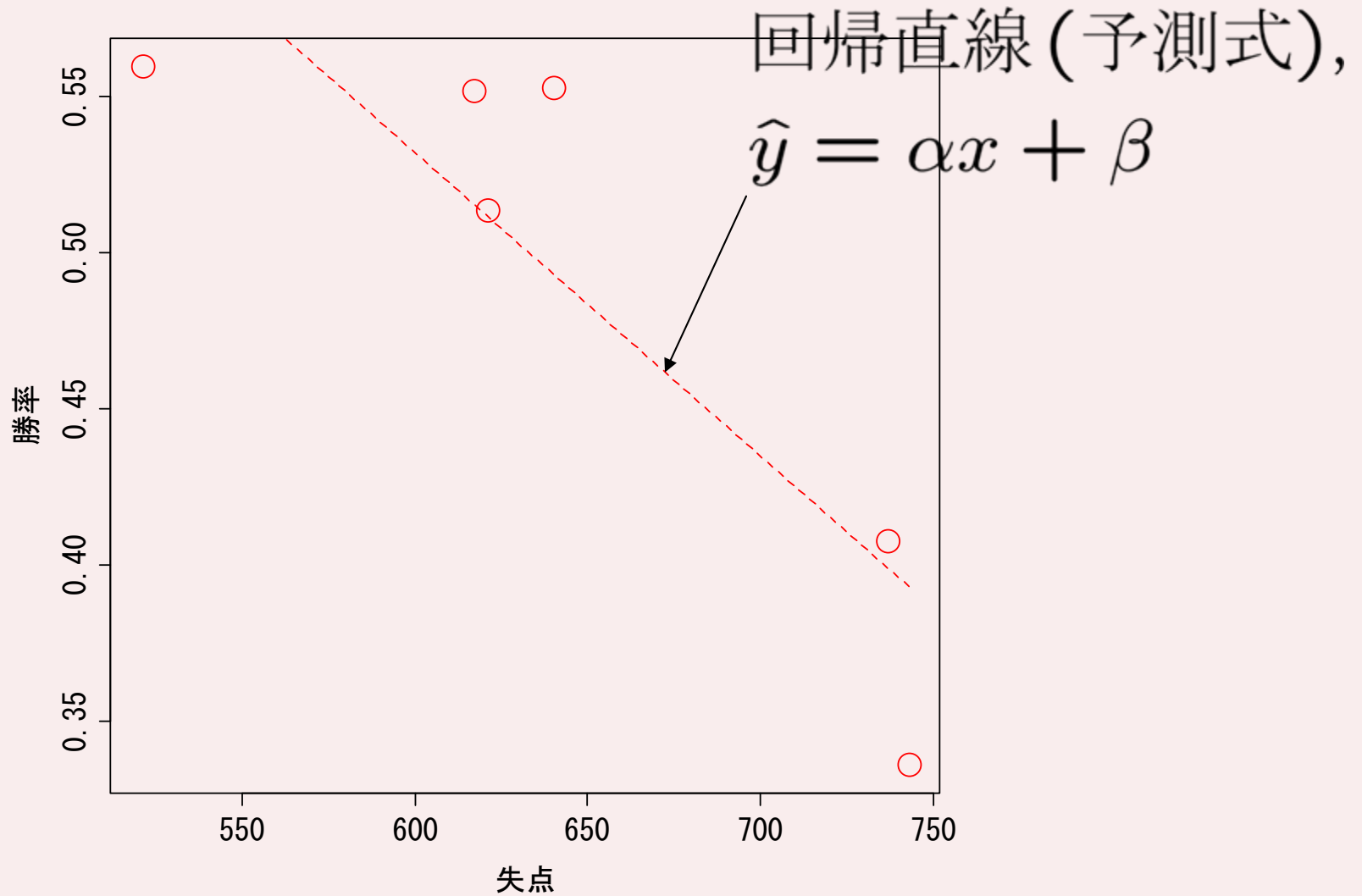
$$y_i = \alpha x_i + \beta + e_i$$

予測部分

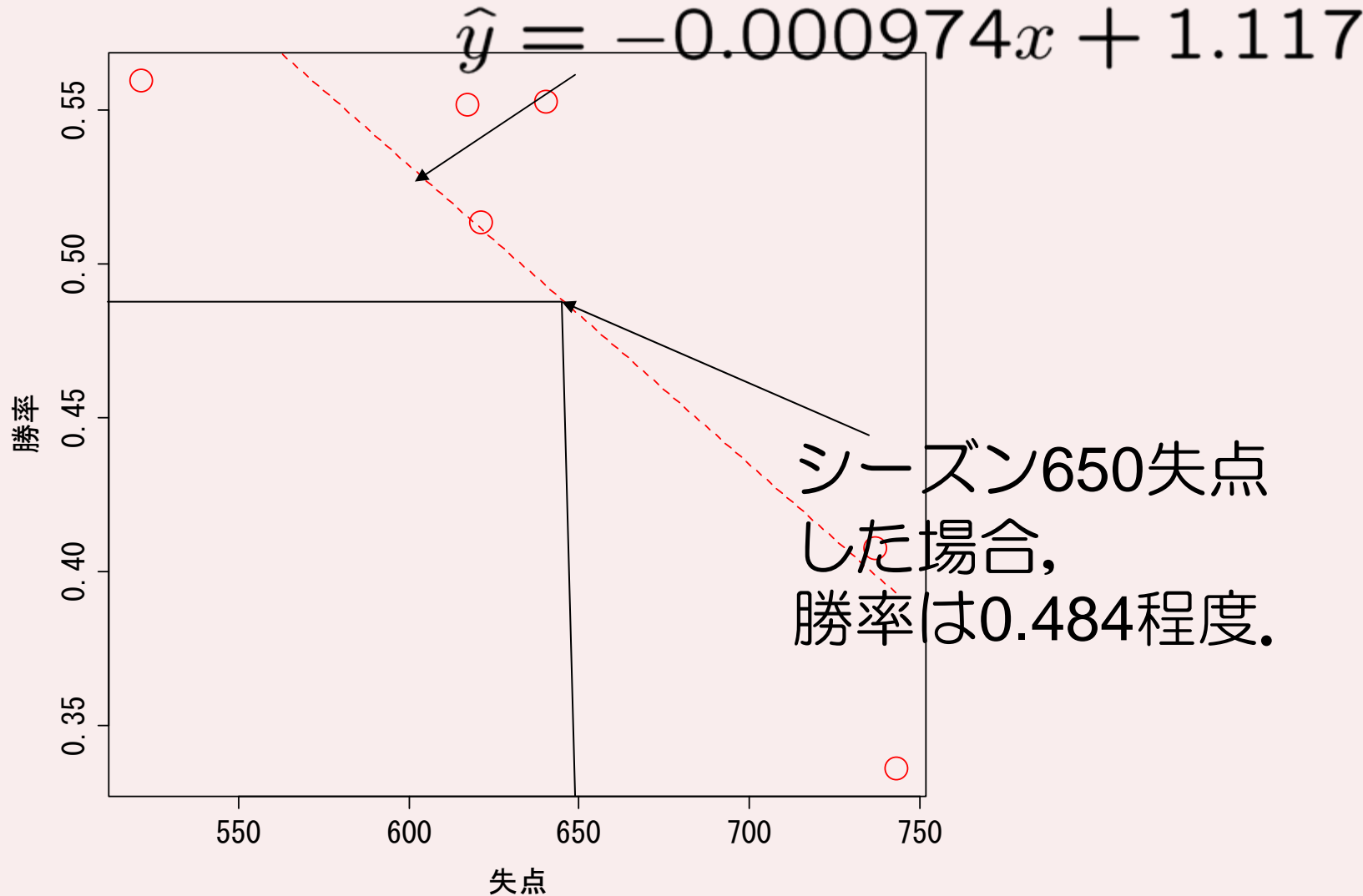
誤差

- この式は、 x_i を利用して y_i を、 $\hat{y}_i = \alpha x_i + \beta$ という形で予測，説明する事を意味している。
- しかし，予測には誤差が必ず含まれる（同じチーム得点数であったとしても，同じ勝率になるとは限らない）その誤差が e_i で表現される。
- この α, β の値をデータから推定することで，予測式を完成させる。

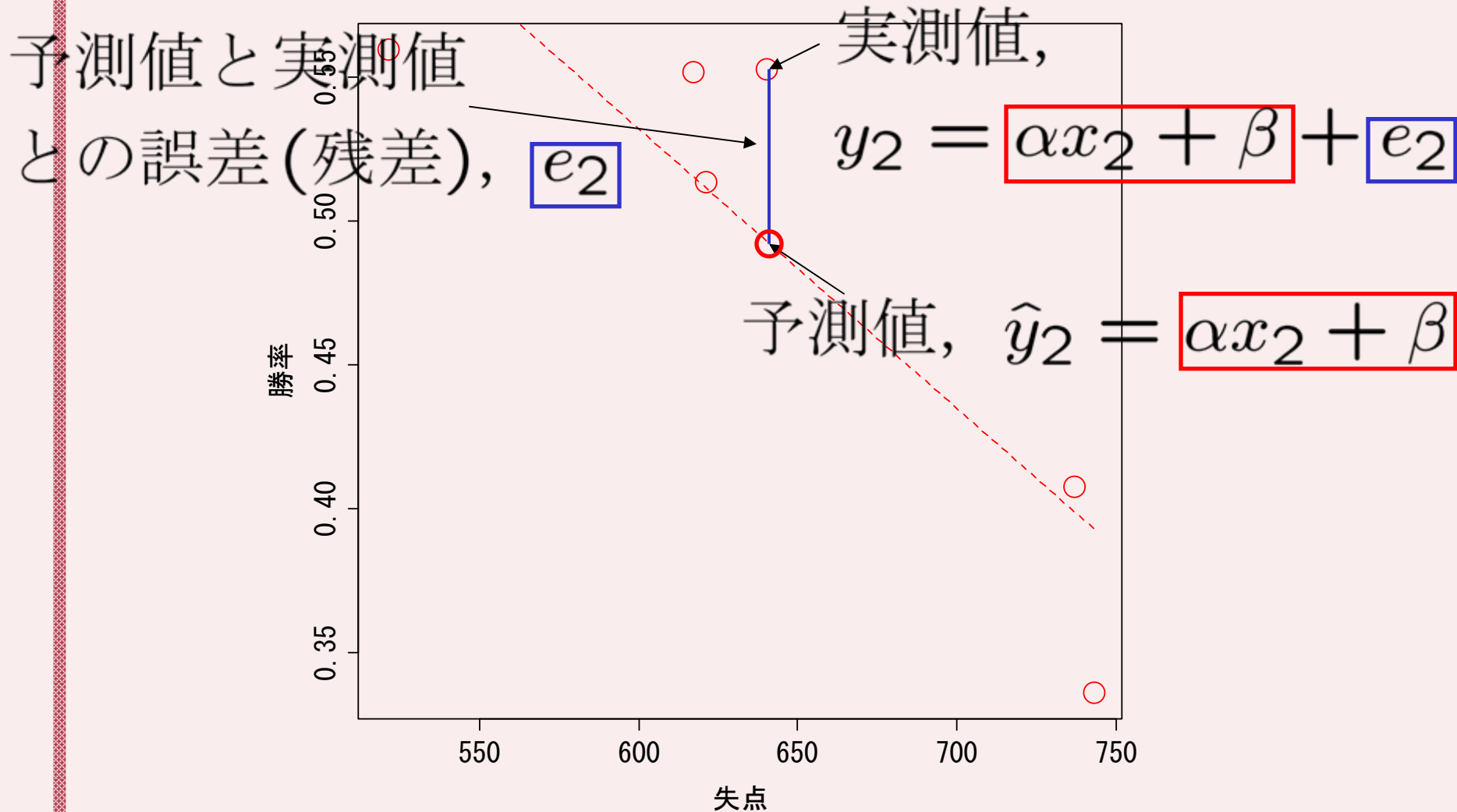
单回归分析



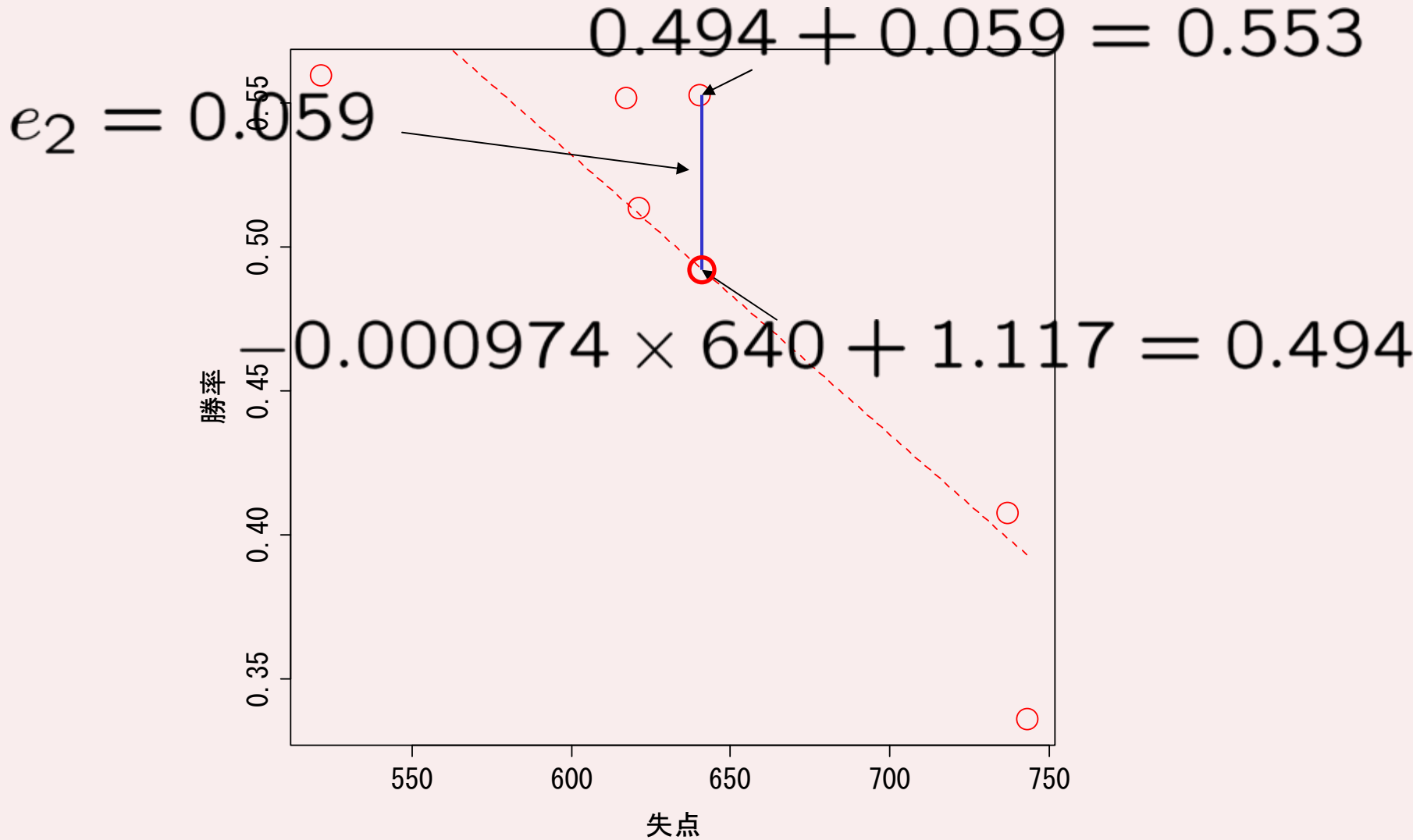
単回帰分析



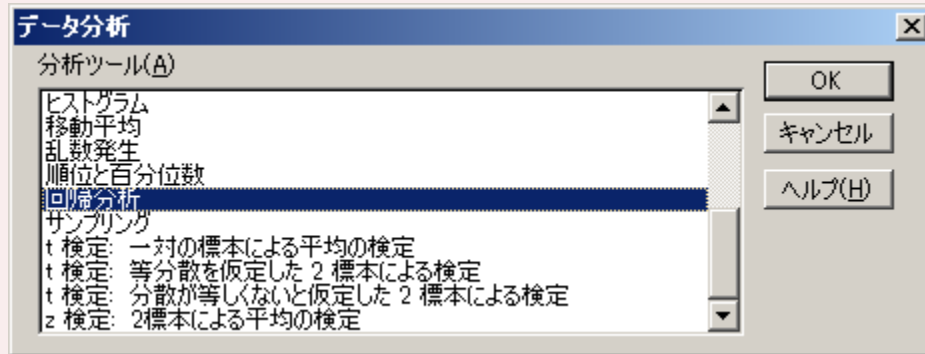
単回帰分析



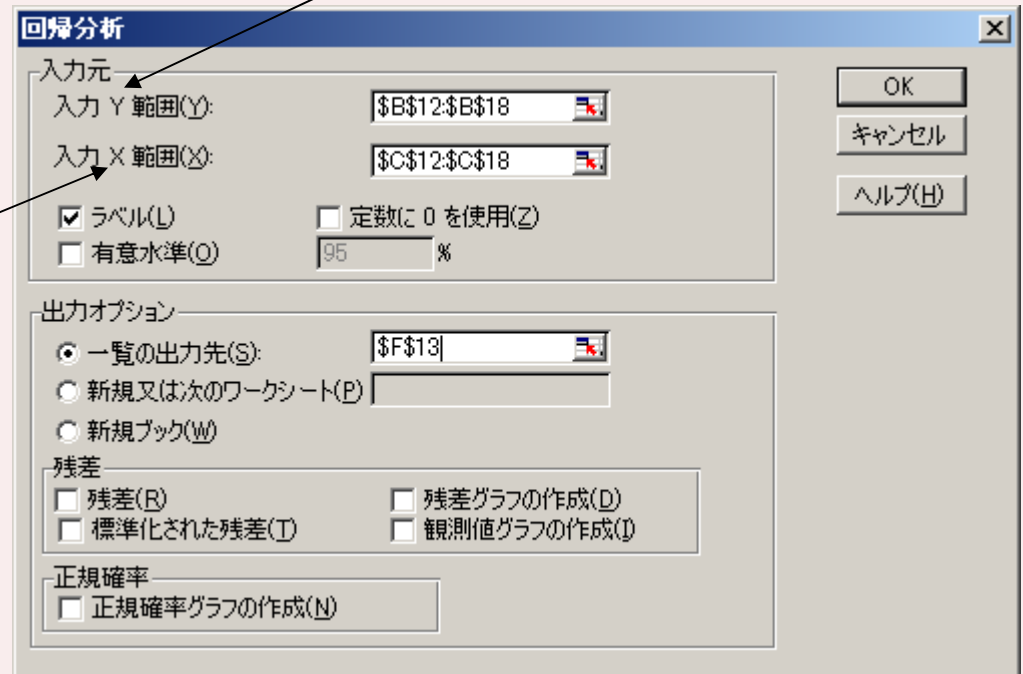
单回归分析



Excelによる単回帰分析



目的変数



予測変数

単回帰分析の出力

回帰統計	
重相関 R	0.869745234
重決定 R2	0.756456773
補正 R2	0.695570966
標準誤差	0.051570755
観測数	6

予測値と実測値の間の相関係数

重相関Rの2乗。2乗しているのので、正の値。データの変動の76%を予測式で説明できていると解釈する。決定係数と呼ばれる。高い方が望ましい。

予測値と実測値との間の標準誤差関数“STEYX”のヘルプなど参照。

単回帰分析の出力

分散分析表					
	自由度	変動	分散	測された分散	有意 F
回帰	1	0.033043	0.033043	12.424189	0.024344
残差	4	0.010638	0.00266		
合計	5	0.043681			

分散分析と見方はほぼ一緒。有意Fが5%以下ならば、この回帰式は5%で有意と考える。つまり、変数「失点」の多寡が、変数「勝率」の値と関係していると考ええる。

単回帰分析の出力

	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	1.116991421	0.17992	6.208264	0.00342492	0.617453	1.61653	0.617453	1.61653
失点	-0.00097421	0.000276	-3.5248	0.02434448	-0.00174	-0.00021	-0.00174	-0.00021

係数の推定値

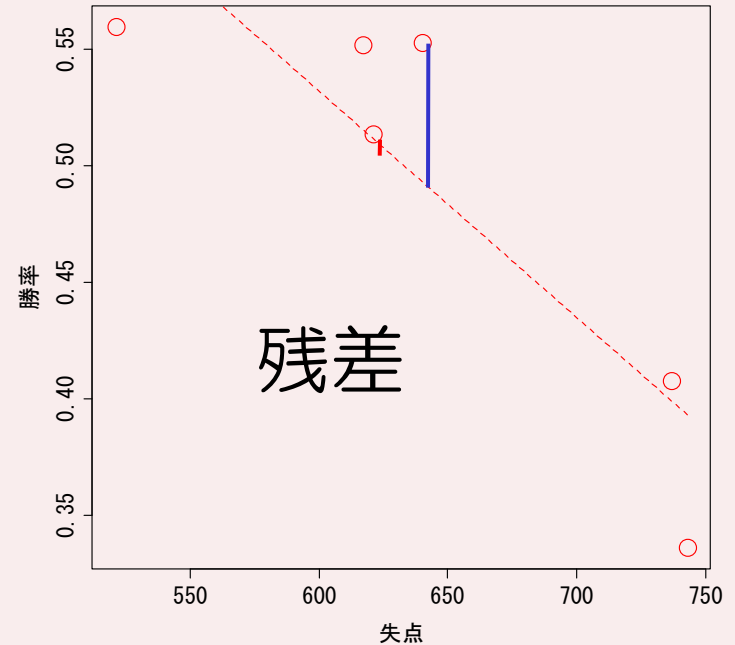
$$\hat{y} = -0.000974x + 1.117$$

係数の信頼区間

切片，傾きが0という帰無仮説に関する検定。
傾きに関する帰無仮説が棄却されれば，傾きが
0ではない，即ち変数「得点」は「勝率」の
予測に意味があると考ええる。

単回帰分析の出力

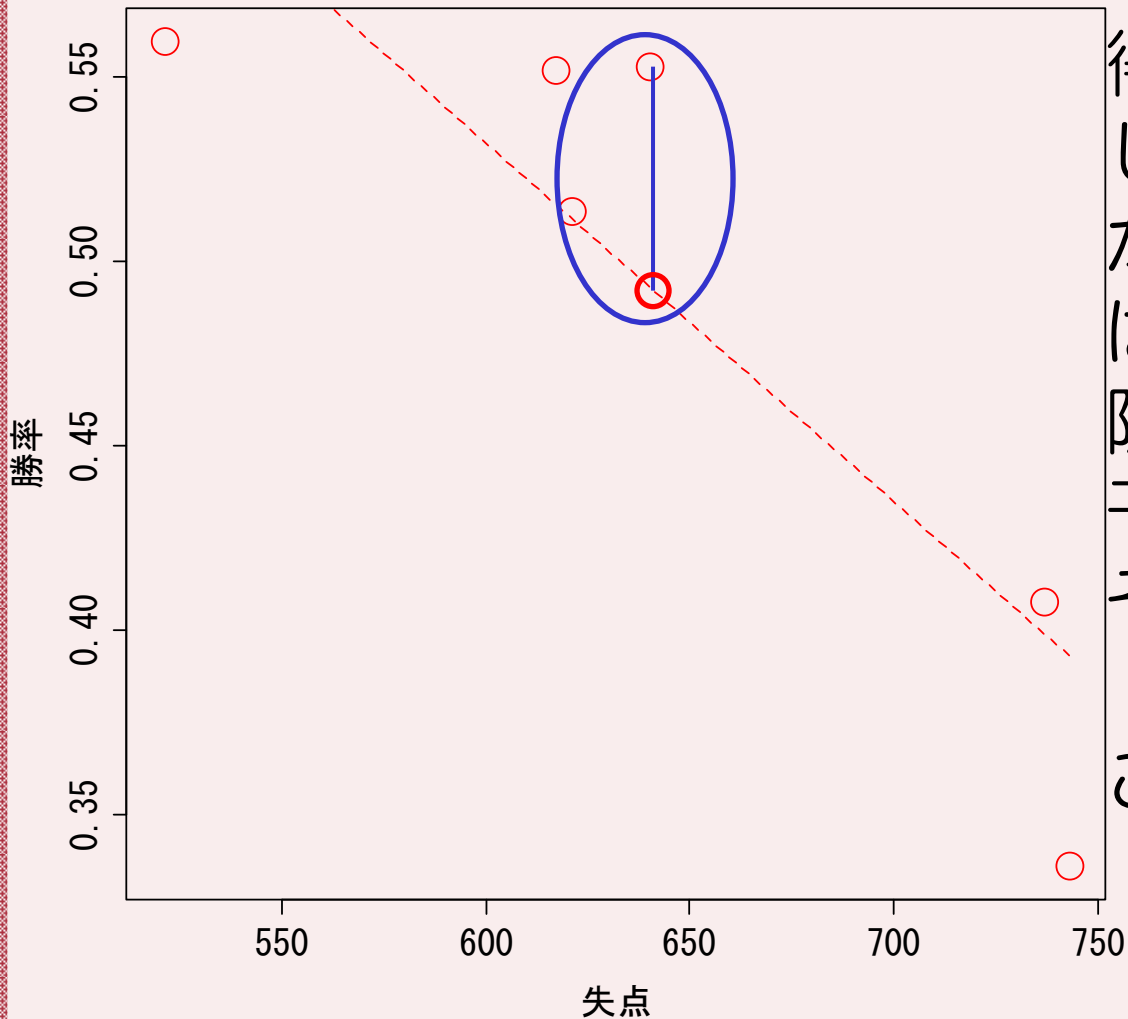
残差出力		
観測値	予測値：勝率	残差
1	0.609429631	-0.04943
2	0.493499011	0.059501
3	0.51590577	0.036094
4	0.512008942	0.001991
5	0.399000943	0.008999
6	0.393155702	-0.05716



予測値と実測値との差が**残差**。概ね0.06以下の範囲に収まっている。4番目、**ヤクルト**の残差が一番小さく2番目、**阪神**の残差が一番大きい。

重回帰分析

重回帰分析



得点を予測変数とした単回帰分析でかなりの部分勝率は説明できたが、阪神の勝率には若干の誤差があった。そこで、予測変数「得点」を加えることを考える。

重回帰分析

- 重回帰分析では、以下のようなモデル式をまず考える。

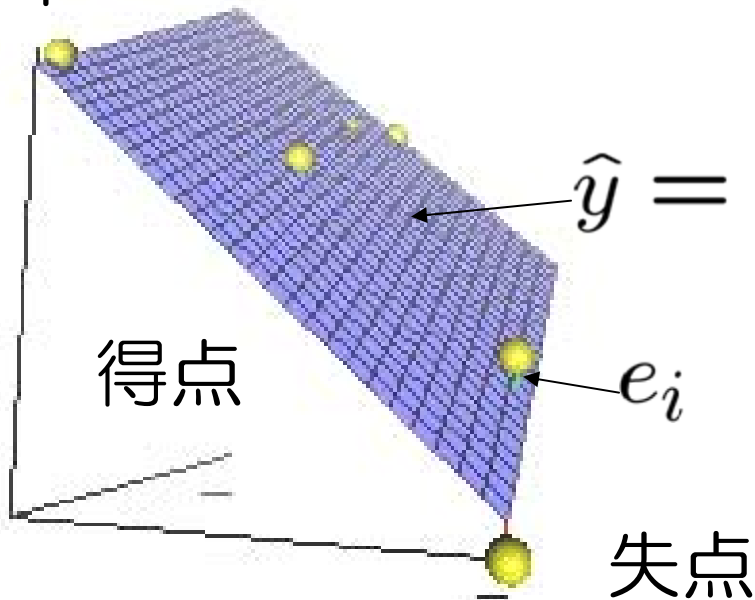
$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \beta + e_i$$

- 目的変数，切片の部分に違いはない．予測部分と誤差との関係も単回帰分析と同じ．
- 予測変数が2つになったのに伴い， α, x の部分が変化している． x_{1i} は，1つ目の予測変数の*i*番目の値（「中日」の「失点数」など）， α_1 はそれにかかる傾き， x_{2i} は2つ目の予測変数の*i*番目の値， α_2 はそれにかかる傾きである．
- 予測変数が3つ以上でも，同じようにモデル化．

重回帰分析

予測変数が2つになったことで、回帰直線は回帰平面となる。

勝率

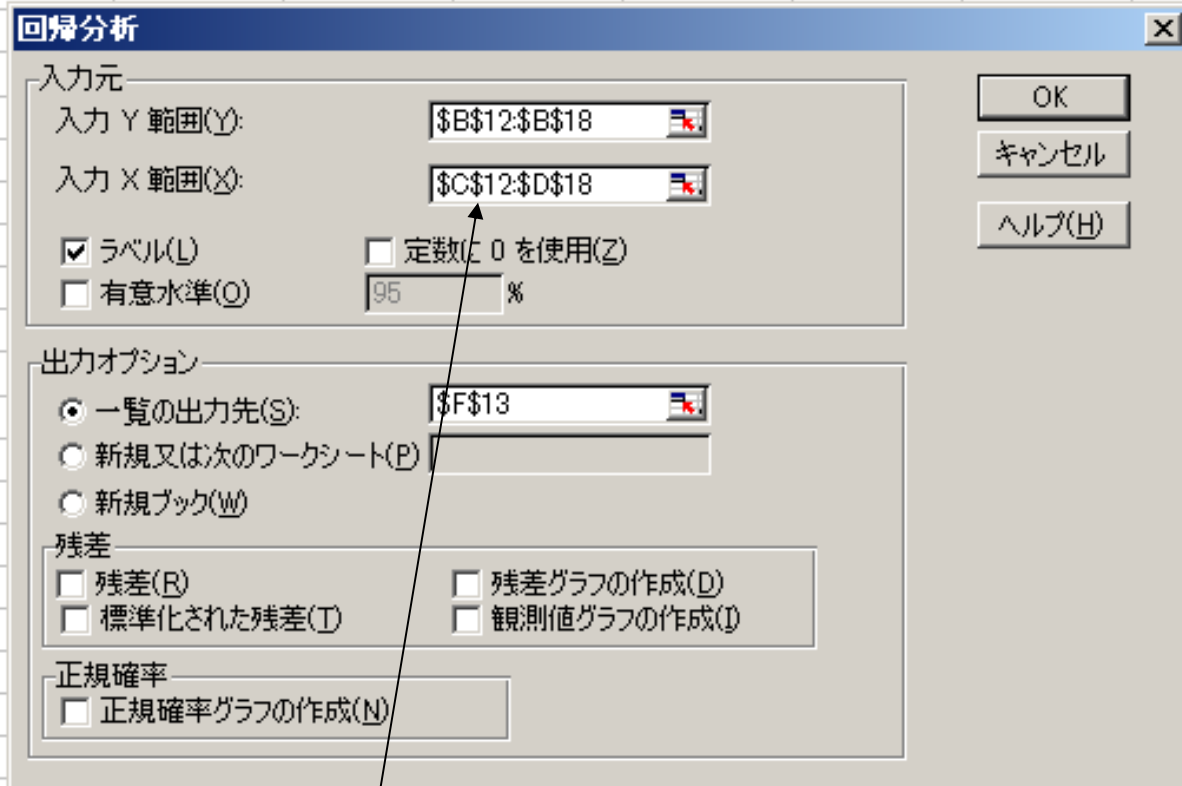


$$\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + \beta$$

失点

Excelによる重回帰分析

勝率	失点	得点
0.659	521	539
0.566	640	740
0.497	617	711
0.479	621	617
0.464	737	596
0.354	743	521



The image shows the 'Regression Analysis' dialog box in Excel. The 'Input Range' (入力元) section has 'Input Y Range' (入力 Y 範囲(Y)) set to '\$B\$12:\$B\$18' and 'Input X Range' (入力 X 範囲(X)) set to '\$C\$12:\$D\$18'. The 'Labels' (ラベル(L)) checkbox is checked. The 'Constant is zero' (定数(0)を使用(Z)) checkbox is unchecked. The 'Confidence Level' (有意水準(O)) is set to 95%. The 'Output Options' (出力オプション) section has 'Output to' (一覧の出力先(S)) set to '\$F\$13'. The 'Residuals' (残差) section has 'Residuals' (残差(R)), 'Standardized Residuals' (標準化された残差(I)), 'Residuals Plot' (残差グラフの作成(D)), and 'Residuals Plot' (観測値グラフの作成(I)) all unchecked. The 'Normal Distribution' (正規確率) section has 'Normal Distribution Plot' (正規確率グラフの作成(N)) unchecked. An arrow points from the text below to the 'Input X Range' field.

予測変数が2つに

Excelによる重回帰分析

回帰統計	
重相関 R	0.869745234
重決定 R ²	0.756456773
補正 R ²	0.695570966
標準誤差	0.051570755
観測数	6

単回帰分析

1行目を見ると、予測値と実測値の相関は向上している。それに伴い、決定係数の値も向上している。しかし、決定係数は、予測変数が多ければ多いほどいい値が出る傾向がある。それを修正したのが補正R²の部分である。補正R²は若干悪化しており、ここから、予測変数を増やしてもあまり意味がないと解釈できる。

回帰統計	
重相関 R	0.873416
重決定 R ²	0.762855
補正 R ²	0.604758
標準誤差	0.064506
観測数	6

重回帰分析

Excelによる重回帰分析

	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	1.103187	0.323922	3.405715	0.042283	0.072322	2.134052	0.072322	2.134052
失点	-0.00105	0.000349	-3.00455	0.05746	-0.00216	6.21E-05	-0.00216	6.21E-05
得点	0.000125	0.000327	0.382105	0.727845	-0.00092	0.001165	-0.00092	0.001165

係数の推定値

係数の信頼区間

$$\hat{y} = -0.00105x_1 + 0.000125x_2 + 1.103$$

P値から、「得点」の係数が0であるという帰無仮説は棄却できない。つまり、「得点」を予測変数として使うことが合理的であるという証拠はないことになる。

多重共線性問題

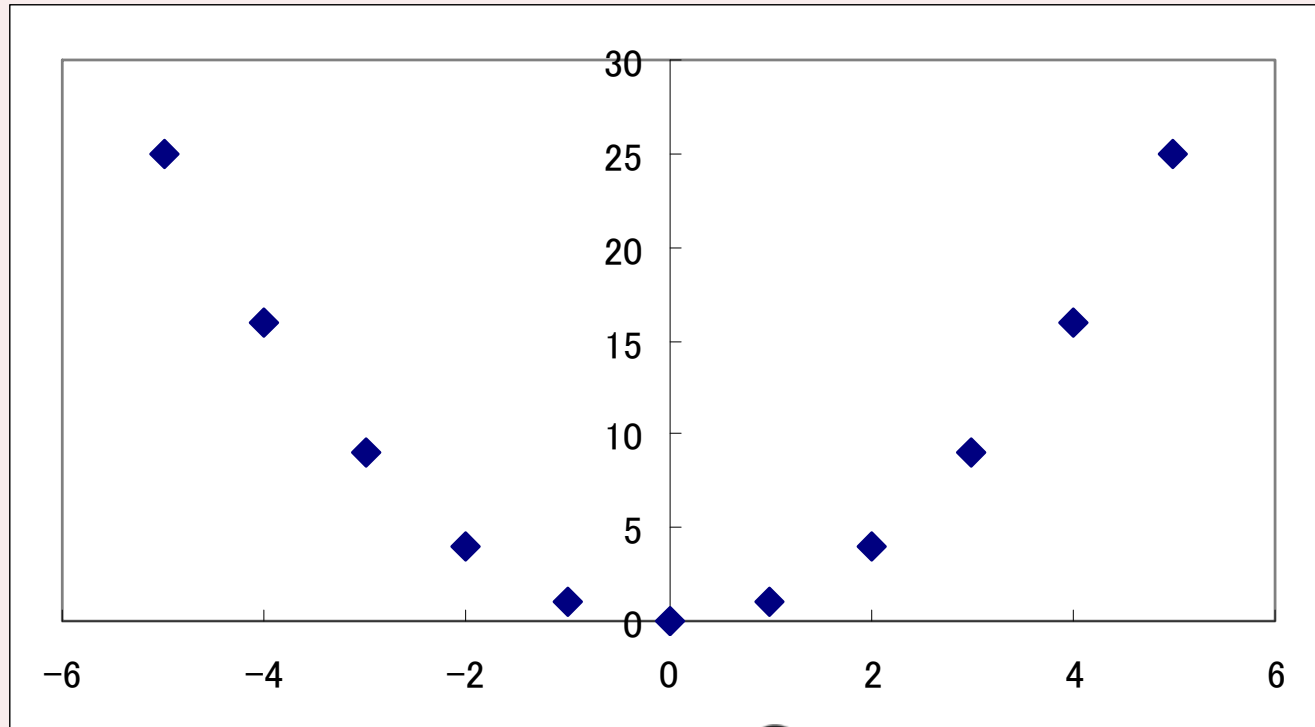
	勝率	打率	得点	本塁打	盗塁	失点	被本塁打
勝率	1						
打率	0.526665	1					
得点	0.593345	0.840853	1				
本塁打	0.554794	0.435002	0.800973	1			
盗塁	-0.15357	0.046678	0.334767	0.171424	1		
失点	-0.86975	-0.12904	-0.13322	-0.26596	0.459415	1	
被本塁打	-0.94209	-0.33726	-0.32737	-0.38831	0.391956	0.972351	1

- 変数「得点」を用いたが、あまり意味はなかった。勝率との相関を見ると、被本塁打の方が相関は高かったなので、そちらを使った方が良かったのでは？
- それはダメ。何故なら、失点と被本塁打、予測変数間の相関が高すぎるから。

多重共線性問題

- 重回帰分析において、予測変数間の相関があまりに高すぎる場合、**多重共線性問題**という問題が起こる可能性があることが知られている。これは、 α, β 等の推定値が不安定になる（僅かなデータの変化で大きく変わってしまう）問題である。これを避けるためには予測変数間で相関の高い変数はあまり使わないこと。
- そもそも、「失点」の情報と「被本塁打」の情報はかなり重複しているので、「被本塁打」の情報を加えたところで、推定精度はあまり向上しない。

回帰分析の問題点



- このデータには、 $y = x^2$ という明確な関連性があるが、相関係数を算出すると $\rho = 0$ となる。こういったデータには、これまでの回帰分析は役に立たない。

回帰分析の問題点

回帰統計									
重相関 R	0								
重決定 R2	0								
補正 R2	-0.11111111								
標準誤差	9.763879011								
観測数	11								
分散分析表									
	自由度	変動	分散	測された分散	有意 F				
回帰	1	0	0	0	1				
残差	9	858	95.33333						
合計	10	858							
	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	10	2.94392	3.396831	0.00791296	3.34039	16.65961	3.34039	16.65961	
X 値 1	0	0.930949	0	1	-2.10595	2.105954	-2.10595	2.105954	

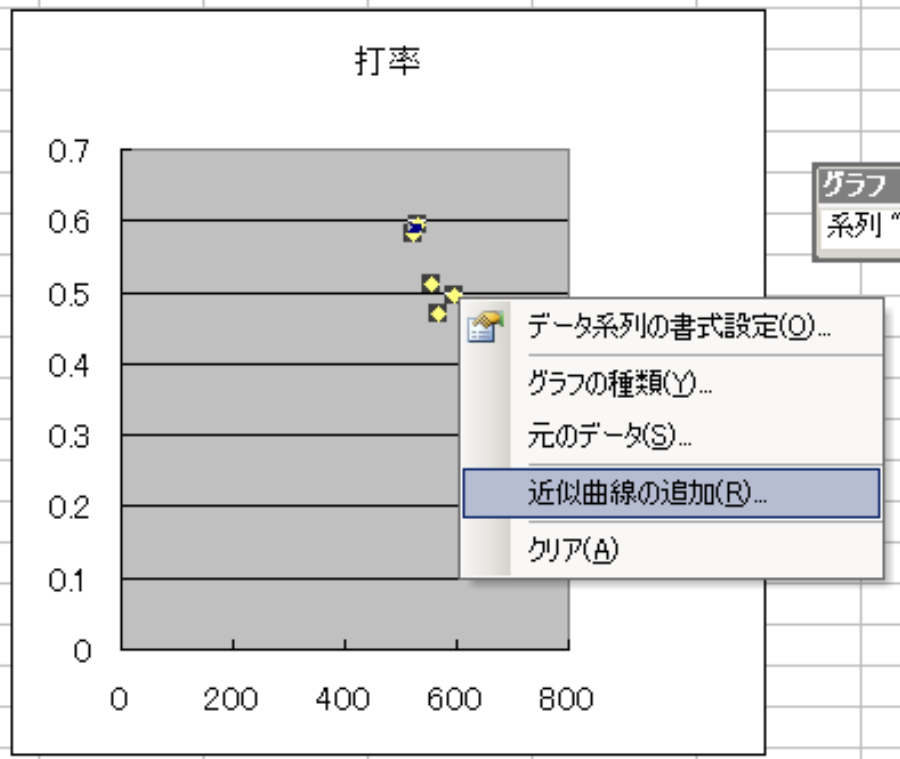
- 全く予測ができていない。これは、これまで述べた回帰分析が、相関係数同様直線的な関係しか表現できないためである。

実習

- 配布したアデレード大学のデータを用い、手の幅から身長を予測するモデルを作成せよ。また、Excelの出力から予測式を作成し、手の幅が17cmであるときの身長の予測値を求めよ。
- 手の幅と心拍数から身長を予測するモデルを作成せよ。また、Excelの出力から予測式を作成し、手の幅が17cm、心拍数が90であるときの身長の予測値を求めよ。
- このデータについて、単回帰分析と重回帰分析ではどちらが適切だったか検討せよ。

補足資料

散布図から回帰分析



散布図から回帰分析

近似曲線の追加

種類 オプション

近似曲線名
 自動(A): 線形(系列1)
 指定(C):

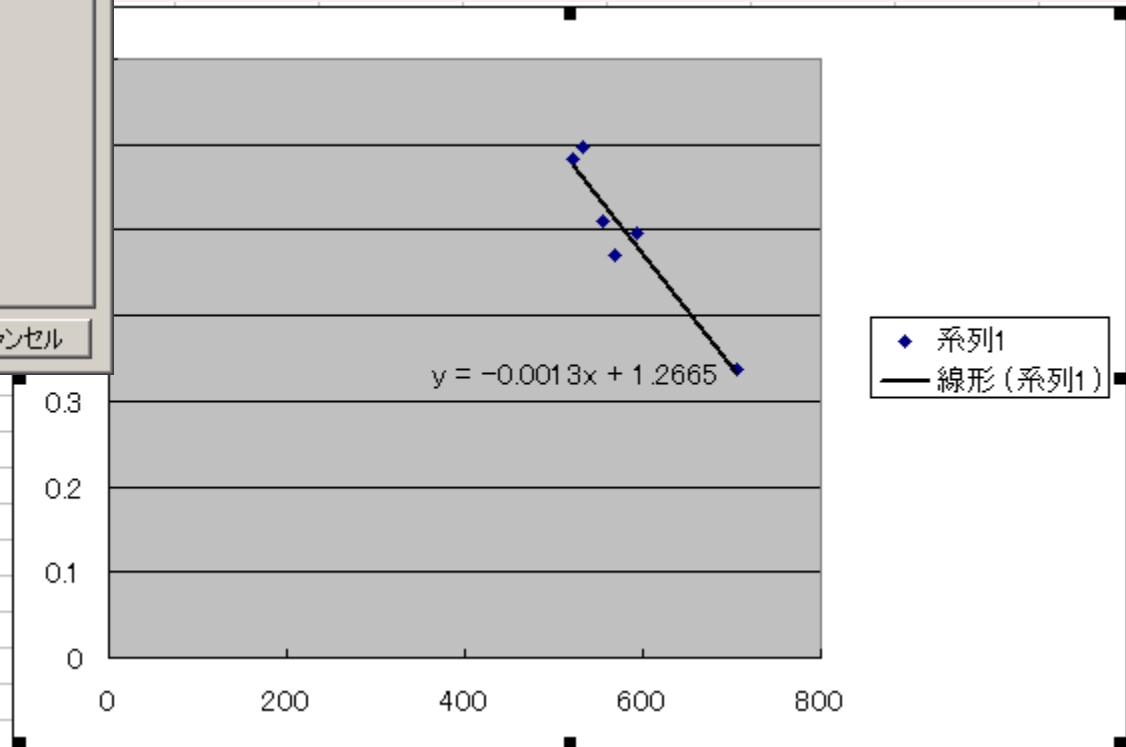
予測
前方補外(F): 0 単位
後方補外(B): 0 単位

切片(S) = 0

グラフに数式を表示する(E)

グラフに R-2 乗値を表示する(R)

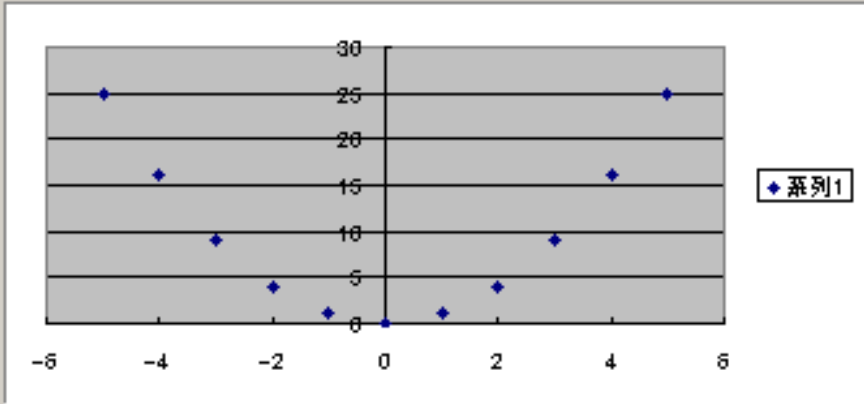
OK キャンセル



多項式による当てはめ

元のデータ

データ範囲 系列



系列(S)

名前(N):	X の値(X):	Y の値(Y):
系列1	=Sheet2!\$A\$1:\$A\$11	=Sheet2!\$B\$1:\$B\$11

追加(A) 削除(B)

キャンセル < 戻る(B) 次へ(N) > 完了(F)

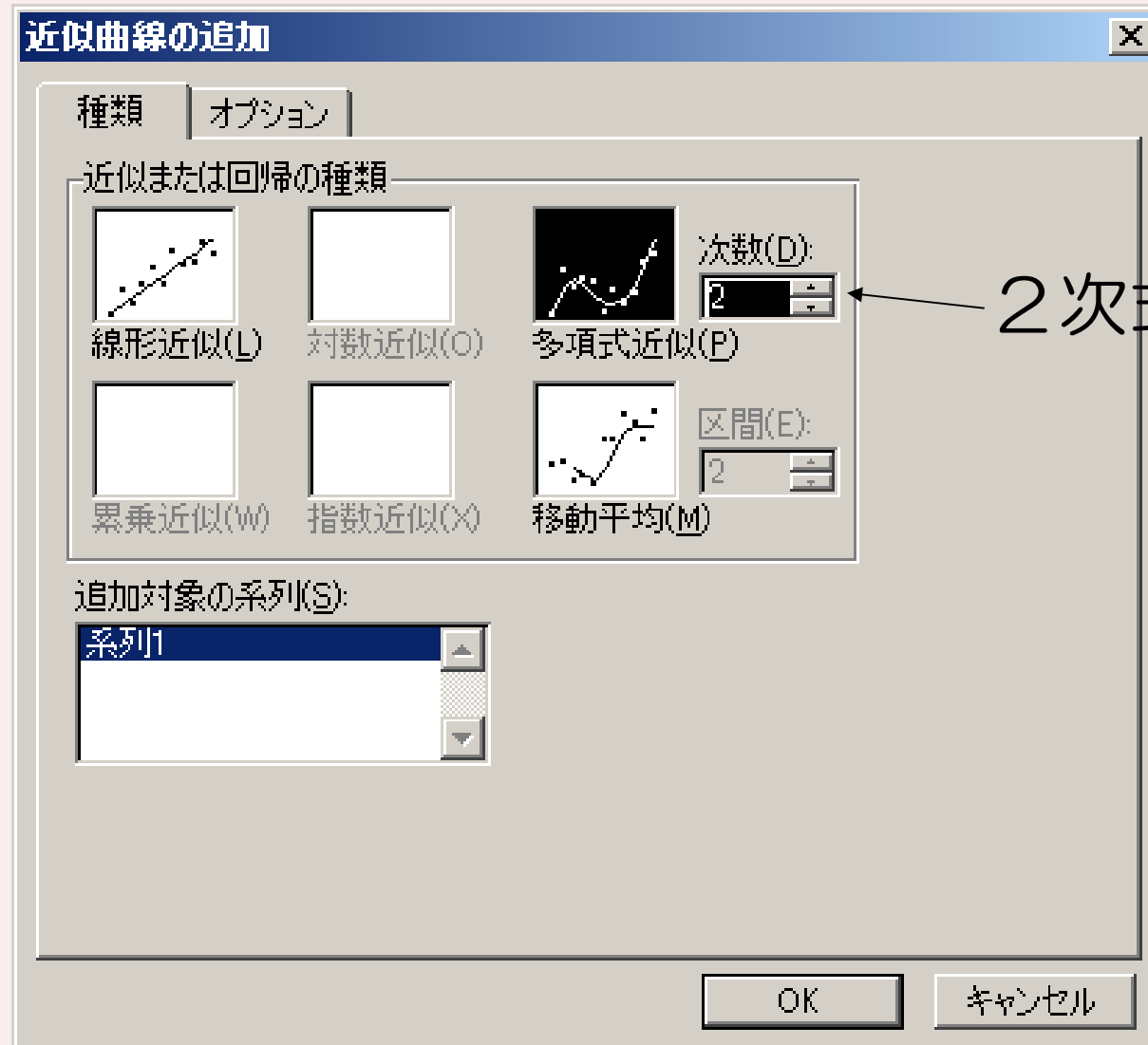
こちらに予測変数

こちらに目的変数

多項式による当てはめ



多項式による当てはめ



2次式で当てはめる

多項式による当てはめ

近似曲線の追加

種類 オプション

近似曲線名

自動(A): 多項式 (系列1)

指定(C):

予測

前方補外(F): 0 単位

後方補外(B): 0 単位

切片(S) = 0

グラフに数式を表示する(E)

グラフに R-2 乗値を表示する(R)

OK キャンセル

予測式を出力

多項式による当てはめ

実質 $y = x^2$

