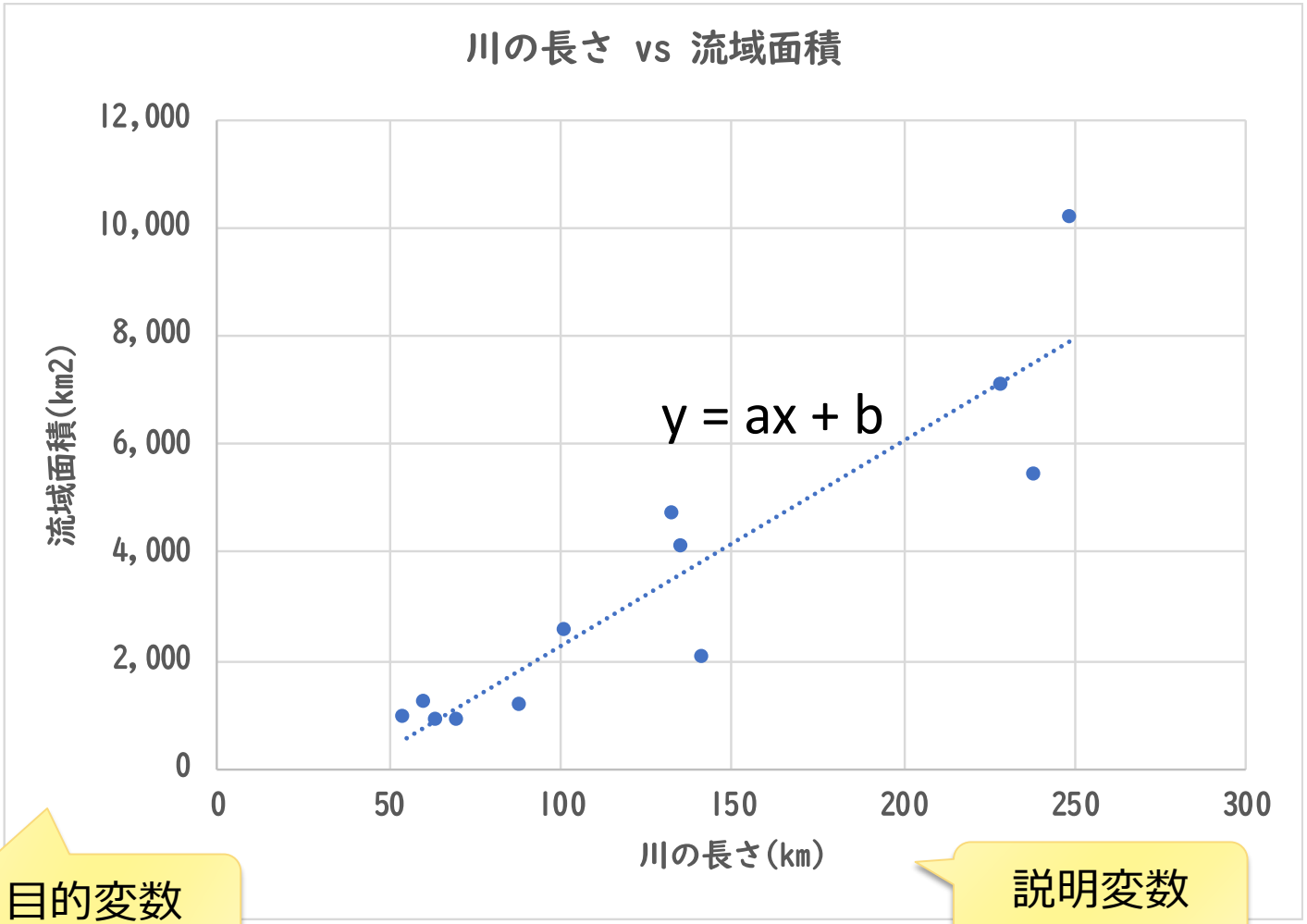
データサイエンス入門B
第11回　回帰分析

髙田　美樹

法政大学

目次

回帰分析とは



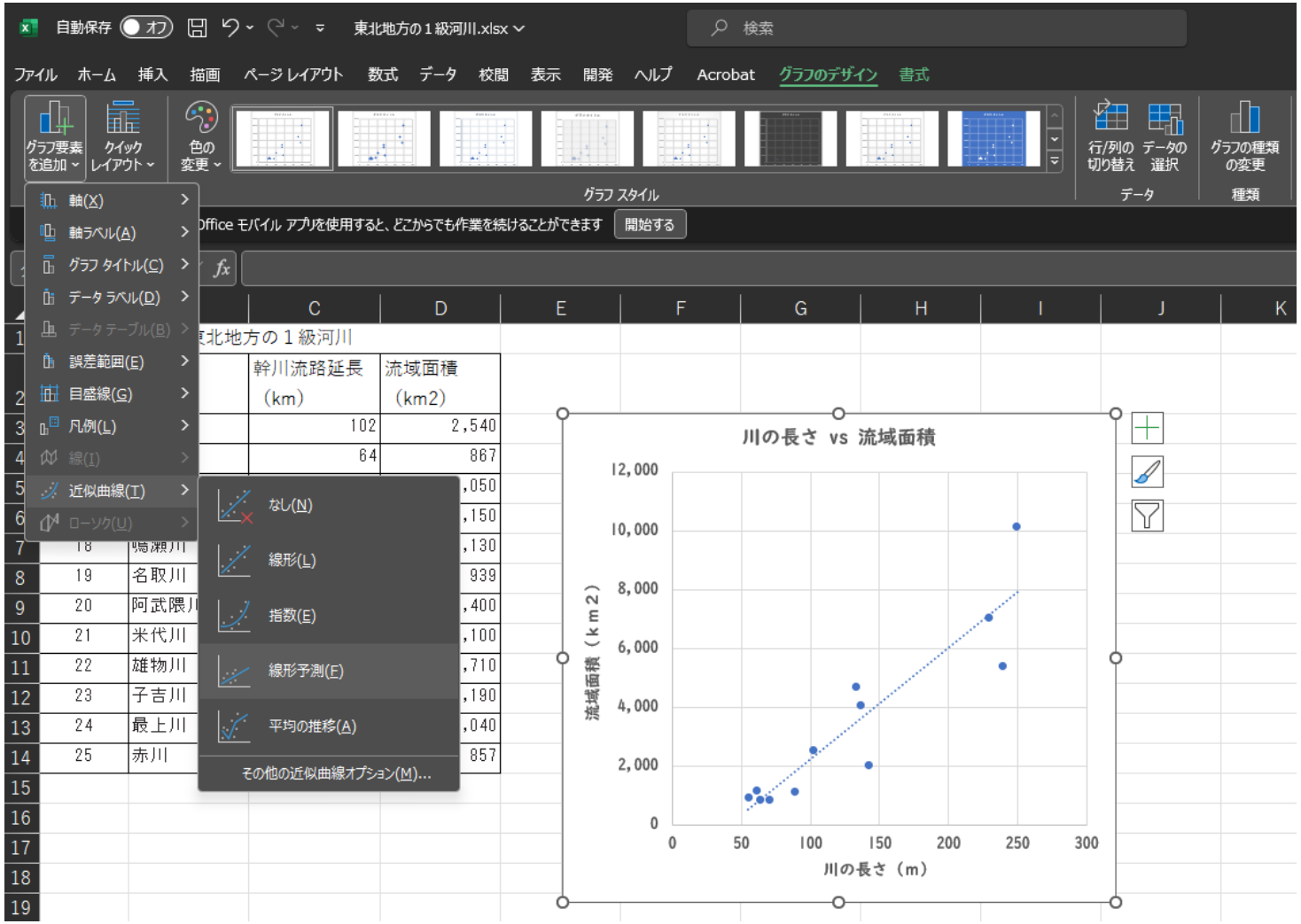
	A	B	C	D
1	東北地方の１級河川			
2	水系番号	河川名	幹川流路延長 (km)	流域面積 (km2)
3	14	岩木川	102	2,540
4	15	高瀬川	64	867
5	16	馬淵川	142	2,050
6	17	北上川	249	10,150
7	18	鳴瀬川	89	1,130
8	19	名取川	55	939
9	20	阿武隈川	239	5,400
10	21	米代川	136	4,100
11	22	雄物川	133	4,710
12	23	子吉川	61	1,190
13	24	最上川	229	7,040
14	25	赤川	70	857

目的変数

説明変数

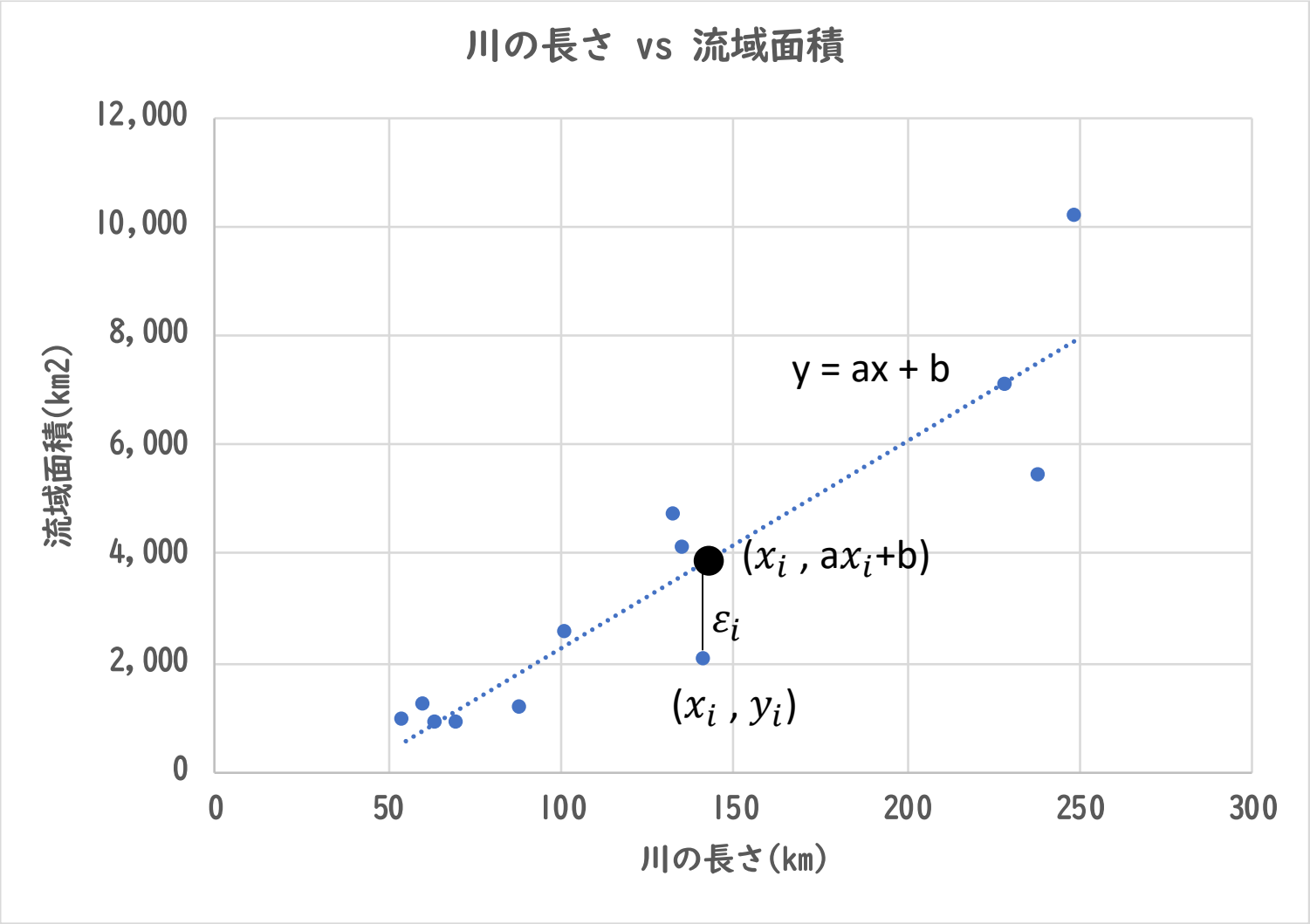
出典：https://www.mlit.go.jp/river/toukei_chousa/kasen_db/pdf/2022/4-1-4.pdf

直線の求め方


自動保存
東北地方の１級河川.xlsx
検索
ファイル ホーム 挿入 描画 ページ レイアウト 数式 データ 校閲 表示 開発 ヘルプ Acrobat グラフのデザイン 書式
グラフ要素を追加
クイックレイアウト
色の変更
グラフ スタイル
行/列の切り替え
データの選択
データ
グラフの種類の変更
種類
軸(X)
軸ラベル(A)
グラフ タイトル(C)
データ ラベル(D)
データ テーブル(B)
誤差範囲(E)
目盛線(G)
凡例(L)
線(I)
近似曲線(T)
ローソク(U)
なし(N)
線形(L)
指数(E)
線形予測(F)
平均の推移(A)
その他の近似曲線オプション(M)...
東北地方の１級河川
幹川流路延長 (km)
流域面積 (km2)
102 2,540
64 867
19 名取川
20 阿武隈川
21 米代川
22 雄物川
23 子吉川
24 最上川
25 赤川
川の長さ vs 流域面積
流域面積（km2）
川の長さ（m）
12,000
10,000
8,000
6,000
4,000
2,000
0
0 50 100 150 200 250 300


近似曲線のオプション
近似曲線のオプション
指数近似(X)
線形近似(L)
対数近似(O)
多項式近似(P)
次数(D)
累乗近似(W)
移動平均(M)
区間(E)
近似曲線名
自動(A)
線形 (系列1
ユーザー設定(C)
予測
前方補外(F)
0.0
後方補外(B)
0.0
切片(S)
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)

最小二乗法



$$\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

回帰直線の求め方

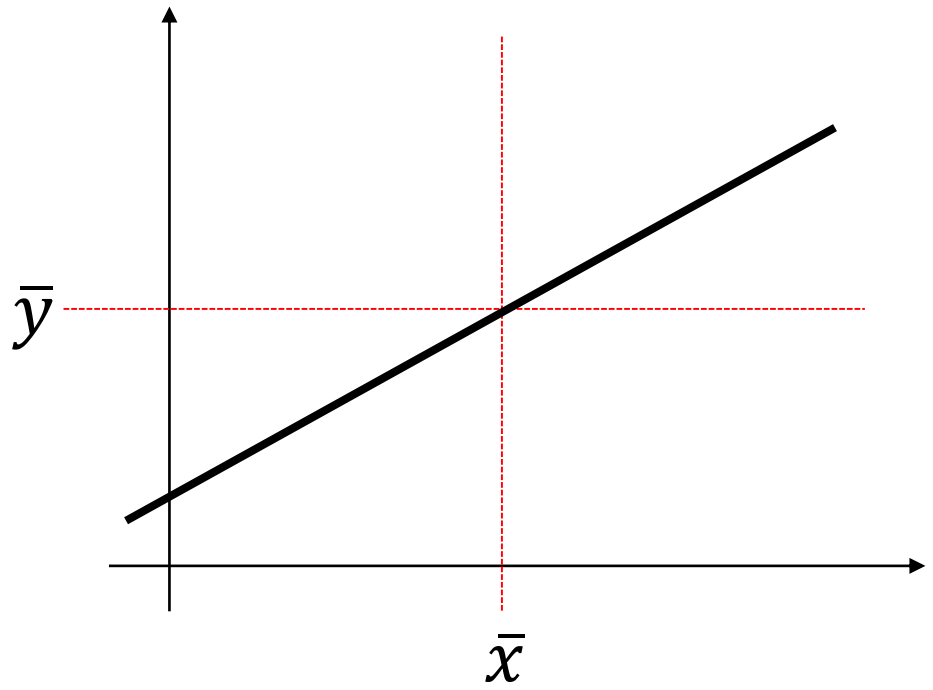
回帰直線

$$y = ax + b$$

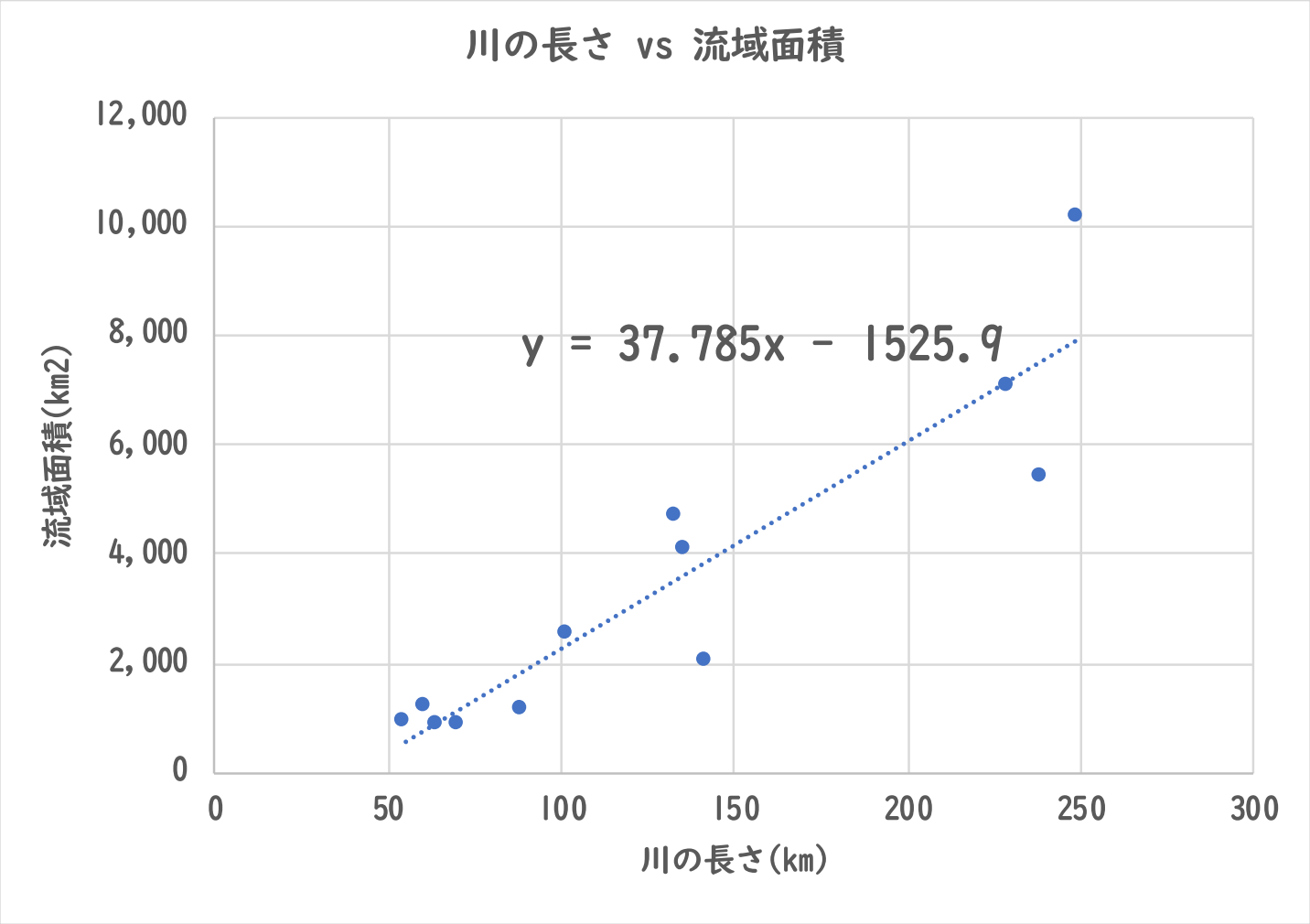
$$a = \frac{\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

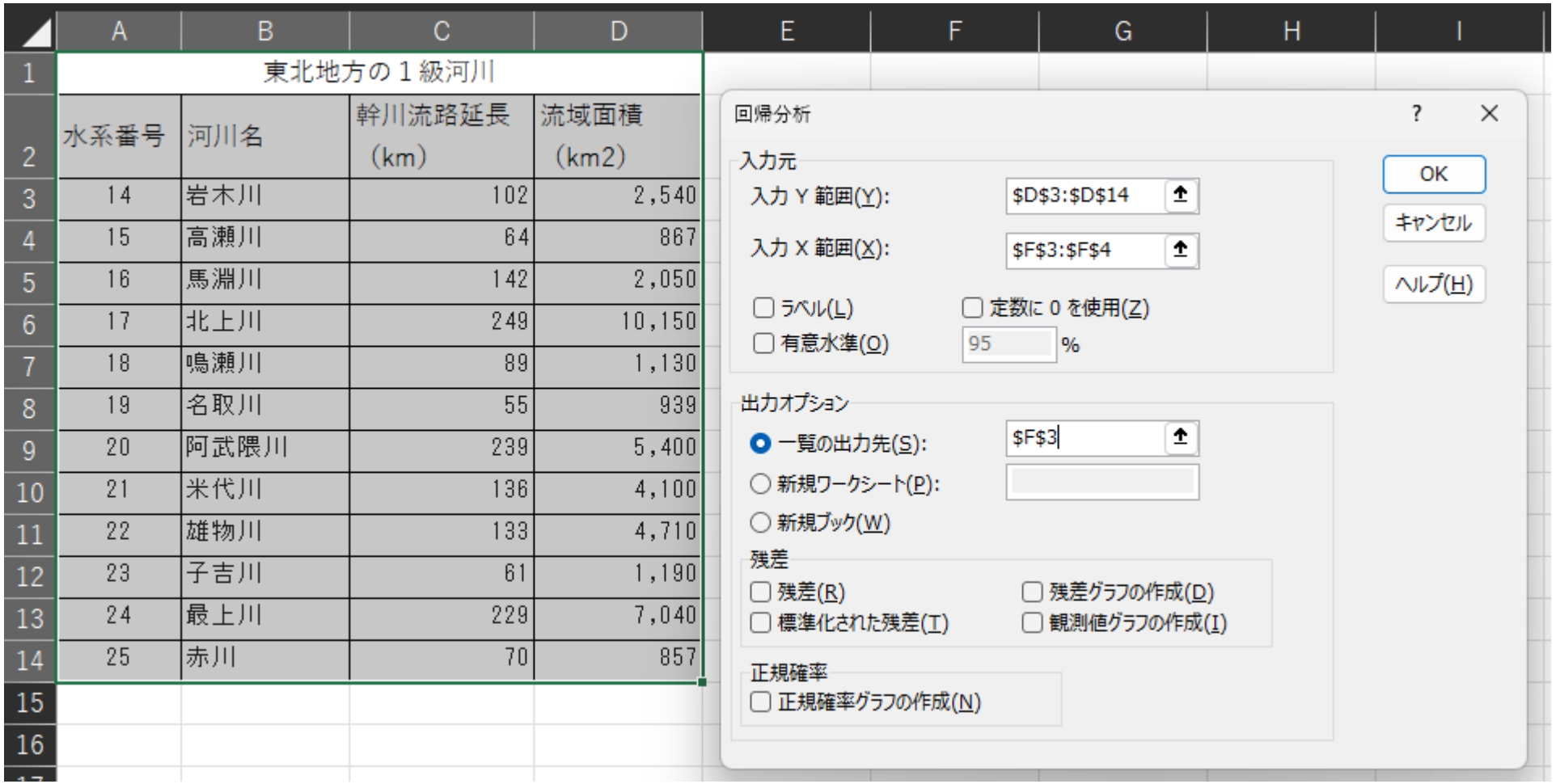
$(\bar{x}, \bar{y}$ ：平均値)



回帰式（単回帰）

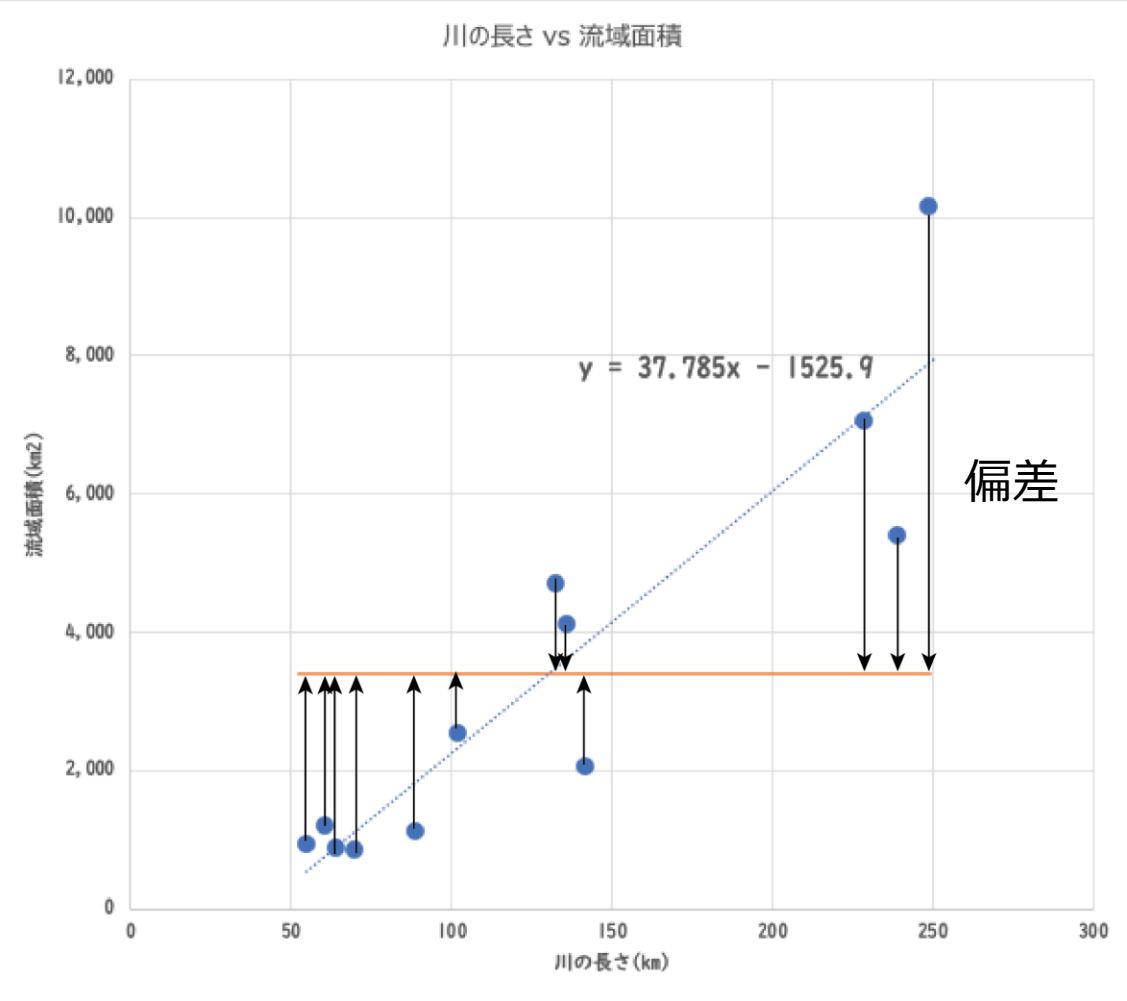

川の長さ vs 流域面積
y = 37.785x - 1525.9
流域面積(km2)
川の長さ(km)
12,000
10,000
8,000
6,000
4,000
2,000
0
0
50
100
150
200
250
300

データ分析（回帰直線）



	A	B	C	D
1	東北地方の１級河川			
2	水系番号	河川名	幹川流路延長（km）	流域面積（km2）
3	14	岩木川	102	2,540
4	15	高瀬川	64	867
5	16	馬淵川	142	2,050
6	17	北上川	249	10,150
7	18	鳴瀬川	89	1,130
8	19	名取川	55	939
9	20	阿武隈川	239	5,400
10	21	米代川	136	4,100
11	22	雄物川	133	4,710
12	23	子吉川	61	1,190
13	24	最上川	229	7,040
14	25	赤川	70	857

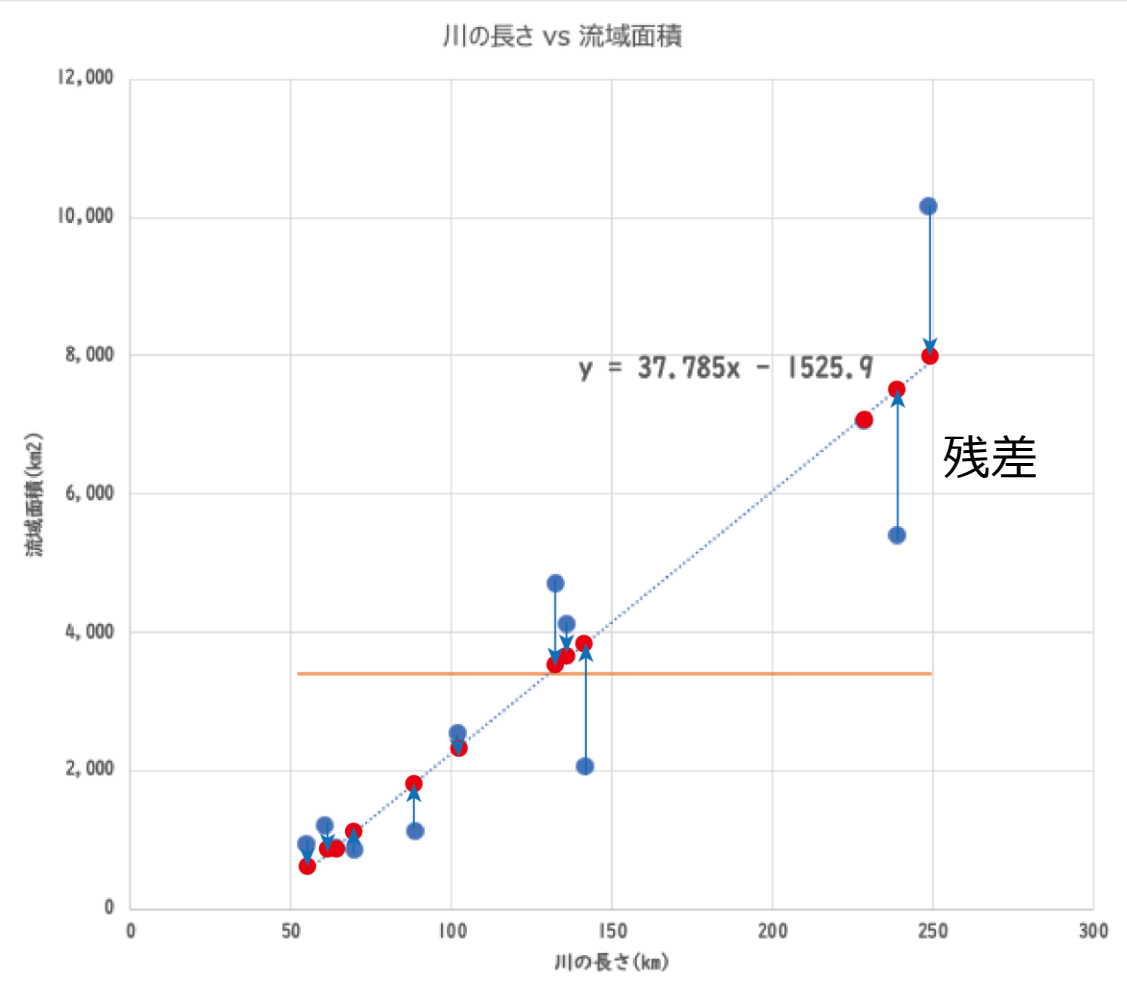
全平方和



- 全平方和
- 観測値と平均値との差の平方和

$S_T = \sum_{i=1}^{n}(\ y_i \ - \ \bar{y} \ )^2$ = 分散 × n

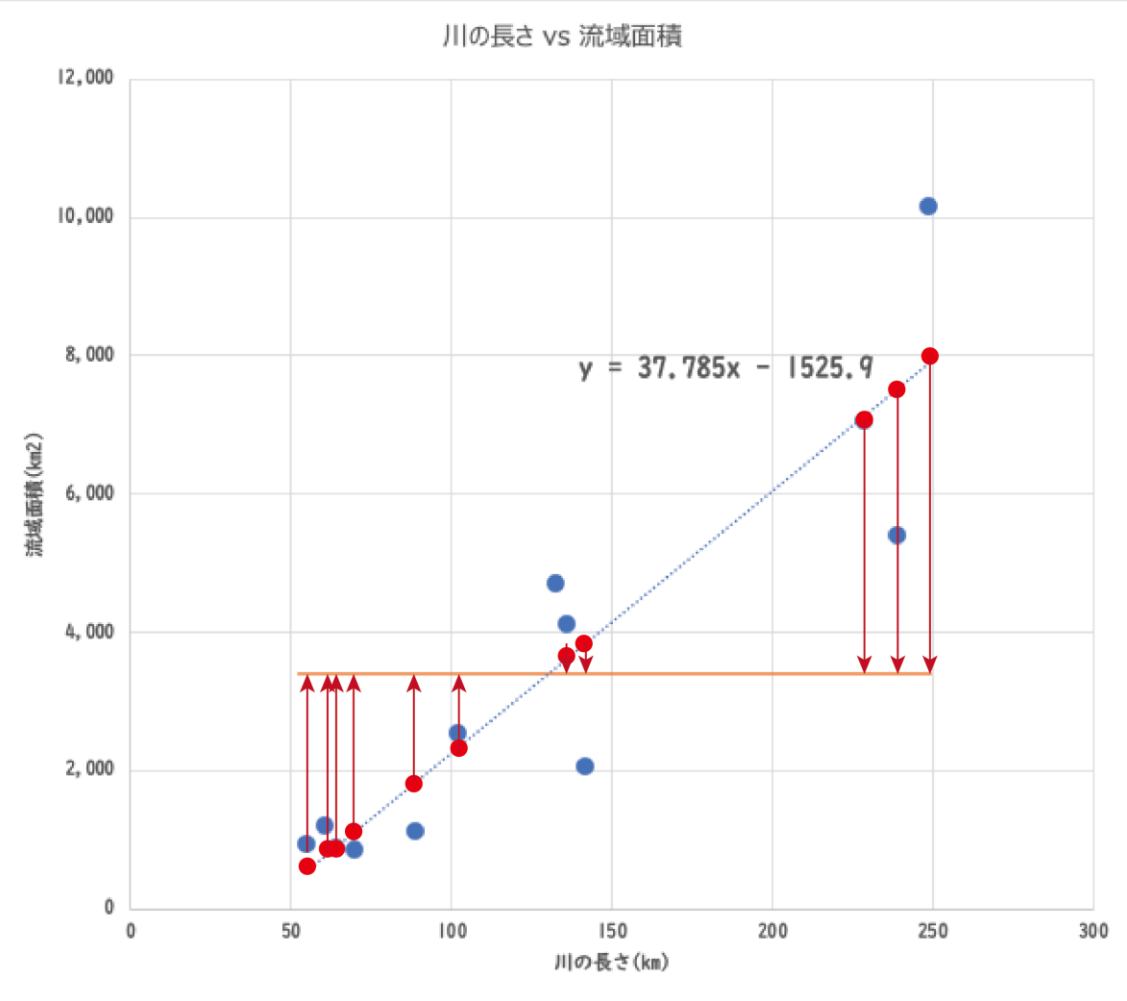
残差平方和



- 残差平方和
- 回帰直線で予測した値と観測値との残差の平方和

$$S_R = \sum_{i=1}^{n}(y - \hat{y}_i)^2$$

回帰平方和



- 回帰平方和：
- 回帰直線で予測した値と平均値との残差の平方和

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

決定係数（寄与率）

- 全平方和：観測値と平均値との差の平方和

$$S_T = \sum_{i=1}^{n}(\ y_i \ - \ \bar{y} \ )^2 = 分散 \times n$$

- 残差平方和：観測値と回帰直線で予測した値との残差の平方和

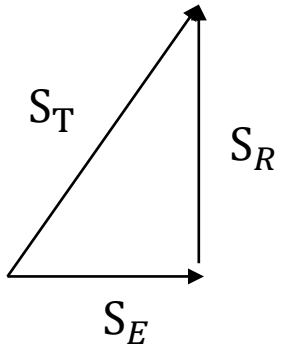
$$S_E = \sum_{i=1}^{n}(\ y_i \ - \ \hat{y_i} \ )^2$$

- 回帰平方和：回帰直線で予測した値と平均値との残差の平方和

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

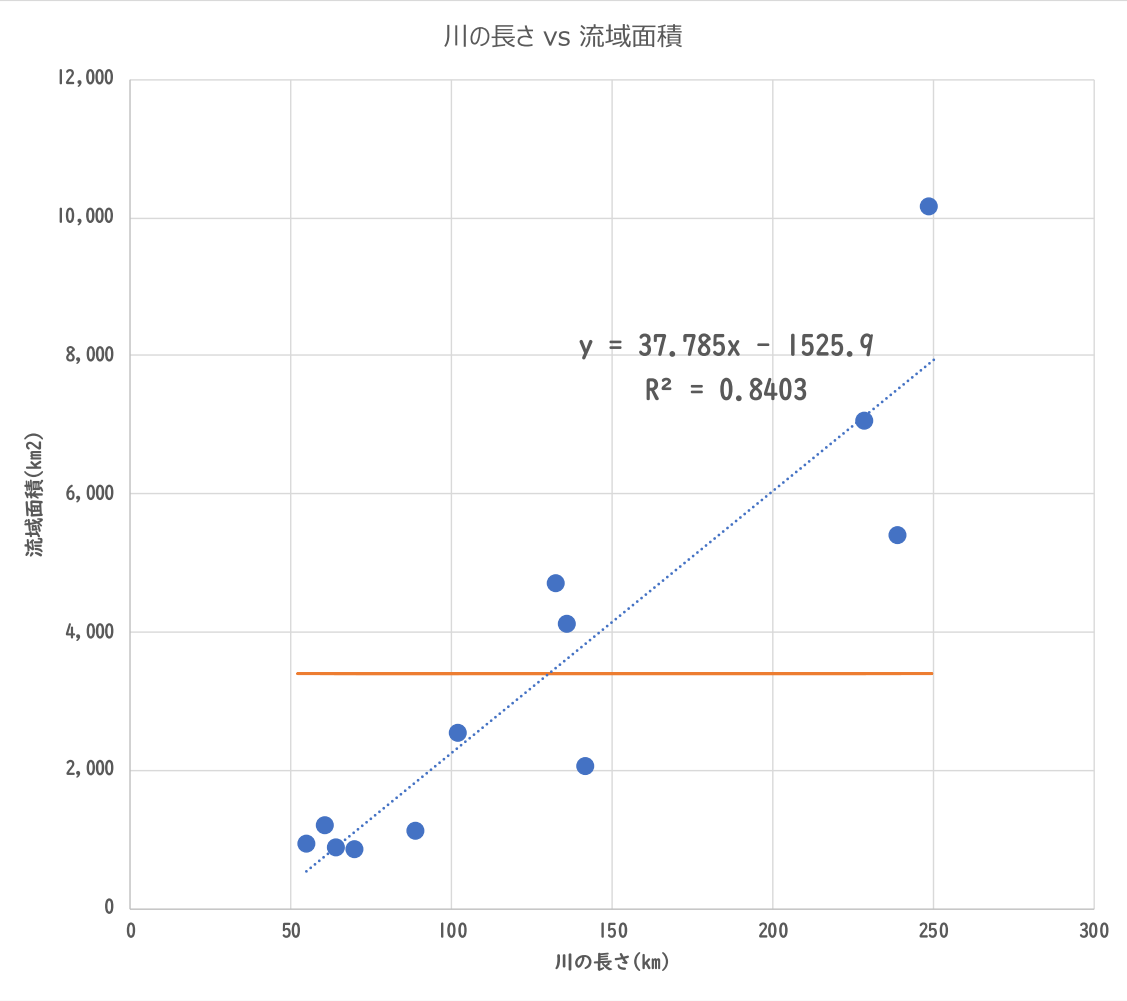
- 決定係数（寄与率）

$$R^2 = 1 - \frac{S_E}{S_T} = \frac{S_R}{S_T}$$



$$S_T = S_E + S_R$$

決定係数



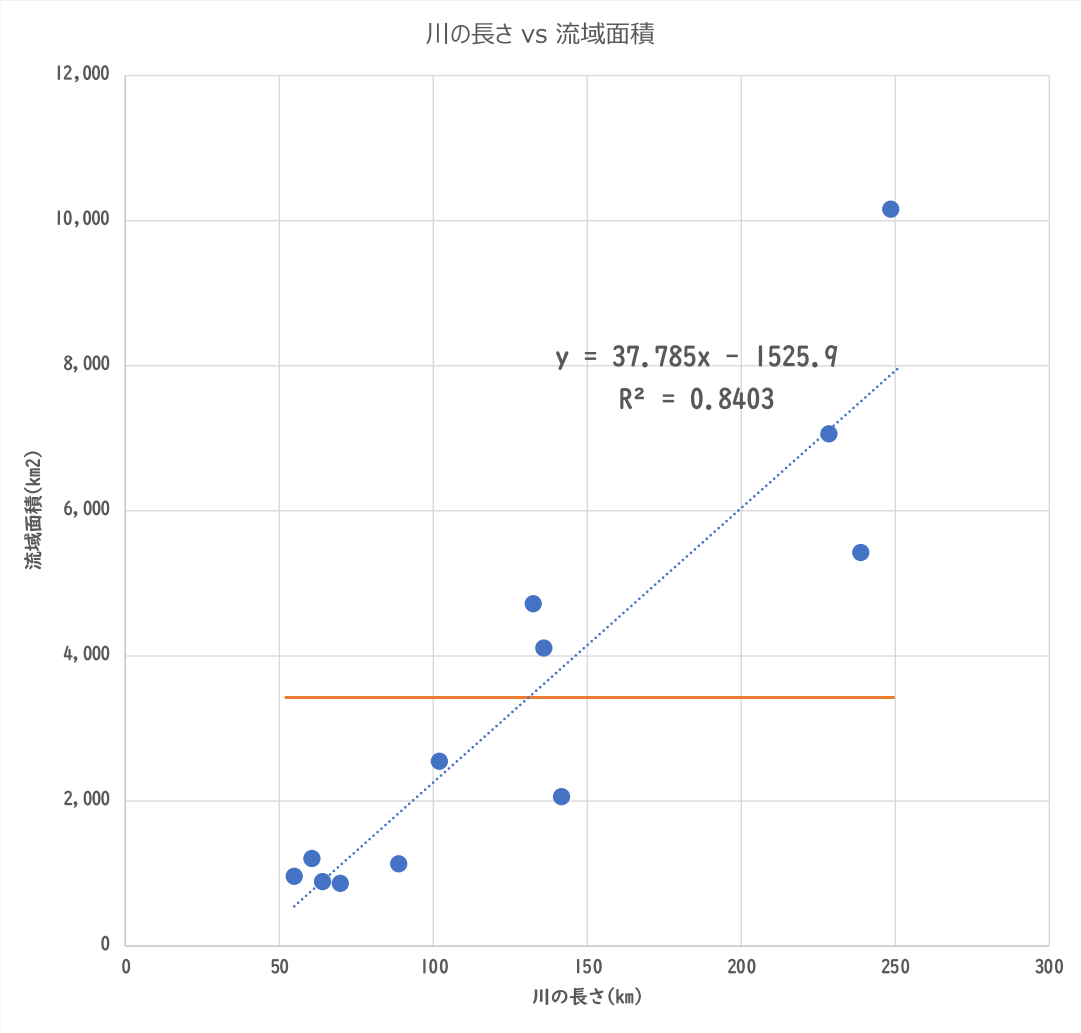
回帰統計	
重相関 R	0.91669826
重決定 R2	0.84033569
補正 R2	0.82436926
標準誤差	1241.62272
観測数	12

分散分析表

	自由度	変動	分散	測された分散
回帰	1	81137995.1	81137995.1	52.6314057
残差	10	15416269.8	1541626.98	
合計	11	96554264.9		

	係数	標準誤差	t	P-値
切片	-1525.9184	769.546457	-1.9828802	0.07551002
X 値 1	37.7845897	5.20825484	7.25475056	2.7432E-05

予測

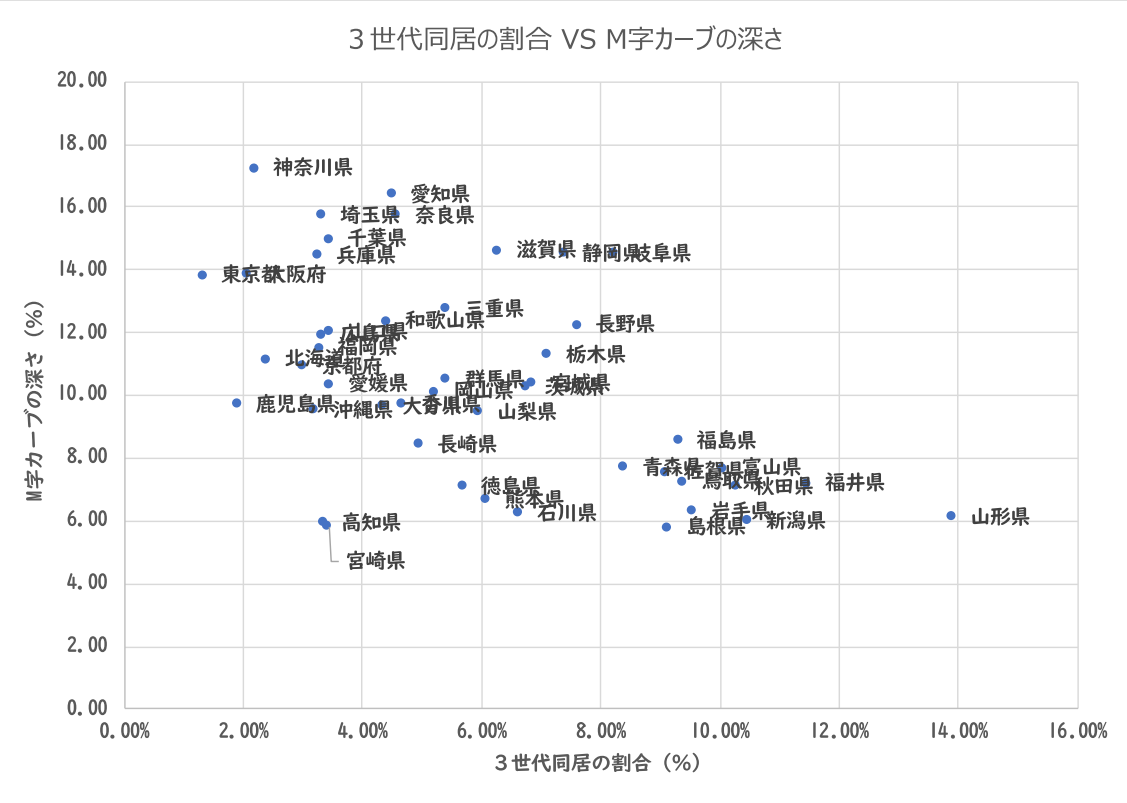


	係数
切片	-1525.9184
X 値 1	37.7845897

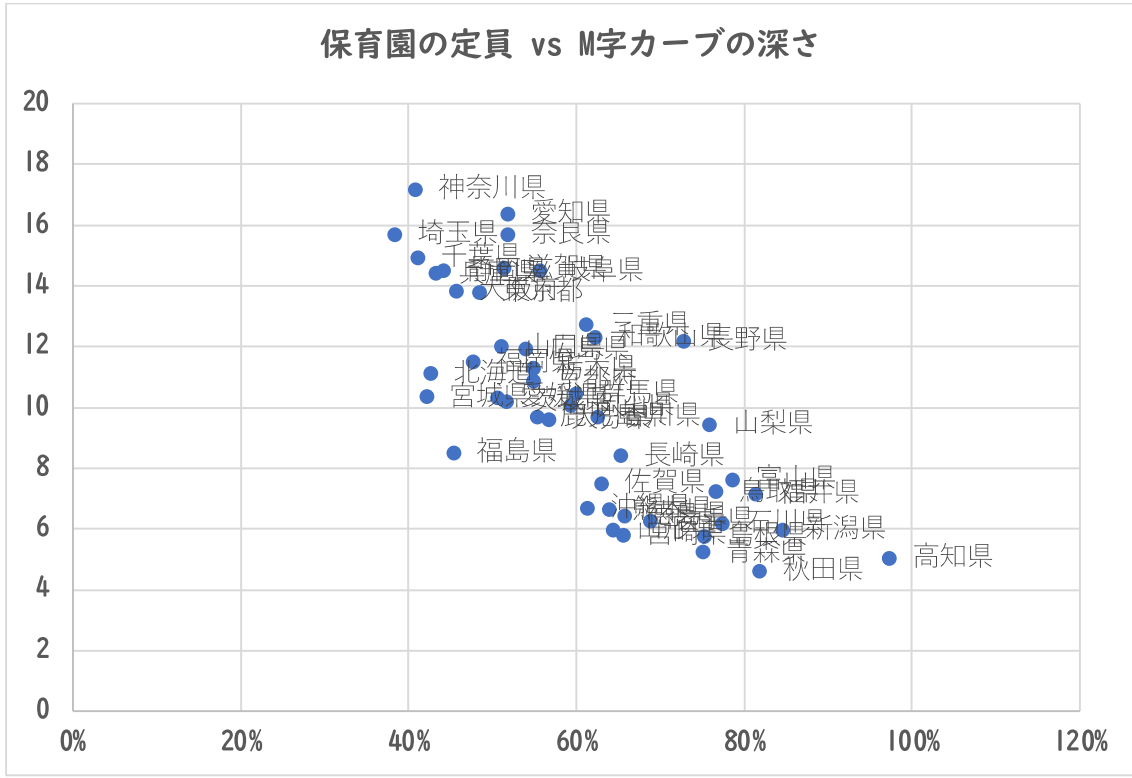
幹川流路延長 (km)	流域面積 (km2)
200	6030.99951
175	5086.38476

y = ax + b

重回帰

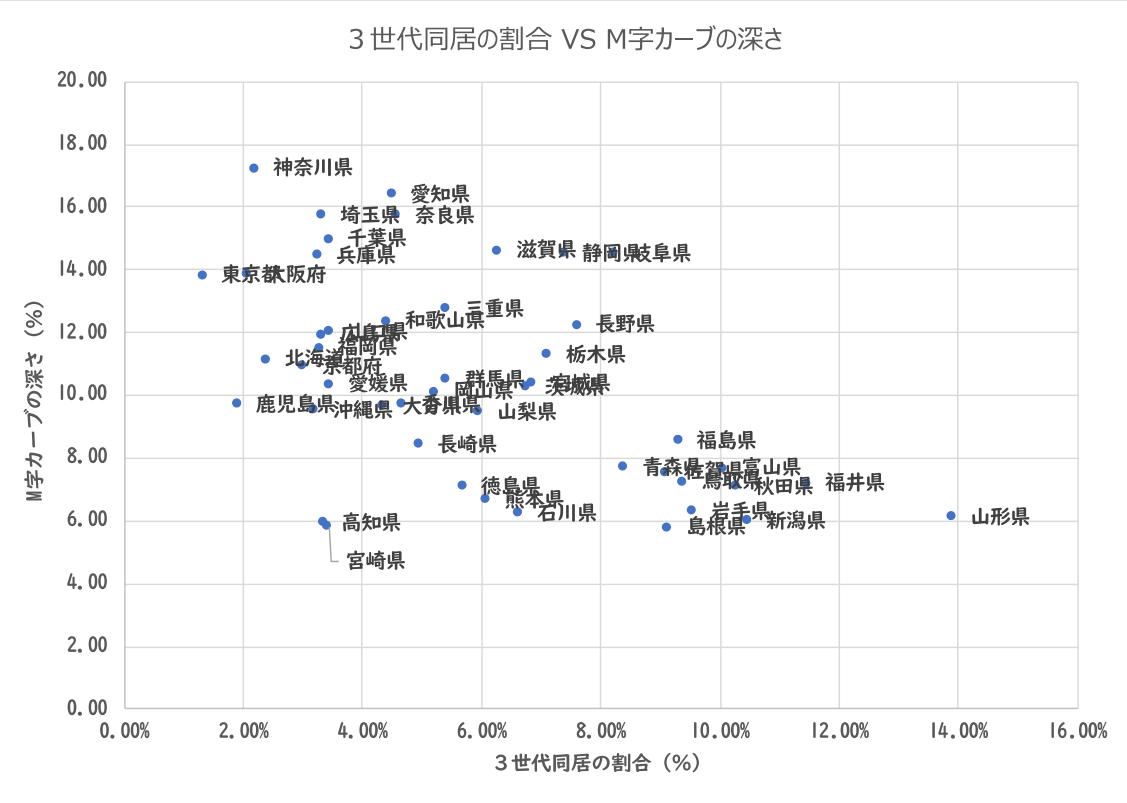


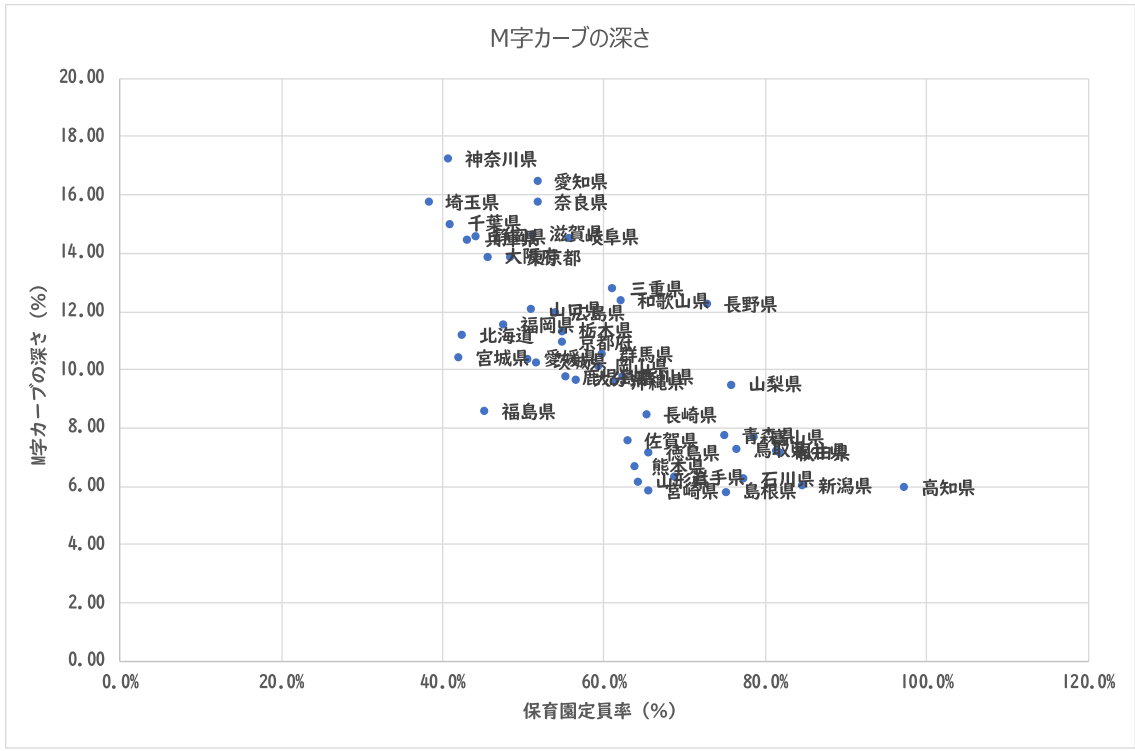
y = -61.879x + 13.827
R^2 = 0.2678



y = -19.479x + 21.884
R^2 = 0.5813

重回帰





$$y = a_1 x_1 + a_2 x_2 + b$$

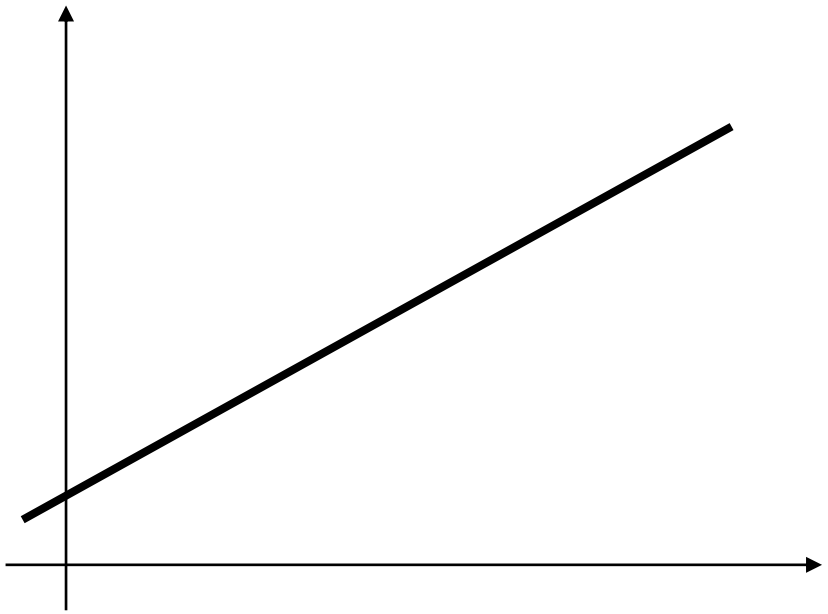
重回帰

	A	B	C	D	E	F	G	H	I
1		深さ	3世帯割合	保育園定員割合					
2	北海道	11.12102	0.02413912	0.42651695					
3	青森県	5.23738	0.08416577	0.75158061		概要			
4	岩手県	6.27019	0.09562209	0.68886185					
5	宮城県	10.35525	0.06884409	0.42181567		回帰統計			
6	秋田県	4.64123	0.10286783	0.81941679		重相関 R	0.77527434		
7	山形県	5.98784	0.13921148	0.64452048		重決定 R2	0.6010503		
8	福島県	8.52165	0.09328878	0.45441345		補正 R2	0.58291622		
9	茨城県	10.2184	0.06777093	0.51800137		標準誤差	2.2572558		
10	栃木県	11.28683	0.07137855	0.54977754		観測数	47		
11	群馬県	10.46665	0.05443374	0.59998734					
12	埼玉県	15.69623	0.03347989	0.38430689		分散分析表			
13	千葉県	14.94784	0.03457757	0.41167501			自由度	変動	分散
14	東京都	13.80184	0.01345638	0.48518317		回帰	2	337.758979	168.87949
15	神奈川県	17.1887	0.02218629	0.4089734		残差	44	224.188964	5.09520373
16	新潟県	5.9787	0.10494601	0.84660466		合計	46	561.947943	
17	富山県	7.64273	0.10082207	0.78713411					
18	石川県	6.20203	0.06651594	0.77522011			係数	標準誤差	t
19	福井県	7.16598	0.11485696	0.81504255		切片	21.7252537	1.50161404	14.4679346
20	山梨県	9.43975	0.05976803	0.75965535		X 値 1	-19.679956	13.3450648	-1.4746991
21	長野県	12.20035	0.07626097	0.729164		X 値 2	-17.285605	2.85124858	-6.0624686
22	岐阜県	14.49487	0.08235765	0.55751465					
23	静岡県	14.5099	0.07404476	0.44258035					

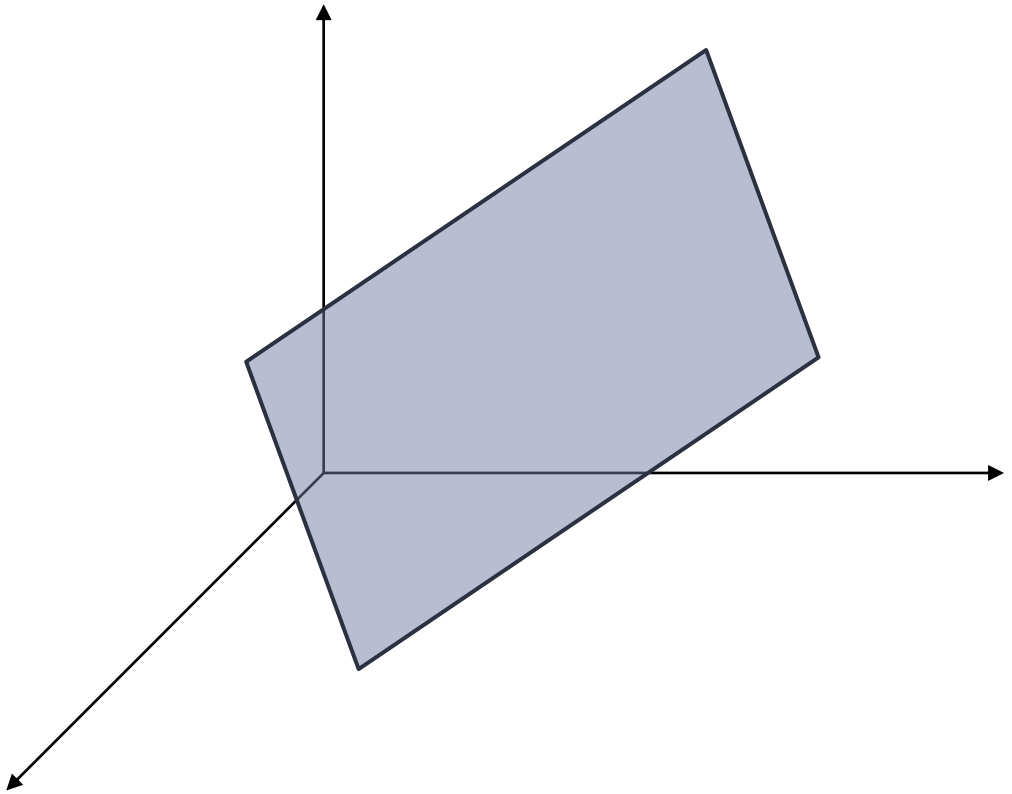
$y = -19.68 \cdot x_1 + \quad -17.28 \cdot x_2$

自由度調整済決定係数

単回帰の決定係数



重回帰の決定係数



自由度調整済決定係数

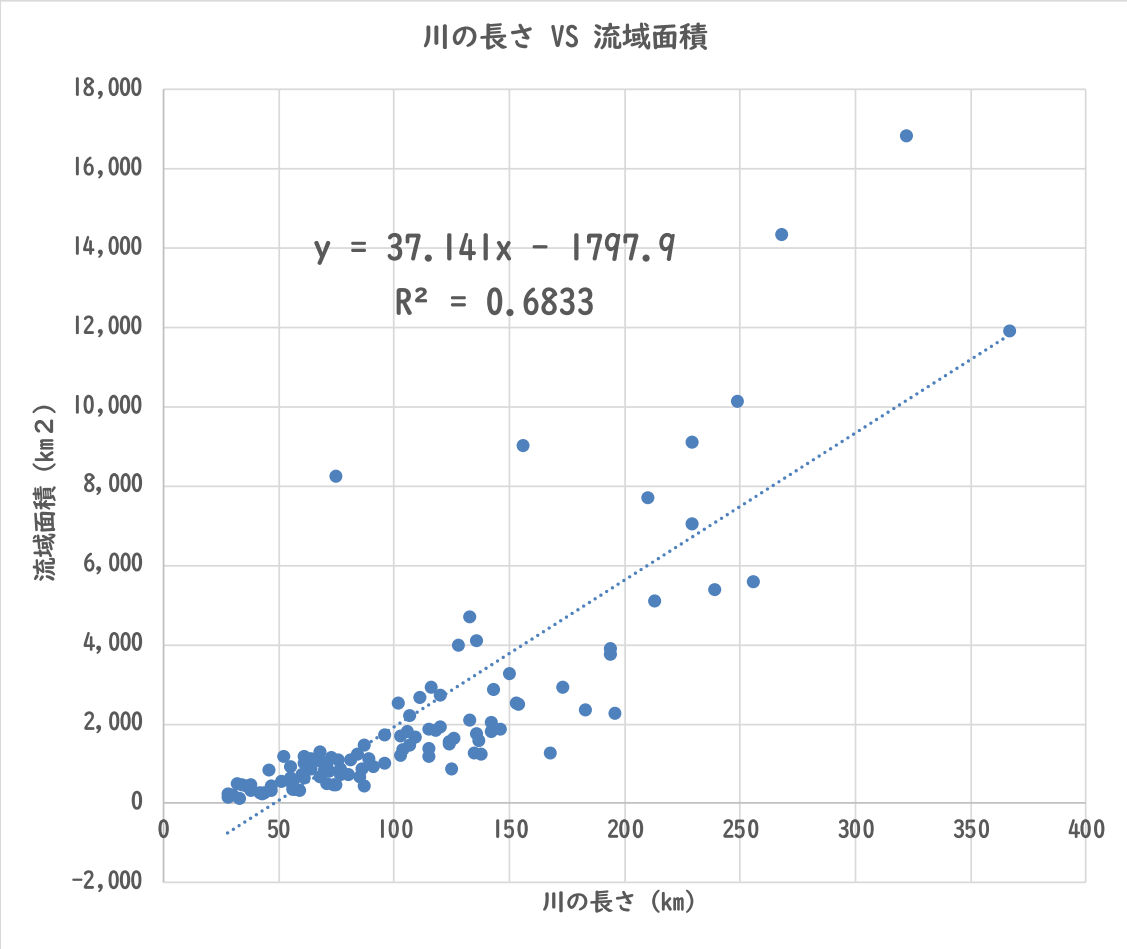
$$\text{自由度調整済}R^2 = 1 - \frac{S_E \frac{1}{N-k-1}}{S_T(N-1)} = 1 - \frac{S_E}{S_T} \times \frac{N-1}{N-k-1}$$

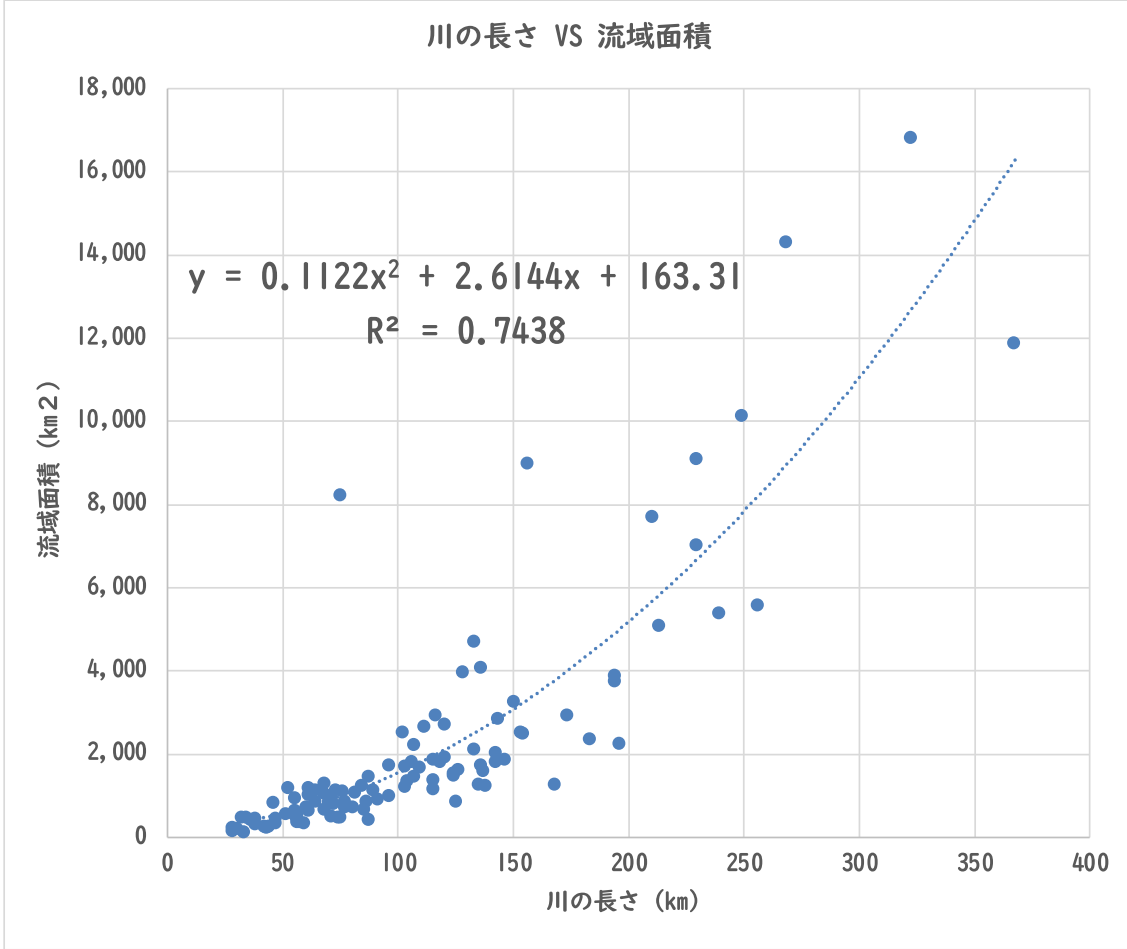
N：データ数　　k：説明変数の数

重回帰分析の注意点

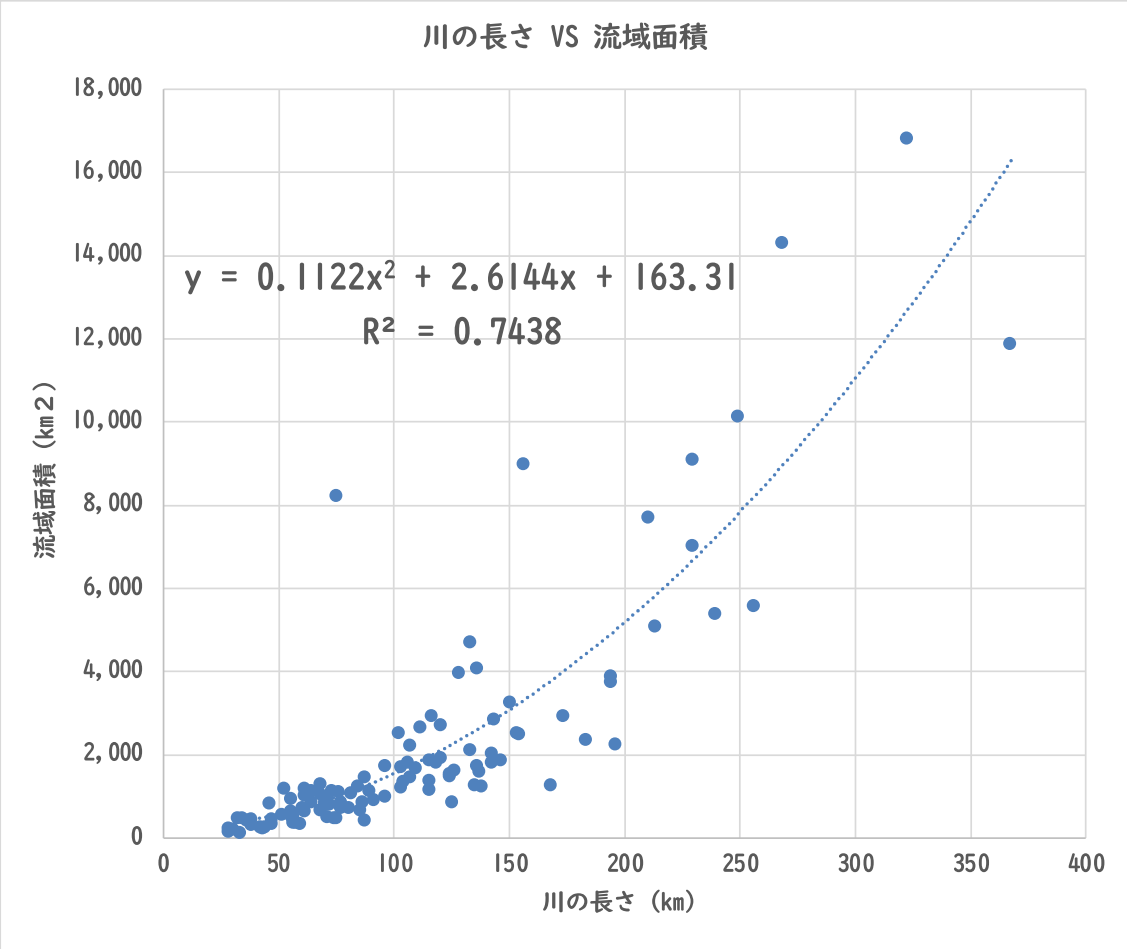
- 線形の関係しか予測できない
- 説明変数を増やすと決定係数が大きくなる
- 説明変数どうしに相関関係があると推定精度が悪くなる

二次曲線の回帰


川の長さ VS 流域面積
y = 37.141x - 1797.9
R² = 0.6833
流域面積（km２）
川の長さ（km）
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
-2,000
0
50
100
150
200
250
300
350
400


川の長さ VS 流域面積
y = 0.1122x² + 2.6144x + 163.31
R² = 0.7438
流域面積（km２）
川の長さ（km）
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
0
50
100
150
200
250
300
350
400

二次曲線の回帰



概要

回帰統計

重相関 R	0.86245253
重決定 R2	0.74382436
補正 R2	0.73899086
標準誤差	1472.46543
観測数	109

分散分析表

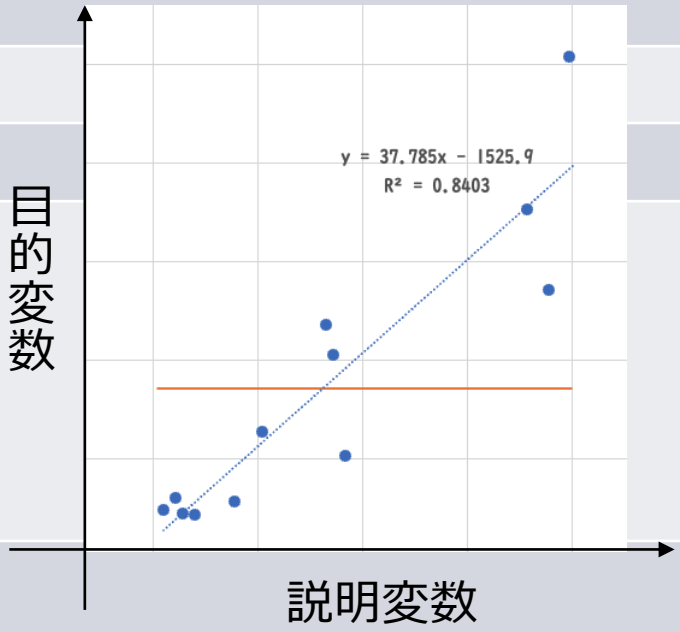
	自由度	変動	分散	則された分散	有意 F
回帰	2	667311571	333655786	153.889307	4.4927E-32
残差	106	229824372	2168154.45		
合計	108	897135943			

	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	163.313871	479.968895	0.34025928	0.73433465	-788.27115	1114.89889	-788.27115	1114.89889
X 値 1	0.11220585	0.02243016	5.00245431	2.2575E-06	0.06773587	0.15667582	0.06773587	0.15667582
X 値 2	2.61439727	7.24669291	0.36077109	0.71898855	-11.752876	16.9816708	-11.752876	16.9816708

$$y = 0.112x^2 + 2.614x + 163.31$$

まとめ

項目	内容
回帰分析	予測モデル（式）を求める
単回帰	説明変数１つ
重回帰	説明変数２つ以上
決定係数	モデルと観測値とのあてはまりのよさ $R^2 = 1 - \frac{S_E}{S_T} = \frac{S_R}{S_T}$ $0 \leqq R^2 \leqq 1$
自由度調整済決定係数	自由度調整済$R^2 = 1 - \frac{S_E}{S_T} \times \frac{N-1}{N-k-1}$
二次式の係数	2次を１つの説明変数と考えて求める



参考文献

- 「身近な統計」石崎克也・渡辺美智子　放送大学教育振興会
- 「大学生のためのデータサイエンス１」オフィシャル　スタディノート　　総務省統計局