Lecture 2（続）

２変量データの整理

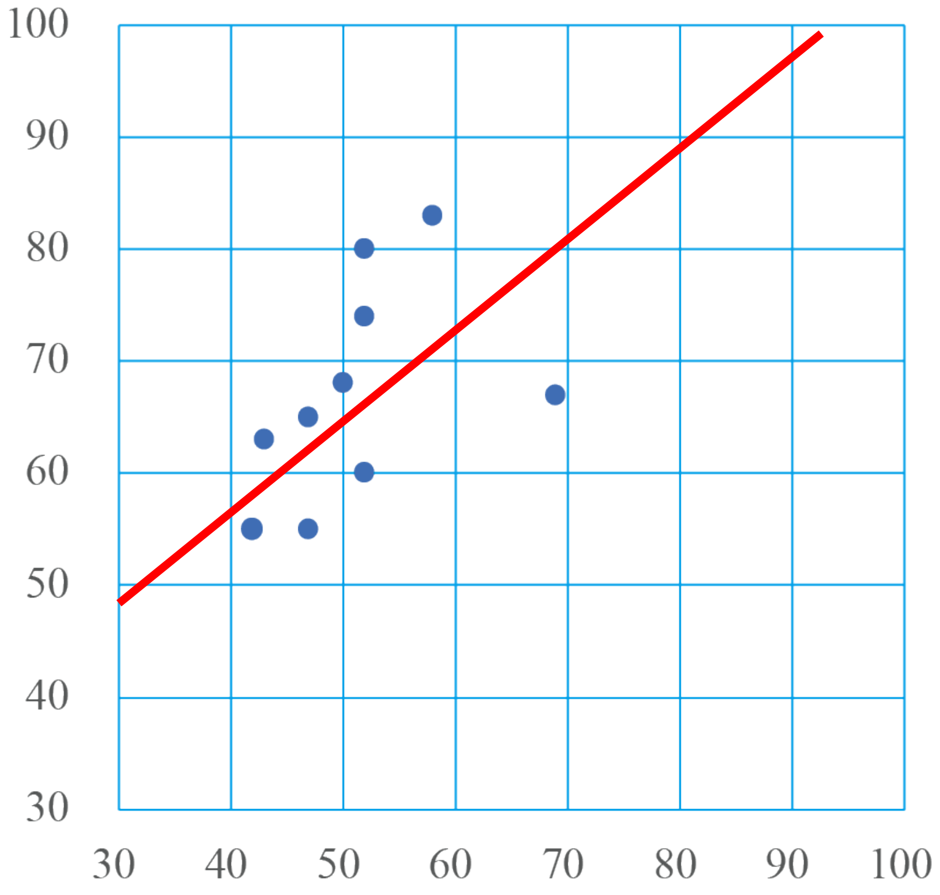
回帰分析

回帰分析

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

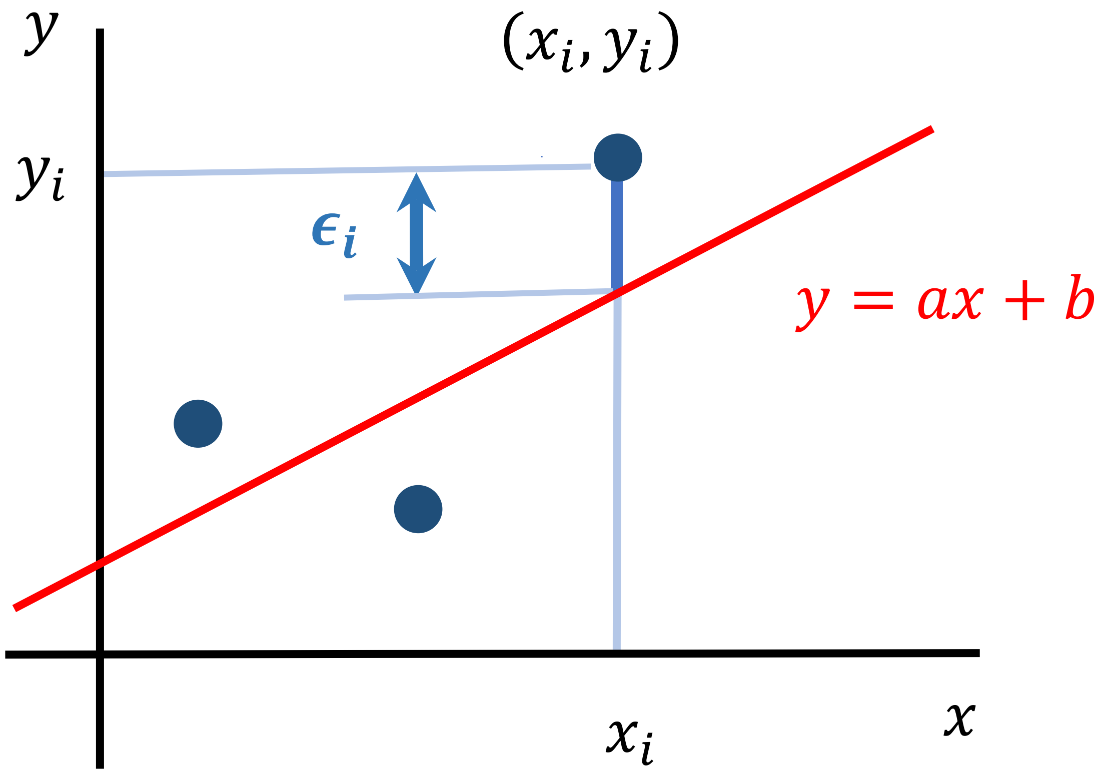
StatData02_2.csv

散布図



データを最適な
1次関数で表現したい

最小２乗法



偏差 ϵ_i

$$y_i = \underbrace{ax_i + b}_{\text{予想値}} + \epsilon_i$$

偏差平方和

$$Q = \sum \epsilon_i^2 = \sum (y_i - ax_i - b)^2$$

- 最小２乗法：偏差平方和を最小にするように a, b を決める
- $Q = Q(a, b)$ は2次式なので、最小化は初等的にできる

線形回帰モデル（回帰直線）

$$\frac{\partial Q}{\partial a} = 2an(\sigma_x^2 + \bar{x}^2) - 2n\left(\sigma_{xy} + \bar{x}\bar{y}\right) + 2bn\bar{x}$$

$$\frac{\partial Q}{\partial b} = 2bn - 2n\bar{y} + 2an\bar{x}$$

連立方程式 $\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial b} = 0$ を解いて

$$a_0 = \frac{s_{xy}}{s_x^2} = \frac{r_{xy}s_y}{s_x}$$

$$b_0 = \bar{y} - a_0\bar{x}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$ （相関係数）

定 理　x を説明変数, y を目的変数とする線形回帰モデル（回帰直線）は

$$\frac{y-\bar{y}}{s_y} = r_{xy}\frac{x-\bar{x}}{s_x}$$

注 意　y を説明変数, x を目的変数とする線形回帰モデル（回帰直線）は

$$\frac{x-\bar{x}}{s_x} = r_{xy}\frac{y-\bar{y}}{s_y}$$

異なる

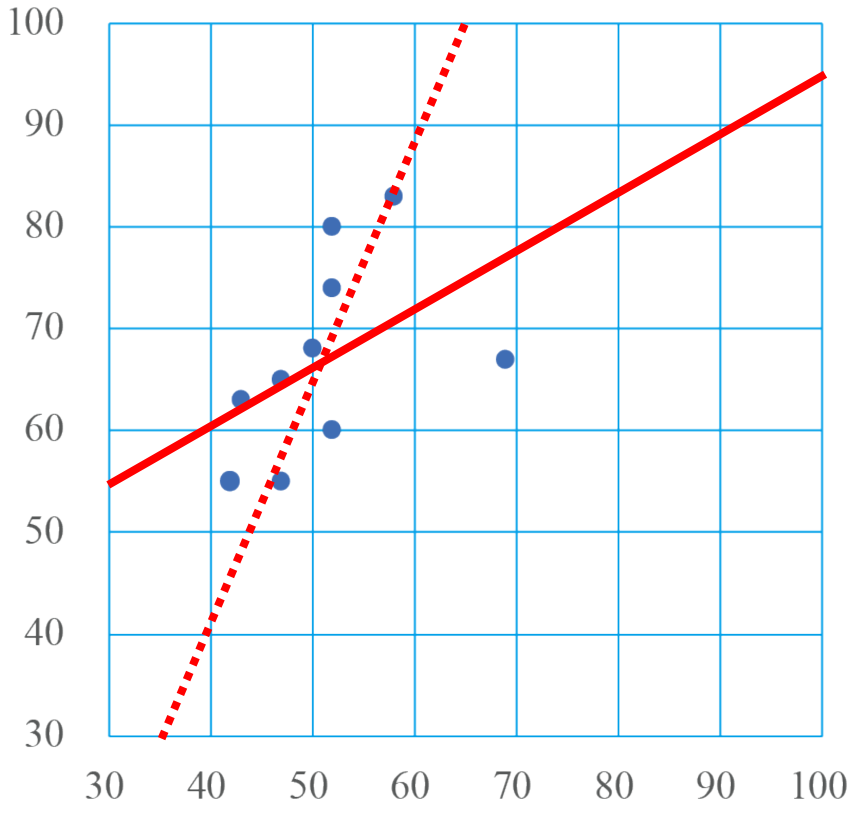
例題 2.2　受講生10名の中間試験と期末試験の結果から回帰直線を求めよ.

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

StatData02_2.csv

例題 2.2　受講生10名の中間試験と期末試験の結果から回帰直線を求めよ．

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55



$\bar{x} = 51.2$　　$\bar{y} = 67.0$

$s_x = 7.44$　　$s_y = 9.12$　　$r_{xy} = 0.47$

回帰直線（ x ：説明変数）

$$\frac{y-\bar{y}}{s_y} = r_{xy}\frac{x-\bar{x}}{s_x} \Longrightarrow y = 0.58x + 37.3$$

回帰直線（ y ：説明変数）

$$\frac{x-\bar{x}}{s_x} = r_{xy}\frac{y-\bar{y}}{s_y} \Longrightarrow x = 0.38y + 25.5$$

身長, 体重, 年齢

番号	選手名	身長	体重	年齢
1	ブラッシュ	196	106	30
2	弓削　隼人	193	105	25
3	清宮　虎多朗	190	84	19
4	J. T. シャギワ	190	90	29
⋮	⋮	⋮	⋮	⋮
81	福山　博之	172	70	30
82	茂木　栄五郎	171	75	26
83	小深田　大翔	168	69	24

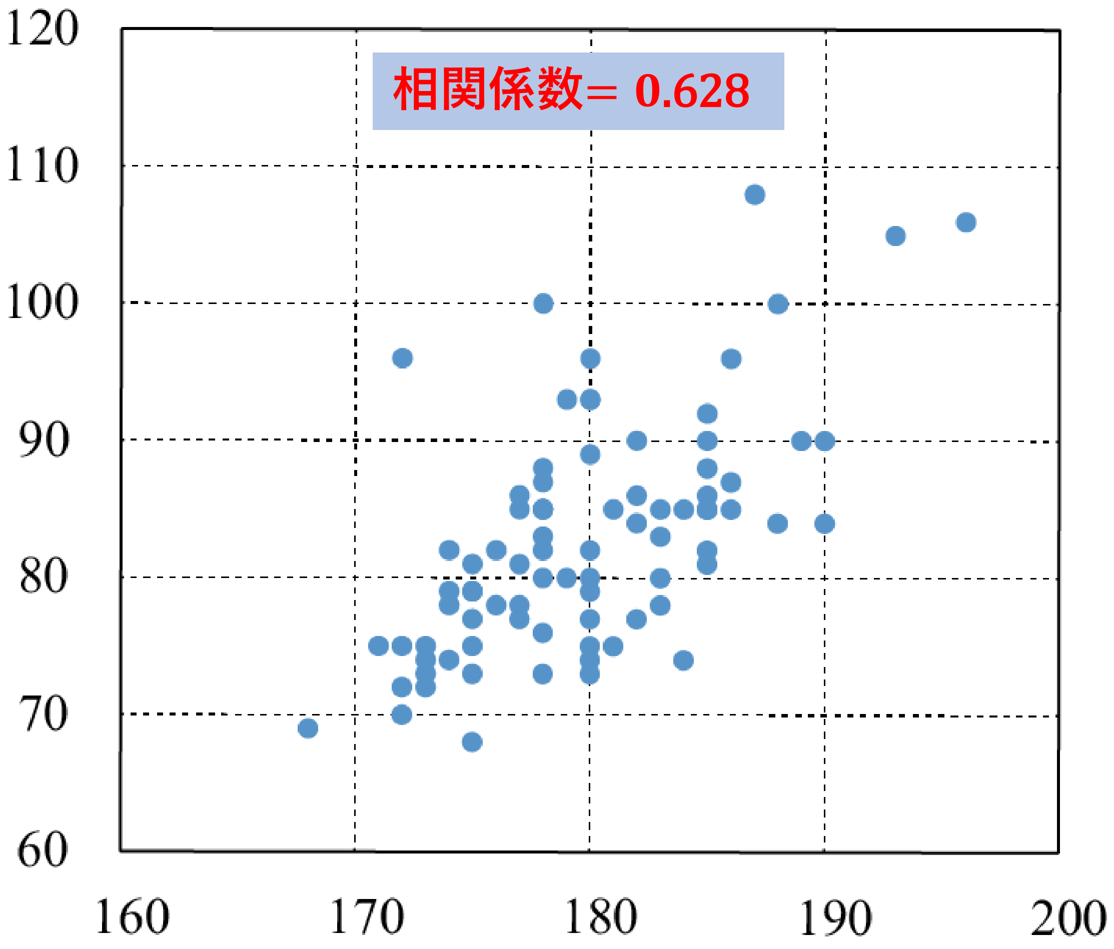
（あるスポーツチームのデータ）

身長, 体重, 年齢

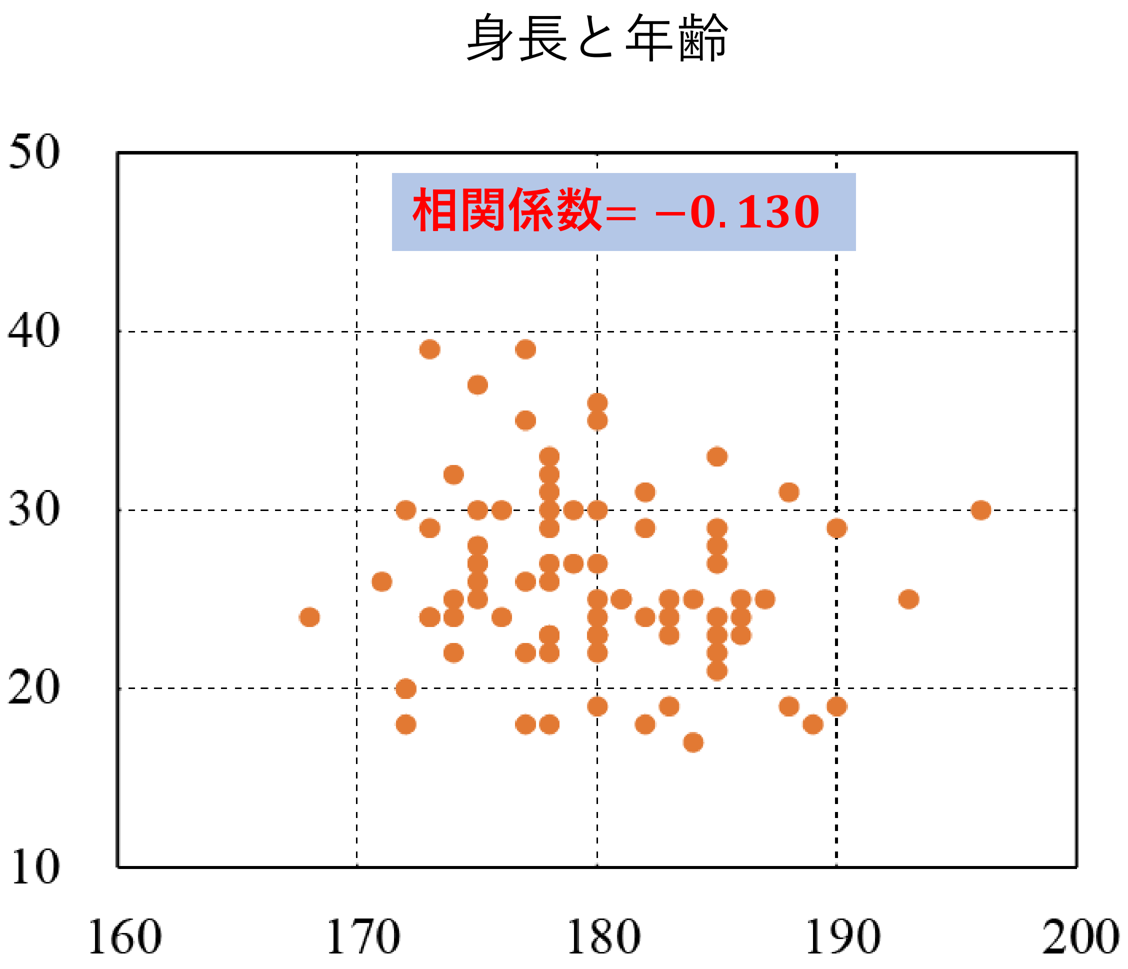
番号	選手名	身長	体重	年齢
1	ブラッシュ	196	106	30
2	弓削　隼人	193	105	25
3	清宮　虎多朗	190	84	19
4	J. T. シャギワ	190	90	29
⋮	⋮	⋮	⋮	⋮
81	福山　博之	172	70	30
82	茂木　栄五郎	171	75	26
83	小深田　大翔	168	69	24

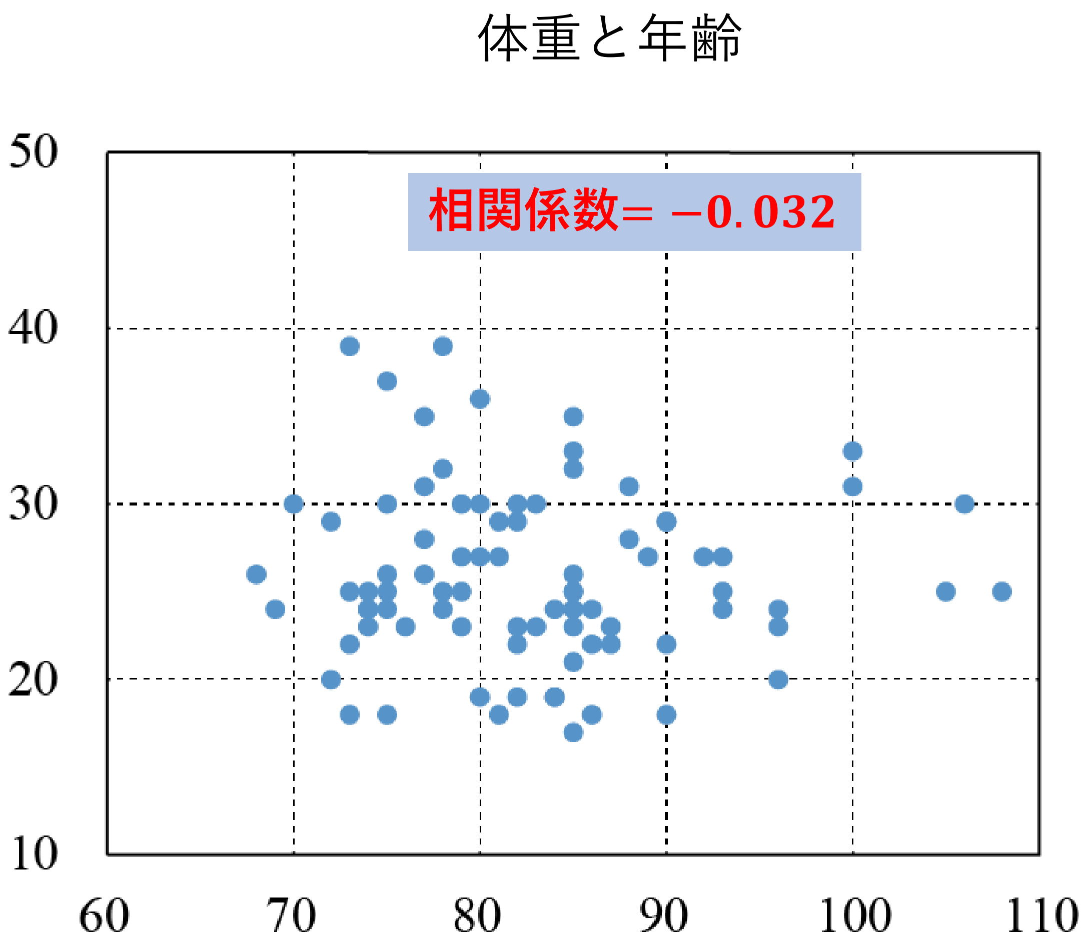
（あるスポーツチームのデータ）

身長と体重



身長, 体重, 年齢


身長と年齢
相関係数= −0.130
50
40
30
20
10
160
170
180
190
200

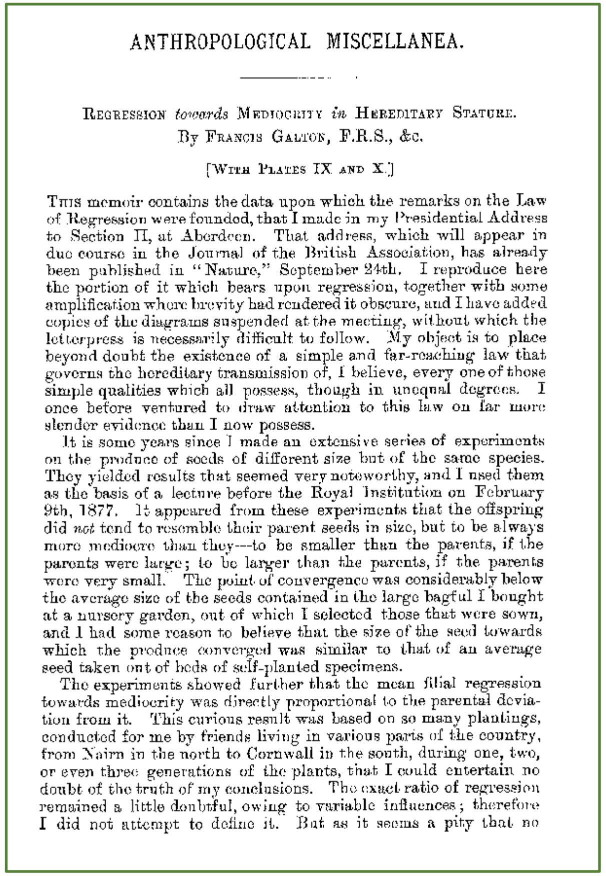

体重と年齢
相関係数= −0.032
50
40
30
20
10
60
70
80
90
100
110

親の身長と子の身長 (x, y)

Heights of Adult Children (y) \ Mid-Heights of Parents (x)	below	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	above	sum
above							5	3	2	4		14
73.2						3	4	3	2	2	3	17
72.2			1		4	4	11	4	9	7	1	41
71.2			2		11	18	20	7	4	2		64
70.2			5	4	19	21	25	14	10	1		99
69.2	1	2	7	13	38	48	33	18	5	2		167
68.2	1		7	14	28	34	20	12	3	1		120
67.2	2	5	11	17	38	31	27	3	4			138
66.2	2	5	11	17	36	25	17	1	3			117
65.2	1	1	7	2	15	16	4	1	1			48
64.2	4	4	5	5	14	11	16					59
63.2	2	4	9	3	5	7	1	1				32
62.2		1		3	3							7
below	1	1	1			1		1				5
sum	14	23	66	78	211	219	183	68	43	19	4	928

F. Galton（ゴルトン）
Regression towards mediocrity in hereditary stature, Anthropological Miscellanea (1886)

回帰分析の始まり



ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

Adult Children (y) \ Mid-height parents (x)	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	sum
73.2					3	4	3	2	2	14
72.2		1		4	4	11	4	9	7	40
71.2		2		11	18	20	7	4	2	64
70.2		5	4	19	21	25	14	10	1	99
69.2	2	7	13	38	48	33	18	5	2	166
68.2		7	14	28	34	20	12	3	1	119
67.2	5	11	17	38	31	27	3	4		136
66.2	5	11	17	36	25	17	1	3		115
65.2	1	7	2	15	16	4	1	1		47
64.2	4	5	5	14	11	16				55
63.2	4	9	3	5	7	1	1			30
62.2	1		3	3						7
sum	22	65	78	211	218	178	64	41	15	892

$\bar{x} = 68.3$　　$\bar{y} = 68.1$

$s_x^2 = 2.77$　　$s_y^2 = 5.62$

$s_x = 1.67$　　$s_y = 2.37$

$s_{xy} = 1.60$

$r_{xy} = 0.41$

回帰直線（ x：説明変数）

$$y = 0.58\,x + 28.36$$

1 inch = 2.54 cm を用いてcm で表すと

$$y = 0.58\,x + 72$$

例（1）　$x = 175 \rightarrow y = 173.5$

例（2）　$x = 160 \rightarrow y = 164.8$

Python を使ってみる

StatData02_1.csv

番号	身長 (x)	体重 (y)
1	46.0	2700
2	49.5	3220
3	50.0	3360
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
60	48.0	2530

StatData02_2.csv

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

StatData02_3.csv

選手名	身長	体重	守備	生年月日	年齢	年数	血液型	投打	出身地	年俸(推定)
ブラッシュ	196cm	106kg	外野手	1989/7/4	30歳	2年	不明	右右	アメリカ	-
弓削隼人	193cm	105kg	投手	1994/4/6	25歳	2年	AB	左左	栃木	-
清宮虎多朗	190cm	84kg	投手	2000/5/26	19歳	2年	A	右左	千葉	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

StatData02_1b - Jupyter Notebook.pdf

StatData02_2b - Jupyter Notebook.pdf

StatData02_3 - Jupyter Notebook.pdf

Lecture 2

2 変量データの整理

おわり