STA302: Regression Analysis

Statistics

- Objective: To draw reasonable conclusions from noisy numerical data
- Entry point: Study relationships between variables

Data File

- Rows are **cases**. There are *n* cases.
- Columns are **variables**. A variable is a piece of information that is recorded for every case.

1	2	2	0	78.0	65	80	39	English	Female	3	3	1
2	2	6	2	66.0	54	75	57	English	Female	3	3	1
3	2	4	4	80.2	77	70	62	English	Male	5	6	1
4	2	5	2	81.7	80	67	76	English	Female	2	2	1
5	2	4	4	86.8	87	80	86	English	Male	5	5	1
6	2	3	1	76.7	53	75	60	English	Male	3	3	1
7	2	3	2	85.8	86	81	54	Other	Female	2	2	1
8	2	4	3	73.0	75	77	17	English	Male	4	5	1
9	2	6	2	72.3	63	60	2	English	Male	4	4	1
10	2	8	6	90.3	87	88	76	English	Male	4	4	1
11	2	8	3	.	.	.	60	English	Male	1	2	1
12	2	6	4	.	.	.	61	Other	Female	1	1	1
13	.	.	.	87.2	84	83	54	English	Male	3	3	1
14	2	2	5	91.0	90	91	84	English	Male	5	5	1
15	2	3	1	72.8	53	74	.	English	Female	3	3	1
16	.	.	.	80.7	72	84	14	English	Male	3	3	1
17	2	5	0	82.5	82	85	75	Other	Female	2	2	1
18	2	4	6	91.5	95	81	94	English	Female	3	3	1
19	2	3	2	78.3	77	74	60	English	Female	3	3	1
20	.	.	.	74.5	0	85	.	English	Male	4	4	1
21	2	3	3	80.7	71	78	53	Other	Female	1	3	1
22	2	5	3	88.3	80	85	63	English	Female	3	3	1
23	2	4	2	76.8	82	64	82	Other	Female	2	2	1

Skipping ....

570	2	5	4	84.8	88	68	80	English	Male	1	1	1
571	2	4	3	78.3	83	84	56	English	Male	4	2	1
572	2	6	3	88.3	81	90	70	English	Female	5	5	1
573	2	3	1	.	.	.	.	English	Male	3	3	1
574	2	5	9	77.0	73	79	60	English	Female	2	2	1
575	.	.	.	78.7	80	73	.	English	Female	6	3	1
576	2	5	2	80.7	80	70	50	Other	Male	1	1	1
577	2	4	2	80.7	56	81	50	English	Female	2	2	1
578	2	4	3	.	.	.	78	Other	Female	4	4	1
579	1	6	1	82.2	80	86	61	English	Female	2	2	1

Variables can be

- Independent: Predictor or cause (contributing factor)

- Dependent: Predicted or effect

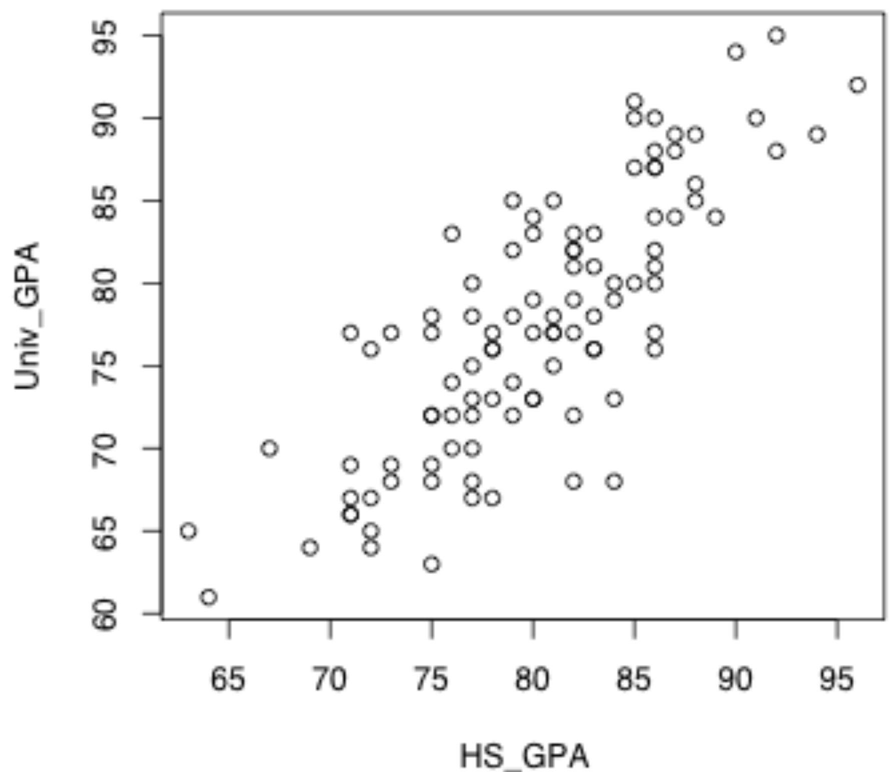
Simple regression and correlation

- Simple means one IV
- DV quantitative
- IV usually quantitative too

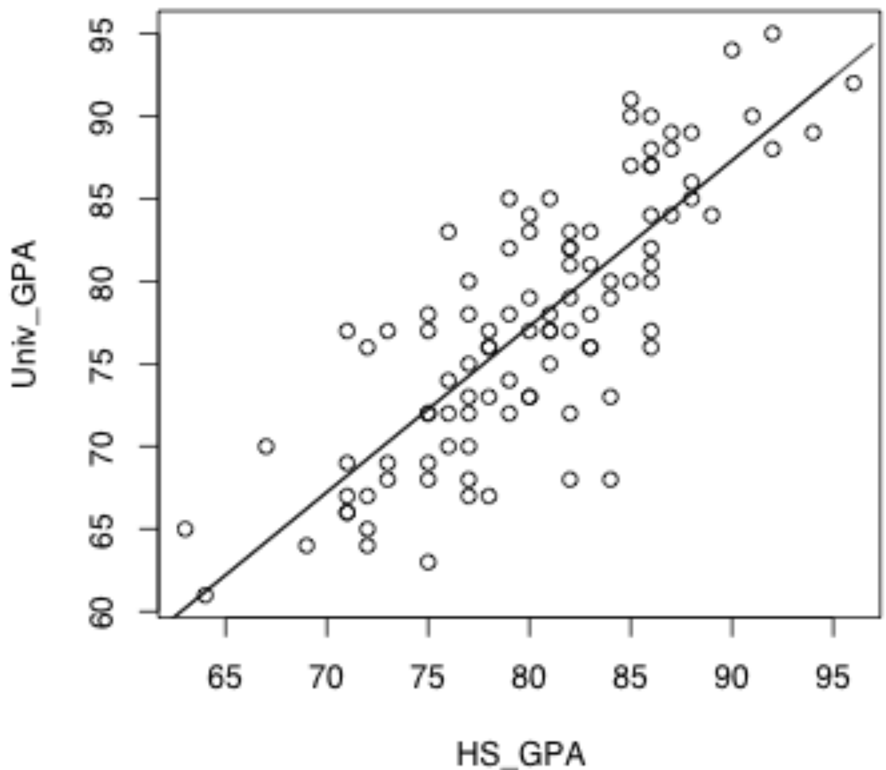
Simple regression and correlation

High School GPA	University GPA
88	86
78	73
87	89
86	81
77	67
…	…

Scatterplot



Least squares line



Correlation between variables

- $$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

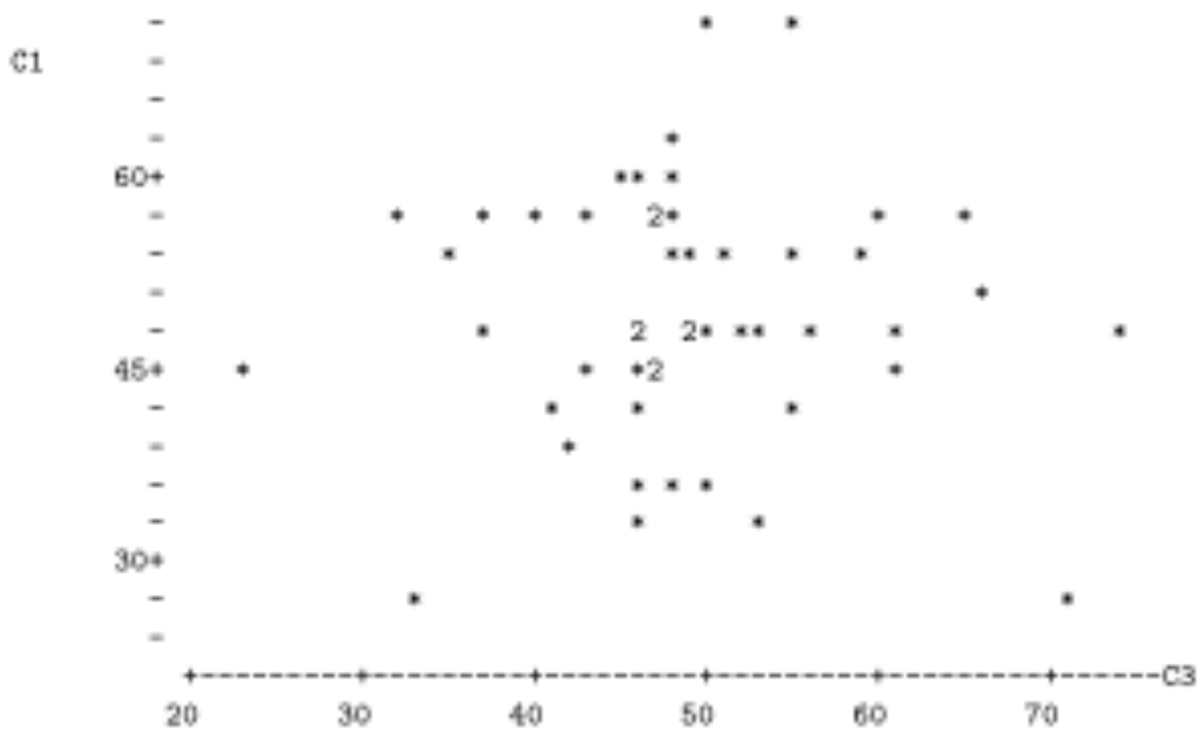
is an estimate of

- $$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Correlation coefficient *r*

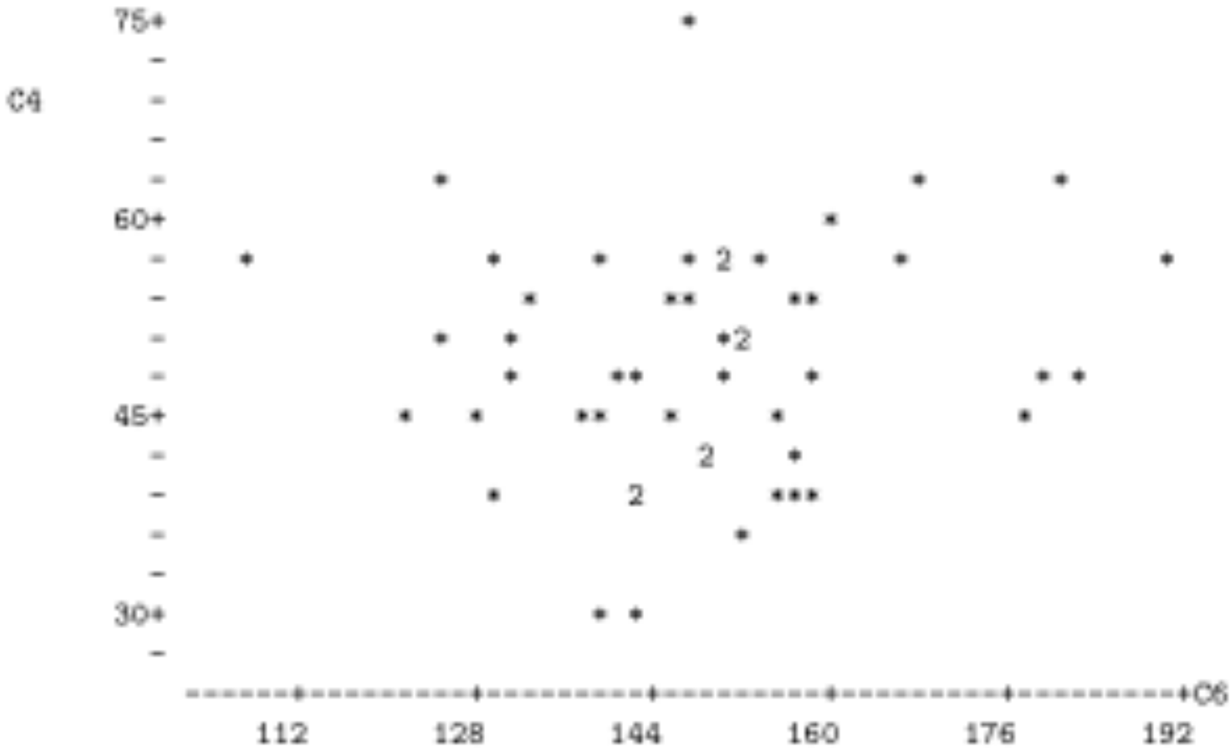
- $-1 \le r \le 1$
- $r = +1$ indicates a perfect positive linear relationship. All the points are exactly on a line with a positive slope.
- $r = -1$ indicates a perfect negative linear relationship. All the points are exactly on a line with a negative slope.
- $r = 0$ means no *linear* relationship (curve possible). Slope of least squares line = 0
- r^2 = proportion of variation explained

r = 0.004



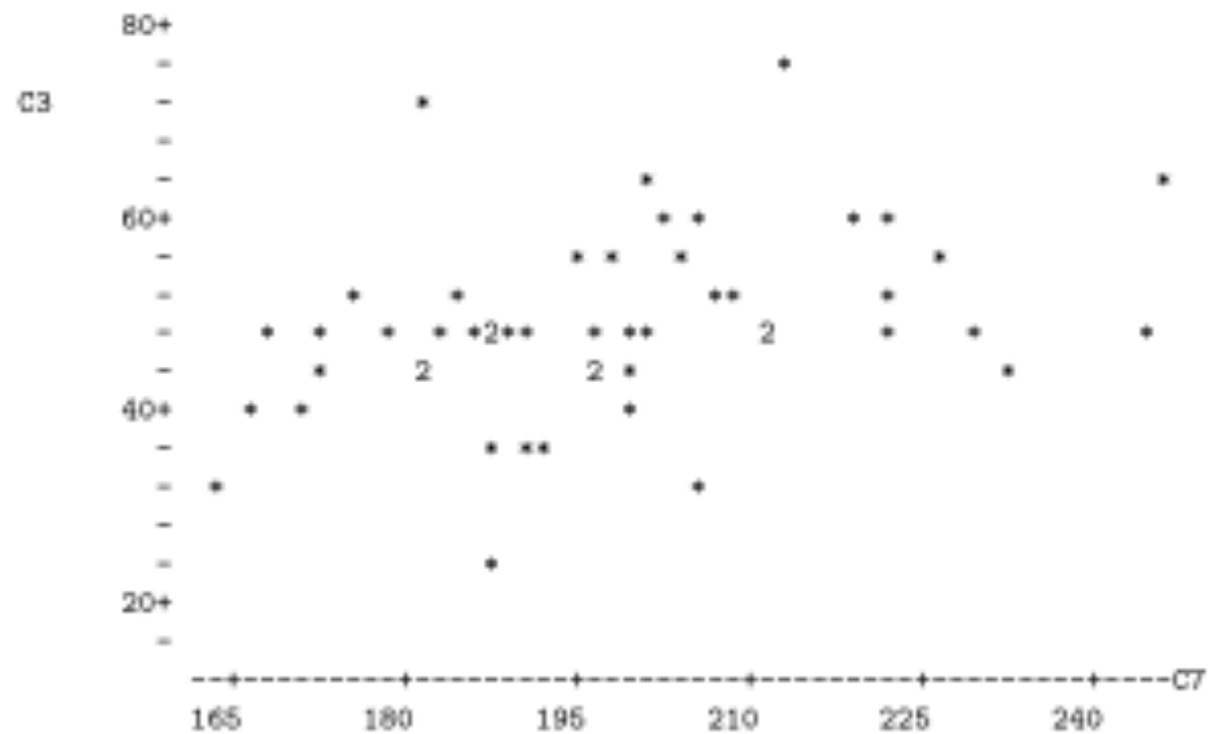
Correlation of C1 and C3 = 0.004

r = 0.112



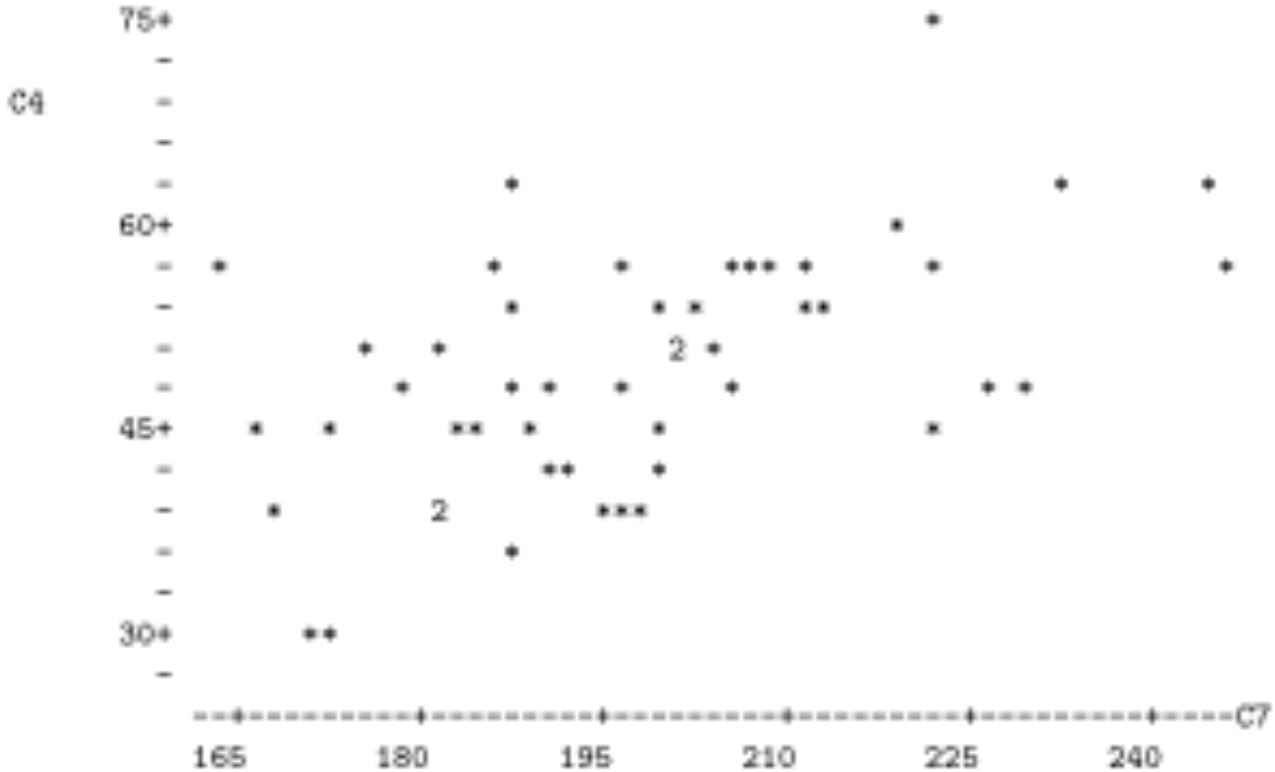
Correlation of C4 and C6 = 0.112

r = 0.368



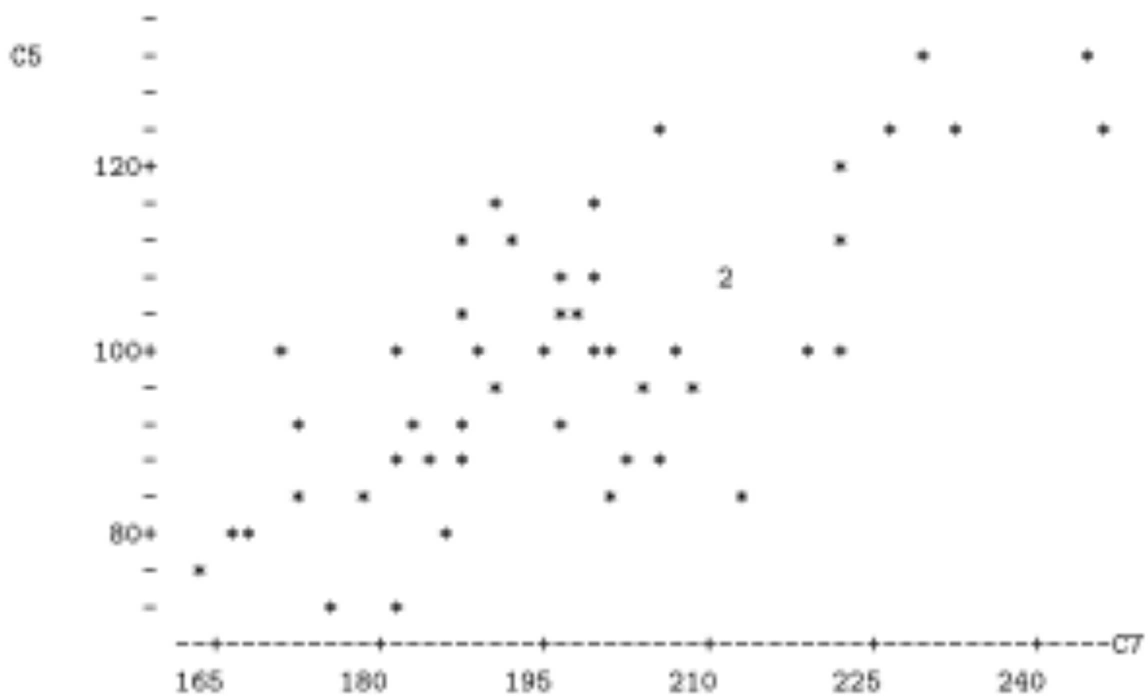
Correlation of C3 and C7 = 0.368

r = 0.547



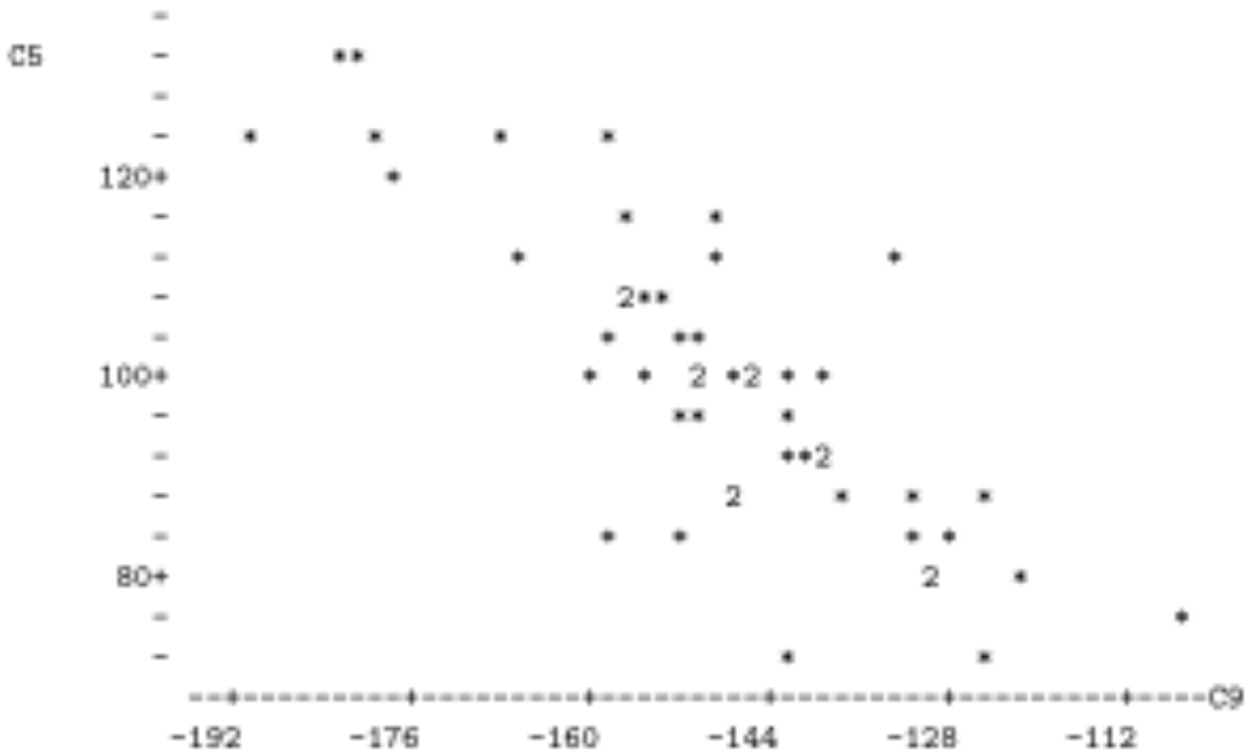
Correlation of C4 and C7 = 0.547

r = 0.733



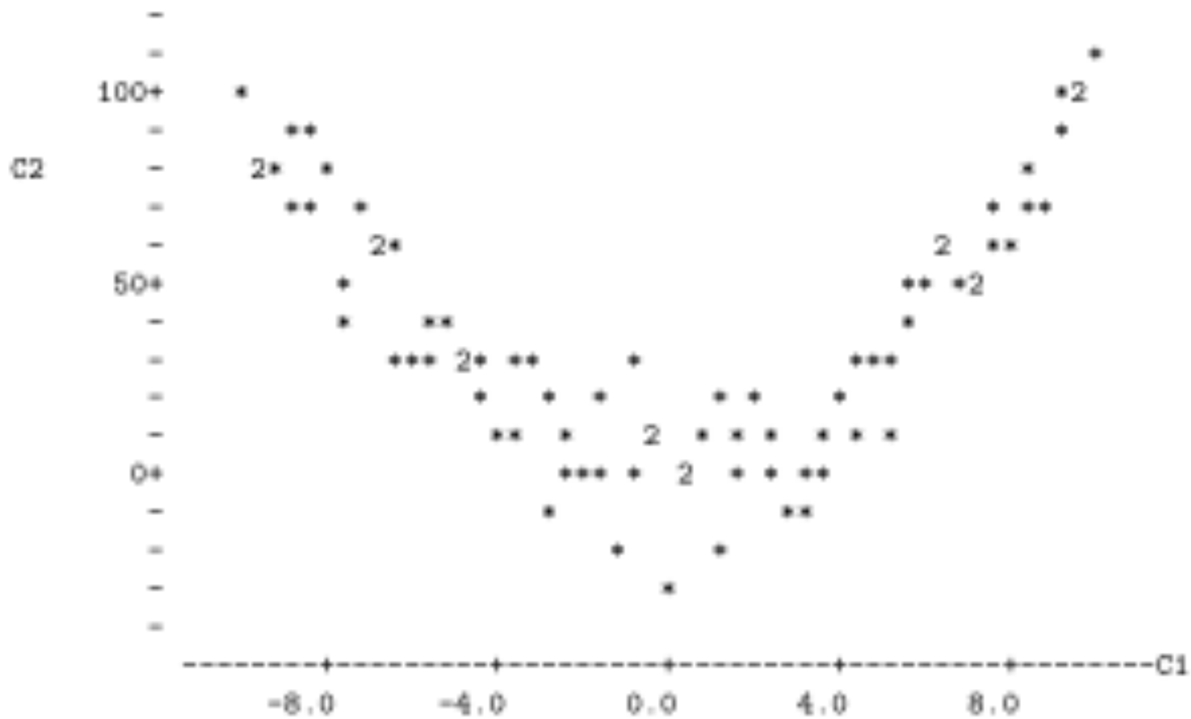
Correlation of C5 and C7 = 0.733

r = - 0.822



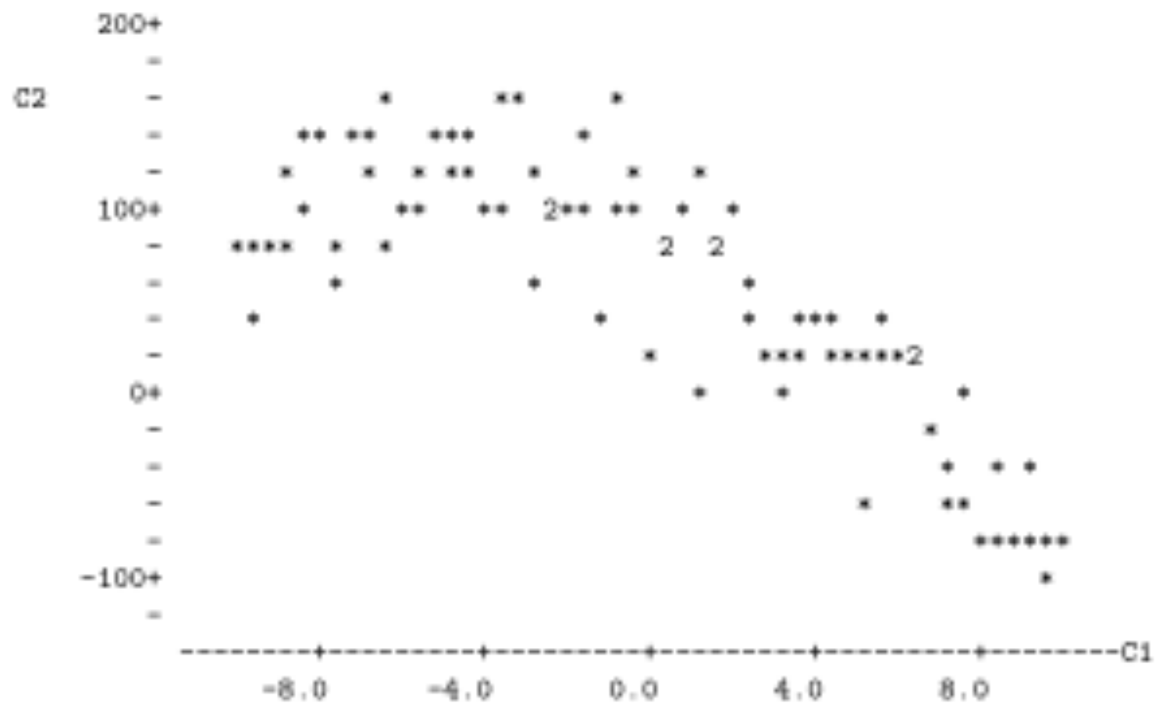
Correlation of C5 and C9 = -0.822

r = 0.025



Correlation of C1 and C2 = 0.025

r = - 0.811



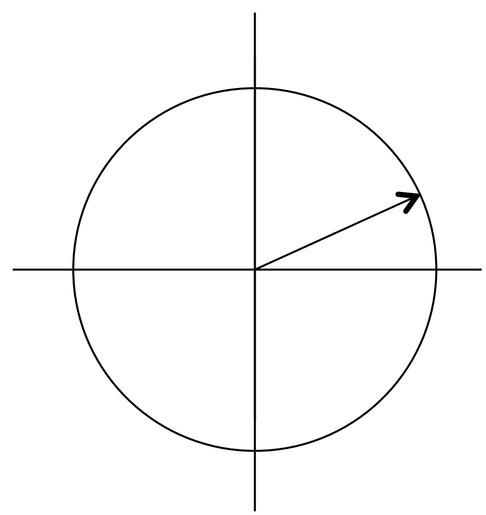
Correlation of C1 and C2 = -0.811

Why -1 ≤ r ≤ 1 ?

- $r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$

- $\cos(\theta) = \frac{\mathbf{a}'\mathbf{b}}{|\mathbf{a}|\ |\mathbf{b}|}$

$$= \frac{\mathbf{a}'\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{a}\ \mathbf{b}'\mathbf{b}}}$$



A Statistical Model

Independently for $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where

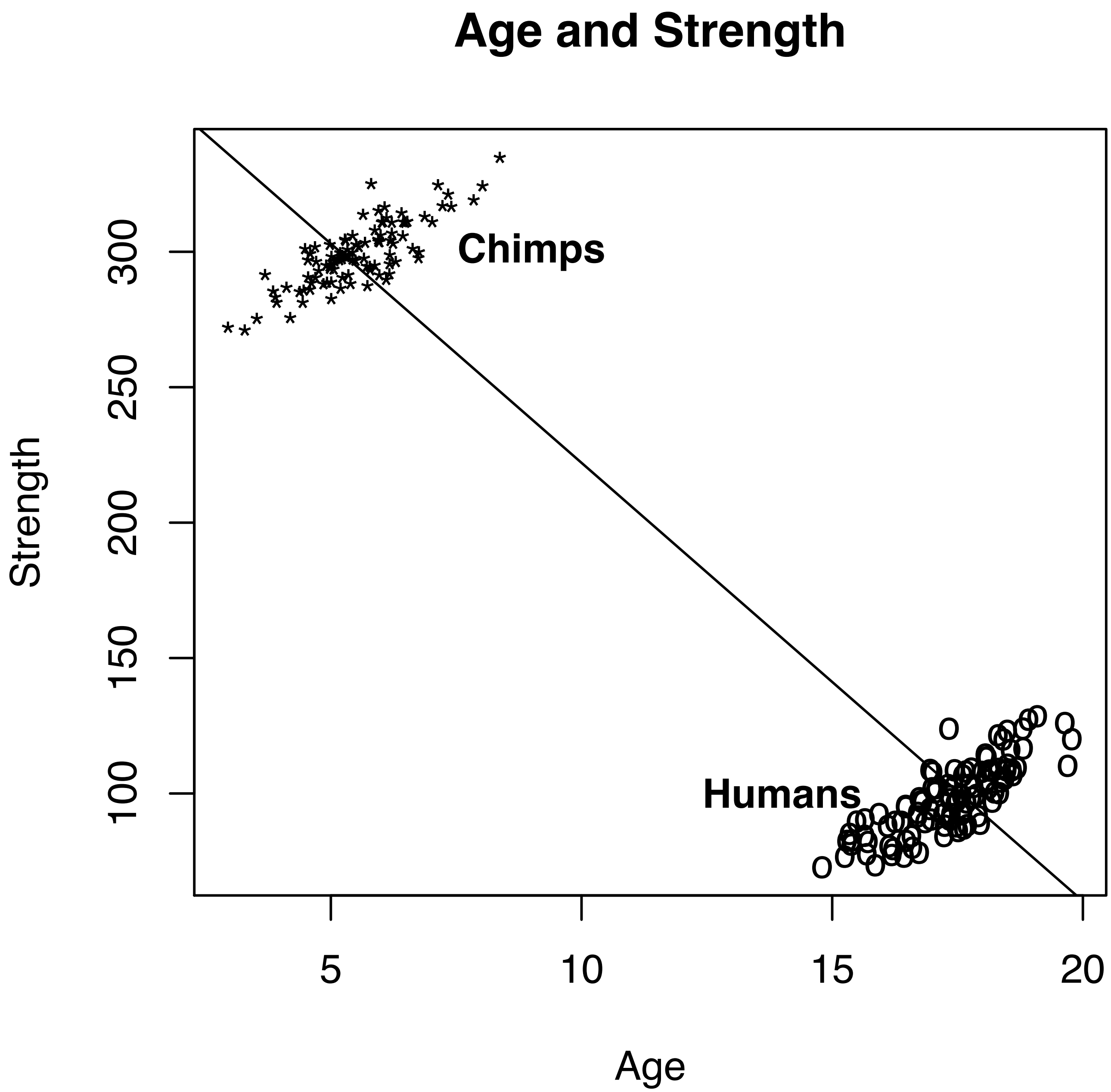
$x_1, \ldots, x_n$ are observed, known constants

$\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables

β_0, β_1 and σ^2 are unknown constants with $\sigma^2 > 0$.

One Independent Variable at a Time Can Produce Misleading Results

- The standard elementary methods all have a single independent variable (at most), so they should be used with caution in practice.
- Example: Artificial and extreme, to make a point:
- Suppose the correlation between Age and Strength is $r = -0.96$


Age and Strength
Chimps
Humans
Strength
300
250
200
150
100
Age
5
10
15
20

Need *multiple* regression

Multiple regression in scalar form

For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$, where

x_{ij} are observed, known constants

$\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables

β_j and σ^2 are unknown constants with $\sigma^2 > 0$.

Multiple regression in matrix form

$$\mathbf{y} \quad = \quad X \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\epsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 14.2 & \cdots & 1 \\ 1 & 11.9 & \cdots & 0 \\ 1 & 3.7 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.2 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

where

X is an $n \times (k+1)$ matrix of observed constants

$\boldsymbol{\beta}$ is a $(k+1) \times 1$ matrix of unknown constants

$\boldsymbol{\epsilon}$ is multivariate normal. Write $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

σ^2 is an unknown constant

So we need

- Matrix algebra
- Random vectors, especially multivariate normal
- Software to do the computation

Copyright Information