Correlation

Sunthud Pornprasertmanit Chulalongkorn University

Correlation and Regression Question

They differ in respect to

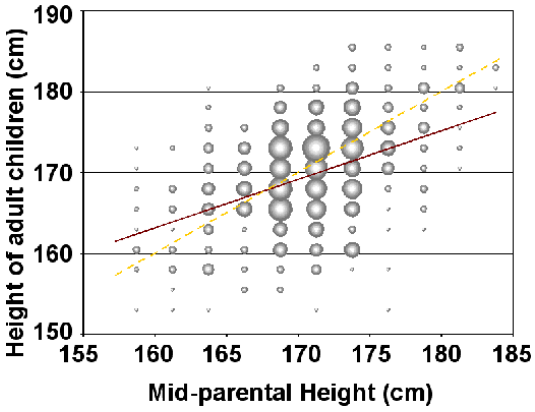
1) The nature of the variables
2) Use of random assignment
3) The researcher's principle interest
4) Kinds of conclusion that can be drawn

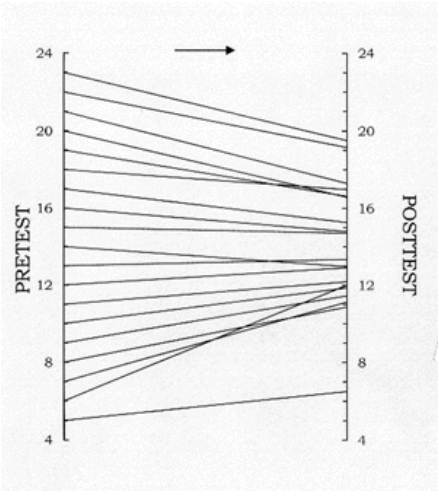
Difference between correlation and regression analysis

History

Sir Francis Galton:

- Regression toward the mean
  - Error of measurement
  - The distribution of repeat measurement of objects selected on the basic of a previous extreme value will not lie symmetrically around the most common value, but will be skewed towards the mean of the population, such as height.
- Regression line

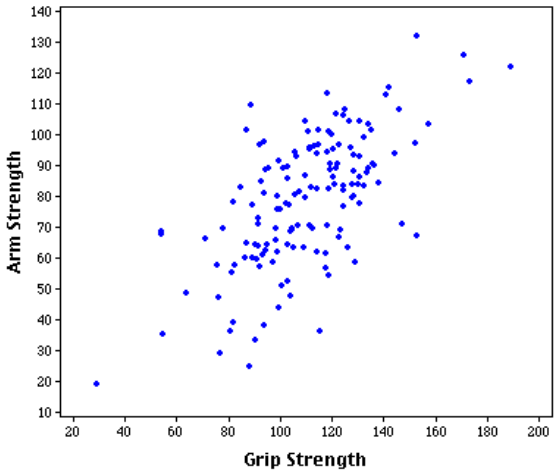




Author Note
This article was written in June 2007 for teaching in Introduction to Statistics in Psychology Class, Faculty of Psychology, Chulalongkorn University
Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com

Scatterplot

Scatterplot or scatter diagram



Changing axis to Mean X and Mean Y

$$X' = X - \bar{X}$$

$$Y' = Y - \bar{Y}$$

Cross product

$$(X - \bar{X})(Y - \bar{Y})$$

Sum of cross products

$$\sum (X - \bar{X})(Y - \bar{Y})$$

Covariance

$$S_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}$$

Correlation

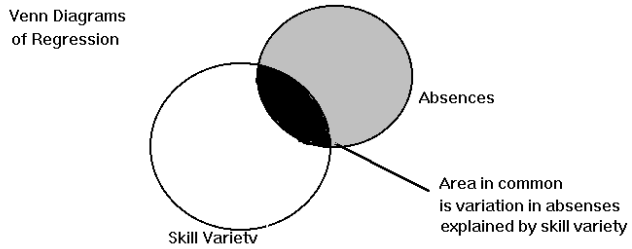
$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

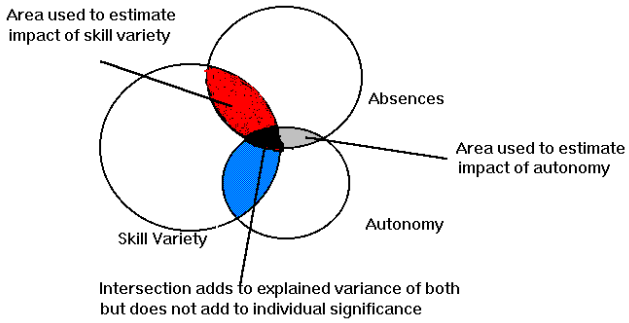
Index of correlation

Pearson product moment correlation

Correlation coefficient (*r* or ρ)

- Direction
- Strength of relationship
- Coefficient of determination (Variance explained)
- Coefficient of nondetermination





Note: Sum of areas from simple will be greater than sum of partial by the size of the intersection

Causal and concomitant relationship

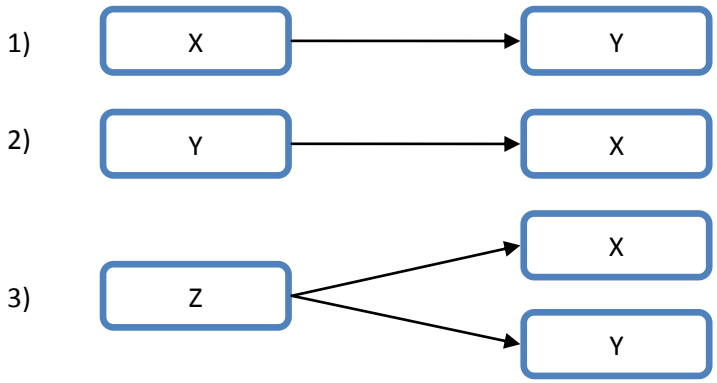
The conditions make causal relation

1) X precedes Y in time
2) Some mechanism explained
3) Change in X is accompanied by change in Y
4) Effect X on Y cannot be explained by other variables

The research design that can prove casual relationship is experimental design.

Interpretation of correlation

1) X causes Y.
2) Y causes X.
3) Z causes both X and Y.



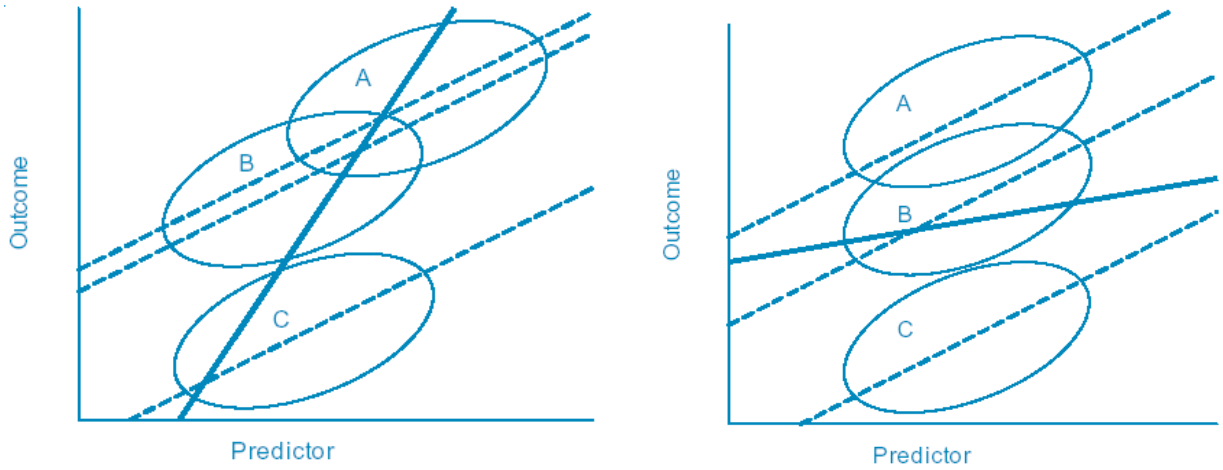
Factor that affect the size of a correlation coefficient

Nonlinear relationship

Range restriction

$$\tilde{r}_{YX} = \frac{r_{YX_C}(\frac{SD_X}{SD_{X_C}})}{\sqrt{1 + r^2_{YX_C}\left(\left(\frac{SD^2_X}{SD^2_{X_C}}\right) - 1\right)}}$$

Spurious effect due to subgroups with different means and standard deviations



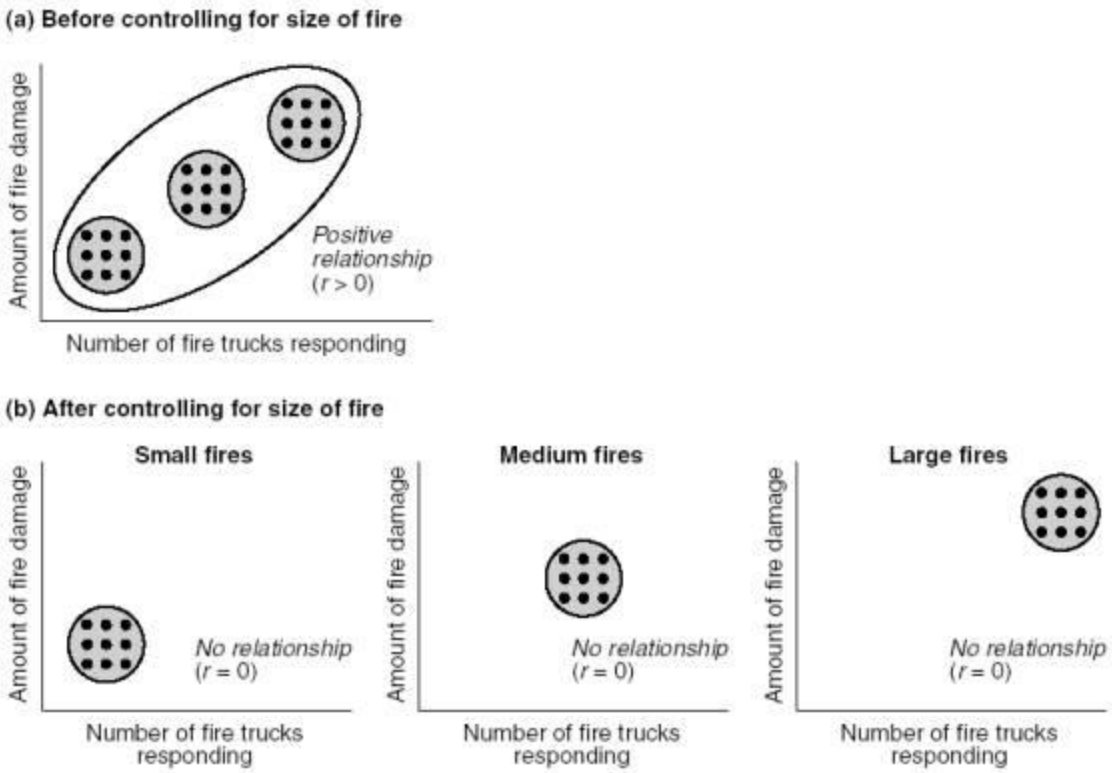
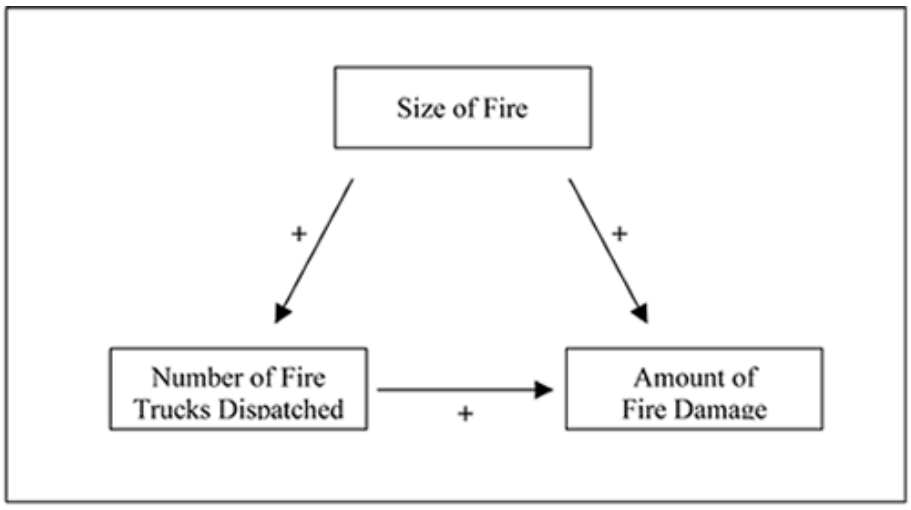


FIGURE 11.2 Relationship between amount of fire damage and number of trucks responding before and after controlling for size of fire. We controlled for size of fire by examining the original relationship at different levels of size of fire. The original relationship disappears.

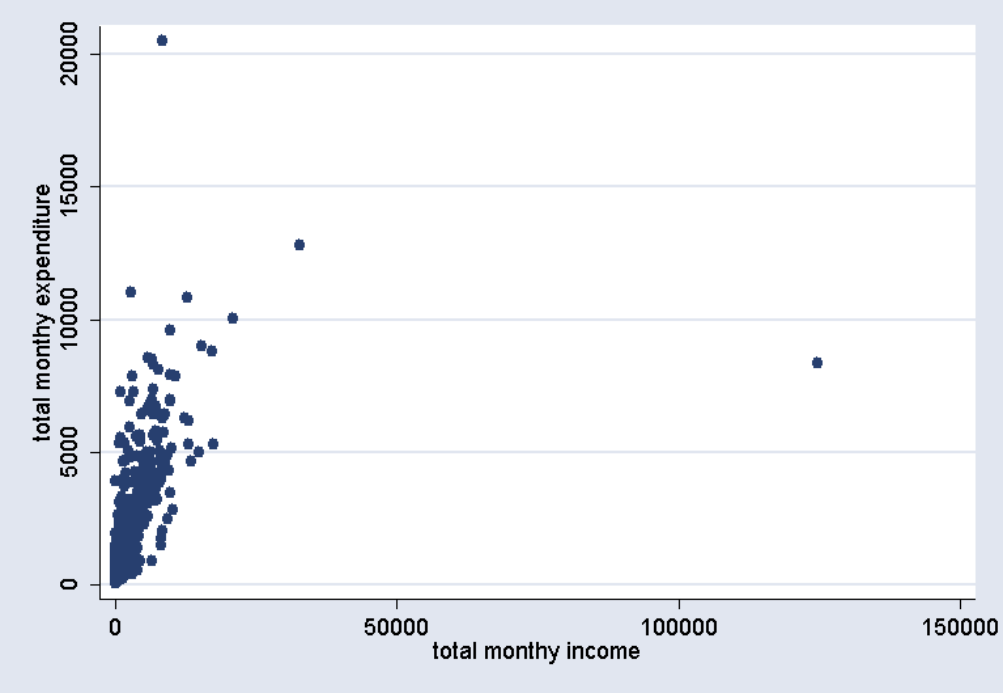


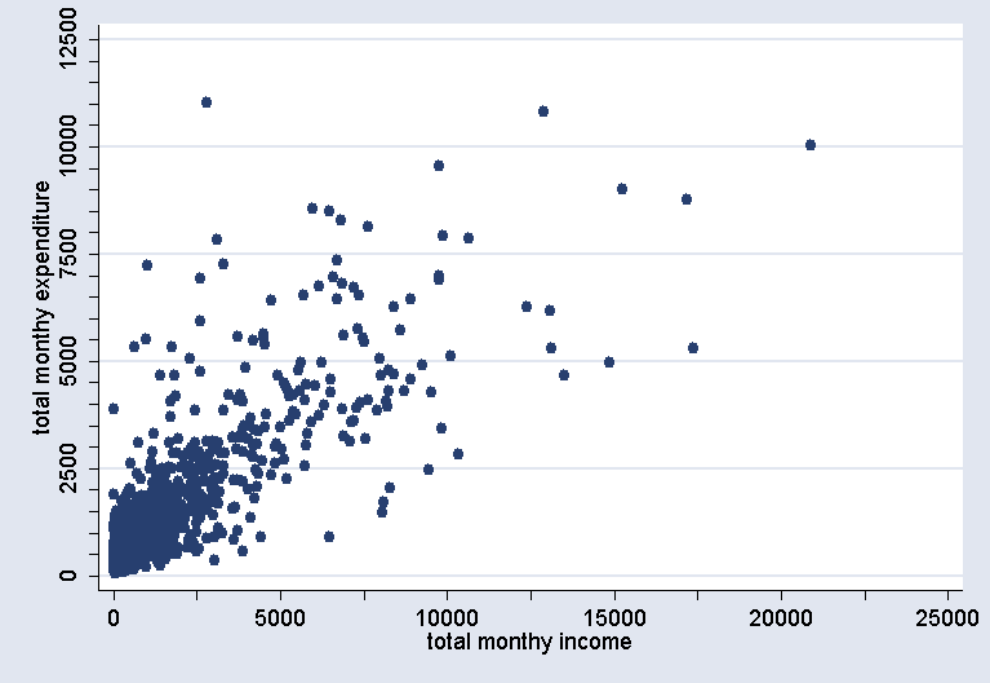
Spurious Relationship.
The "+" sign indicates a positive, significant relationship.

Heteroscadasticity

Discontinuous Distribution

Outlier





Correlation change from .52 to .79.

Correlation Index

Spearman rank correlation

- How to rank data
- Monotonic relationship

$$r_s = 1 - \frac{6\sum(R_{X_i} - R_{Y_i})^2}{n(n^2-1)}$$

Point-biserial correlation

$$r_{pb} = \frac{(M_{Y_i} - M_{Y_0})\sqrt{PQ}}{SD_Y}$$

Phi coefficient

	Politician A	Politician B
Male	A	B
Female	C	D

$$r_{\emptyset} = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Moderator