Psychology 371: Personality Research
Part II: Psychometric Theory

Part I: Overview and the correlation coefficient

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

April, 2019

Outline

What is psychometrics?

In physical science a first essential step in the direction of learning any subject is to find principles of numerical reckoning and methods for practicably measuring some quality connected with it. I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the stage of science, whatever the matter may be. (Thomsom, 1891)

Taken from Michell (2003) in his critique of psychometrics: Michell, J. The Quantitative Imperative: Positivism, Naïve Realism and the Place of Qualitative Methods in Psychology, Theory & Psychology, Vol. 13, No. 1, 5-31 (2003)

What is psychometrics?

The character which shapes our conduct is a definite and durable 'something', and therefore ... it is reasonable to attempt to measure it. (Galton, 1884)

The history of science is the history of measurement" (J. M. Cattell, 1893)

Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality (E.L. Thorndike, 1918)

What is psychometrics?

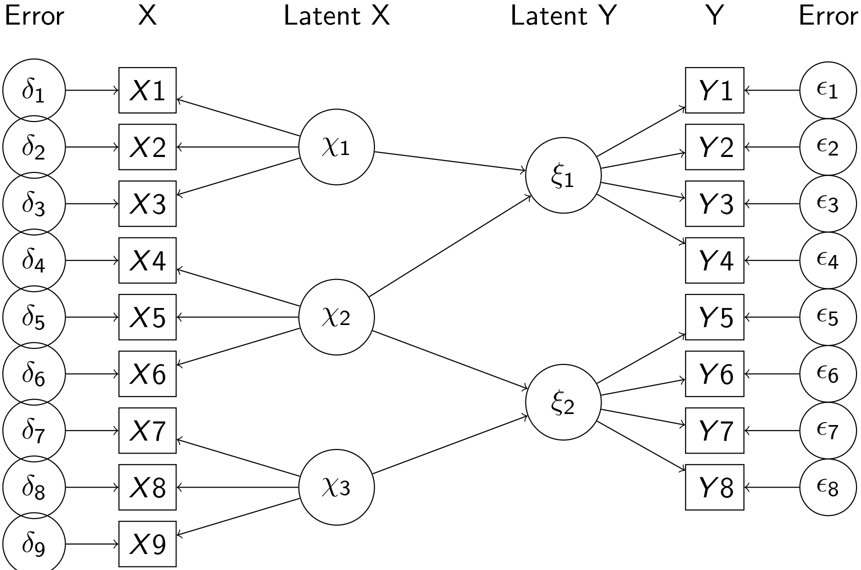
We hardly recognize a subject as scientific if measurement is not one of its tools (Boring, 1929)

There is yet another [method] so vital that, if lacking it, any study is thought ... not be scientific in the full sense of the word. This further an crucial method is that of measurement. (Spearman, 1937)

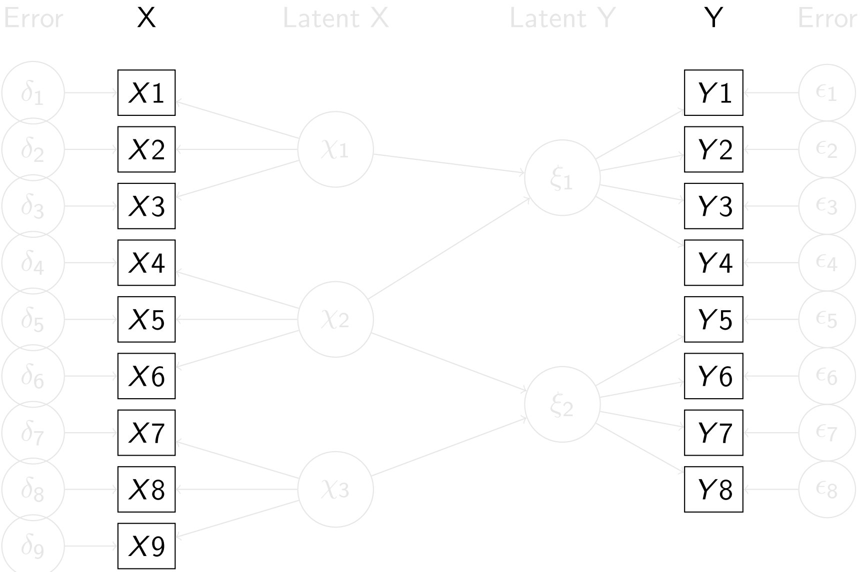
One's knowledge of science begins when he can measure what he is speaking about and express in numbers (Eysenck, 1973)

Psychometrics: the assigning of numbers to observed psychological phenomena and to unobserved concepts. Evaluation of the fit of theoretical models to empirical data.

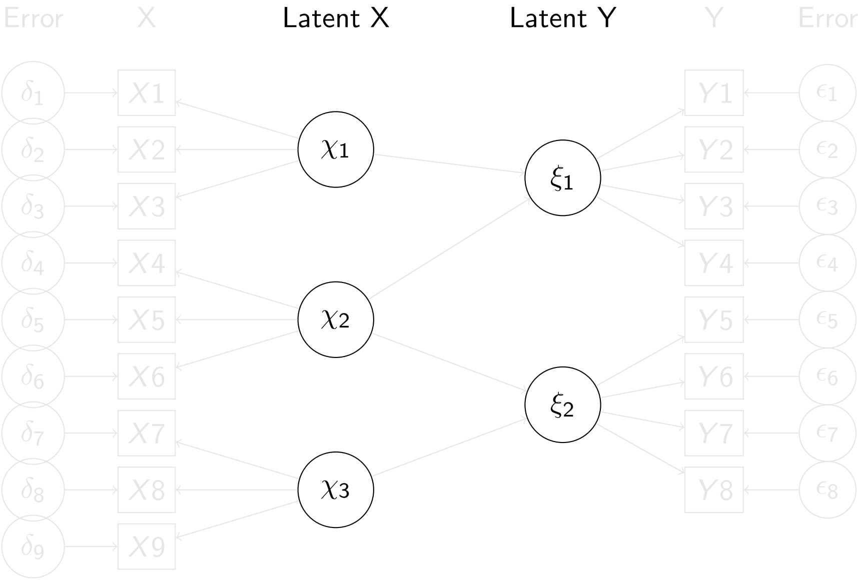
Psychometric Theory: A conceptual Overview



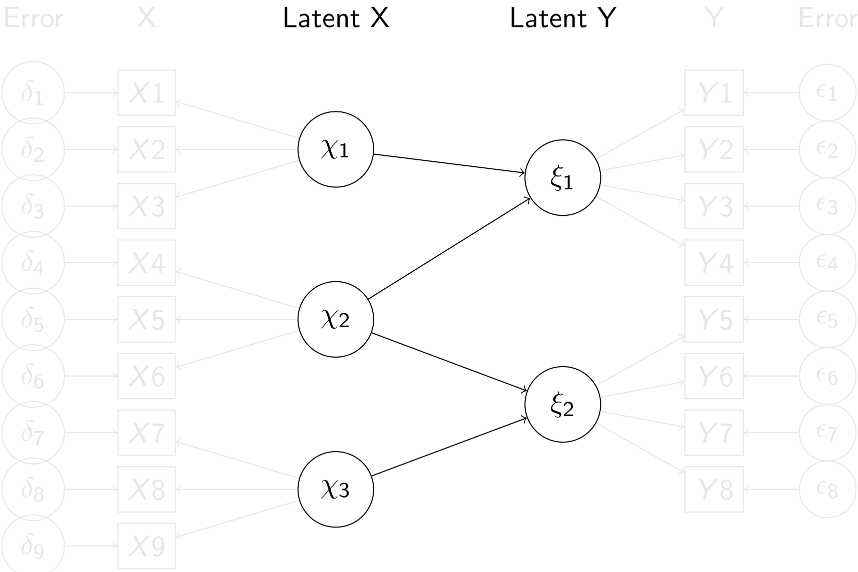
Observed Variables



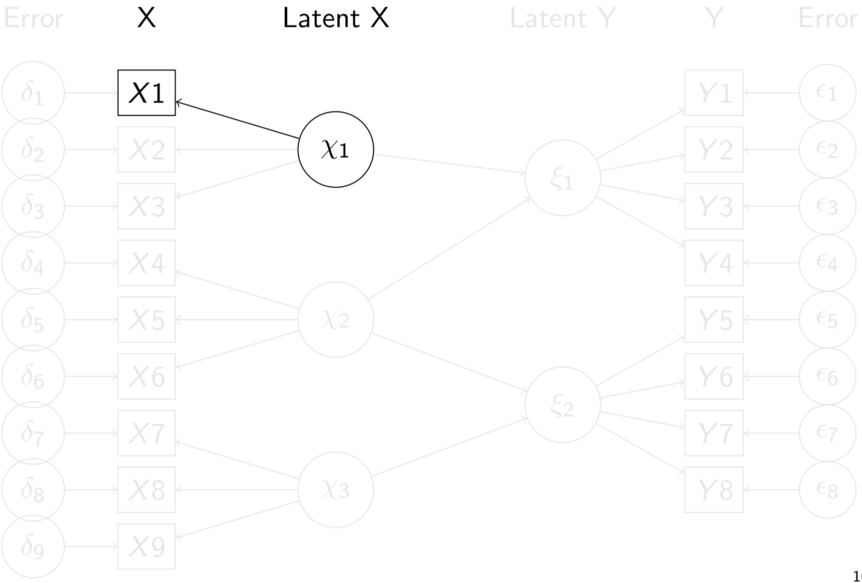
Latent Variables



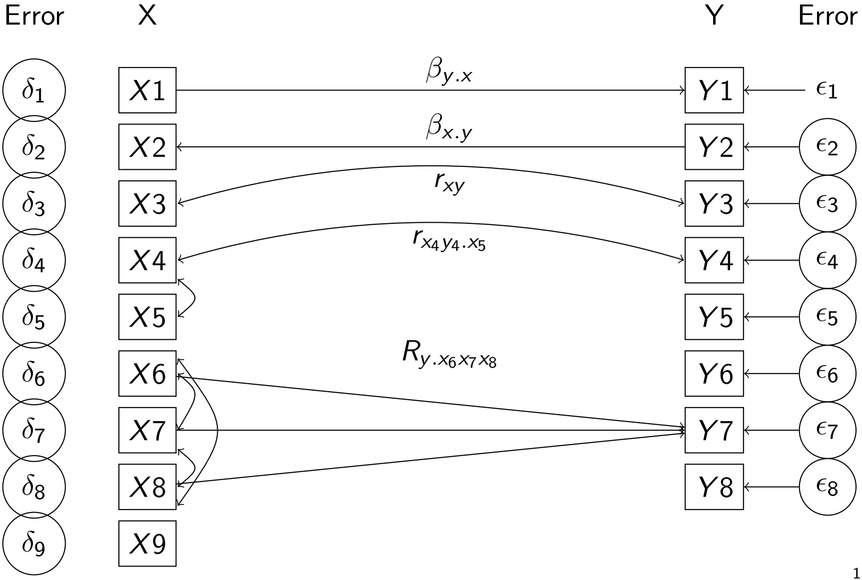
Theory


Error
X
Latent X
Latent Y
Y
Error
δ_1
δ_2
δ_3
δ_4
δ_5
δ_6
δ_7
δ_8
δ_9
X1
X2
X3
X4
X5
X6
X7
X8
X9
χ_1
χ_2
χ_3
ξ_1
ξ_2
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
ϵ_1
ϵ_2
ϵ_3
ϵ_4
ϵ_5
ϵ_6
ϵ_7
ϵ_8

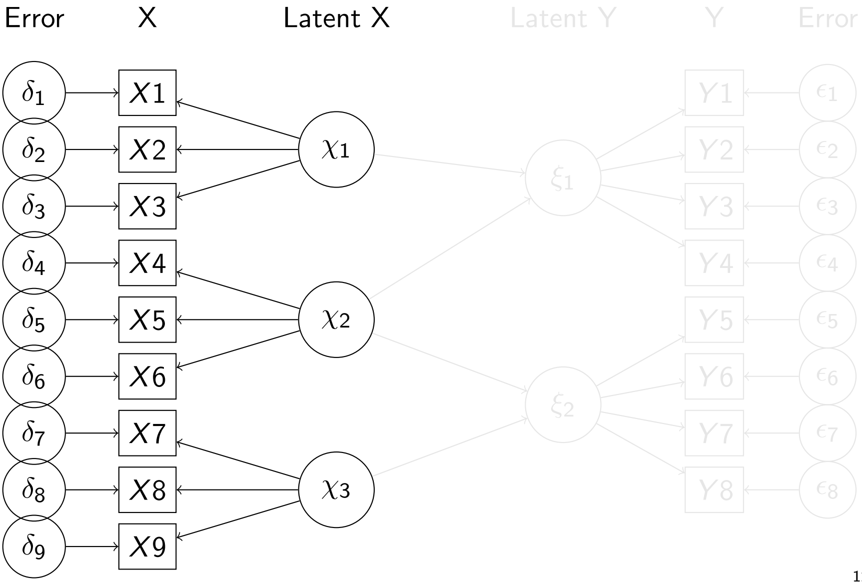
A theory of data and fundamentals of scaling



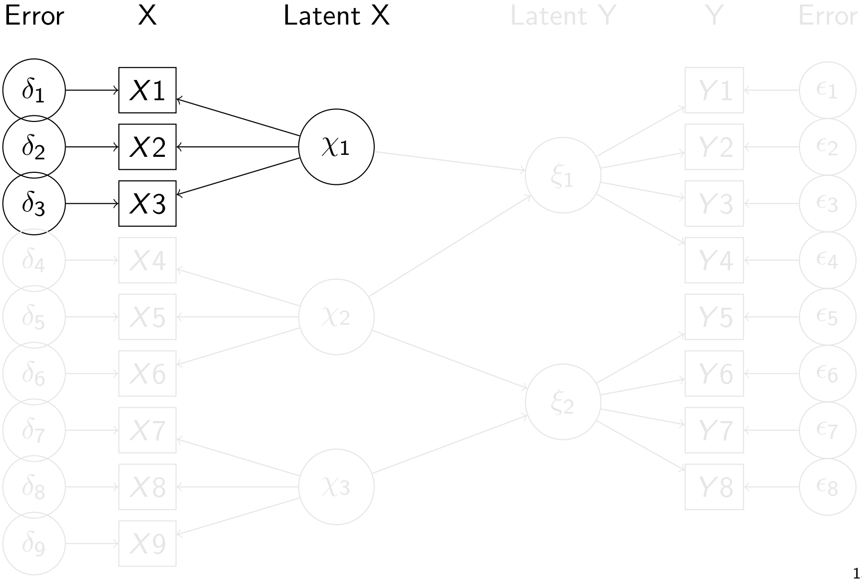
Correlation, Regression, Partial Correlation, Multiple Regression



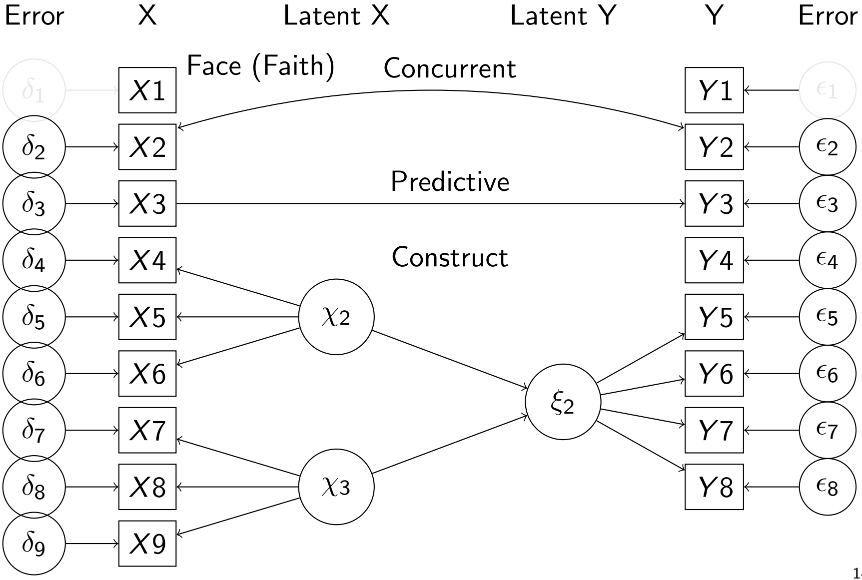
Measurement: A latent variable approach.



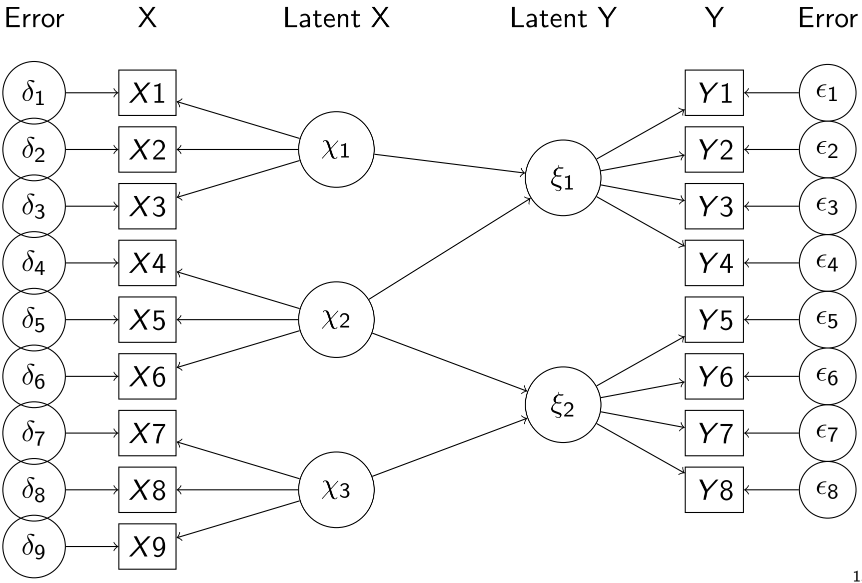
Reliability: How well does a test reflect one latent trait?



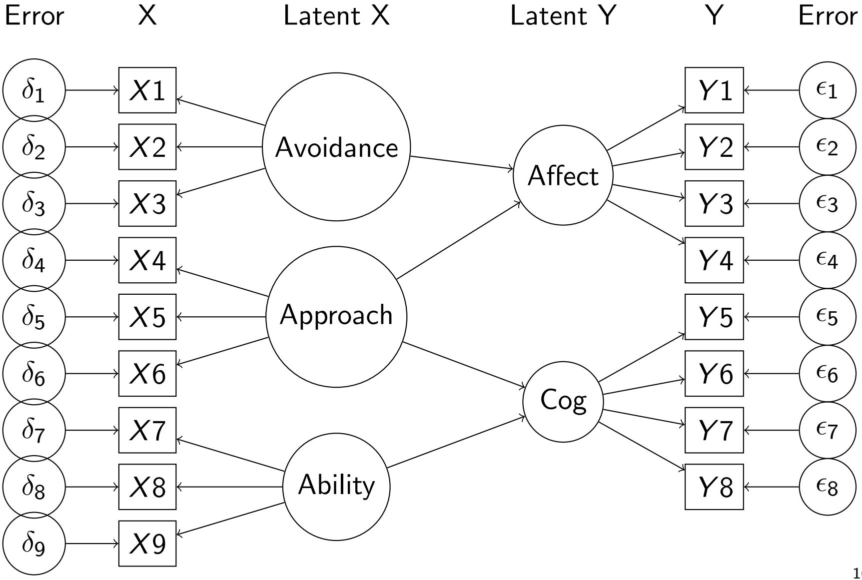
Face, Concurrent, Predictive, Consruct



Psychometric Theory: Data, Measurement, Theory



Psychometric Theory: Data, Measurement, Theory



Data = Model + Residual

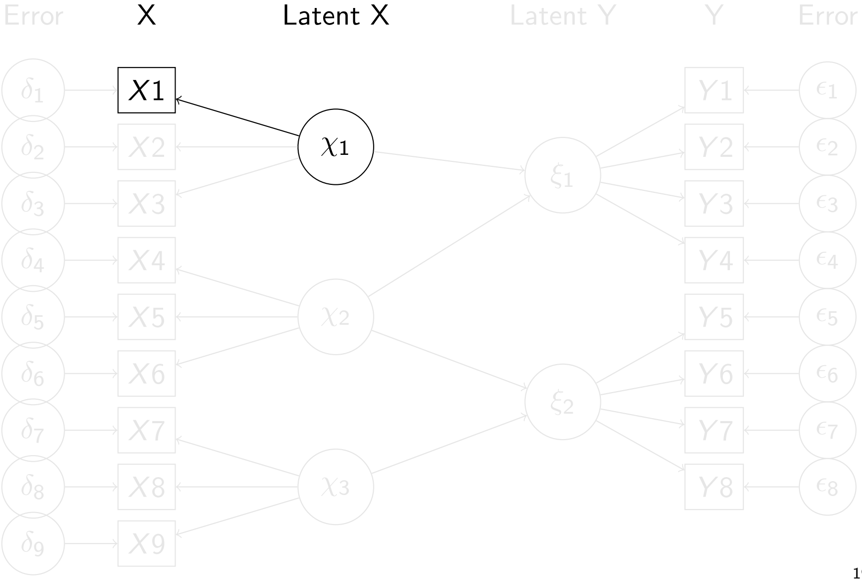
- The fundamental equations of statistics are that
  - Data = Model + Residual
  - Residual = Data - Model
- The problem is to specify the model and then evaluate the fit of the model to the data as compared to other models
  - Fit = f(Data, Residual)
  - Typically: $Fit = f(1 - \frac{Residual^2}{Data^2})$
  - $Fit = f(1 - \frac{(Data - Model)^2}{Data^2})$
- Even for something as simple as the mean is a model of the data. The residual left over after we remove the mean is the variance.

Psychometrics as model estimation and model fitting

We will explore a number of models

1. Modeling the process of data collection and of scaling
   - $X = f(\theta)$
   - How to measure X, properties of the function f.
2. Correlation and Regression
   - $Y = \beta X$
   - $R_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
3. Factor Analysis and Principal Components Analysis
   - $R = FF' + U^2 \quad R = CC'$
4. Reliability $\rho_{xx} = \frac{\sigma^2_\theta}{\sigma^2_X}$
5. Item Response Theory
   - $p(X|\theta, \delta) = f(\theta - \delta)$
6. Structural Equation Modeling
   - $\rho_{yy} Y = \beta \rho_{xx} X$

A theory of data and fundamentals of scaling



Consider the following numbers, what do they represent?

Table: Numbers without context are meaningless. What do these number represent? Which of these numbers represent the same thing?

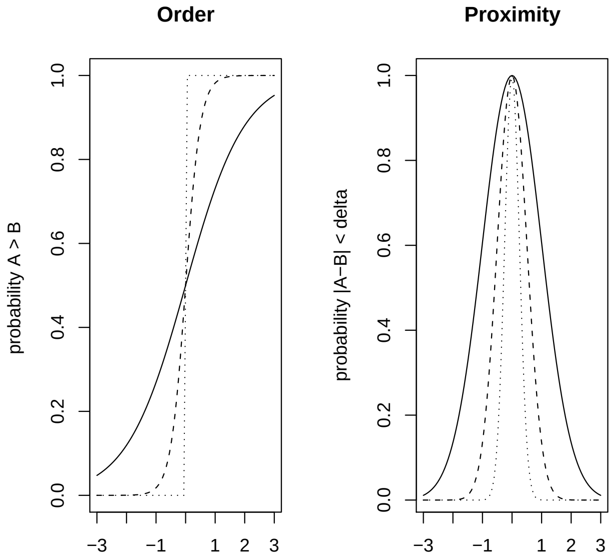
2.718281828459045	3.141592653589793116
24	86,400
37	98.7
365.25	365.25636305
31,557,600	31,558,150
3,412.1416	.4046856422
299,792,458	$6.022141 * 10^{23}$
42	X

Clyde Coombs and the Theory of Data

1. O = the set of objects
   - $O = \{o_i, o_j \dots o_n\}$
2. S = the set of Individuals
   - $S = \{s_i, s_j \dots s_n\}$
3. Two comparison operations
   - order ($x > y$)
   - proximity ($|x - y| < \epsilon$
4. Two types of comparisons
   - Single dyads
     - (s_i, s_j) (s_i, o_j) (o_i, o_j)
   - Pairs of dyads
     - $(s_i, s_j)(s_k, s_l)$ $(s_i, o_j)(s_k, o_l)$ $(o_i, o_j)(o_k, o_l)$

Coombs (1964)

2 types of comparisons: Monotone ordering and single peak proximity



Theory of Data and types of measures

Table: The theory of data provides a 3 x 2 x 2 taxonomy for various types of measures

Elements of Dyad	Number of Dyads	Comparison	Name
People x People	1	Order	Tournament rankings
People x People	1	Proximity	Social Networks
Objects x Objects	1	Order	Scaling
Objects x Objects	1	Proximity	Similarities
People x Objects	1	Order	Ability Measurement
People x Objects	1	Proximity	Attitude Measurement
People x People	2	Order	Tournament rankings
People x People	2	Proximity	Social Networks
Objects x Objects	2	Order	Scaling
Objects x Objects	2	Proximity	Multidimensional scaling
People x Objects	2	Order	Ability Comparisons
People x Objects	2	Proximity	Preferential Choice
People x Objects x Objects	2	Proximity	Individual Differences in Multidimensional Scaling

Tournaments to order people (or teams)

1. Goal is to order the players by outcome to predict future outcomes
2. Complete Round Robin comparisons
    - Everyone plays everyone
    - Requires $N * (N - 1)/2$ matches
    - How do you scale the results?
3. Partial Tournaments – Seeding and group play
    - World Cup
    - NCAA basketball
    - Is the winner really the best?
    - Can you predict other matches

Friendship as proximity

1. Chess or football provides a ranking based upon an ordering relationship ($p_i > p_j$).
2. Alternatively, friendship groups are based upon closeness ($|p_i - p_j| < \delta$)
   2.1 Do you know person j?
   2.2 Do you like person j? or as an alternative:
   2.3 Please list all your friends in this class (and is j included on the list)
   2.4 Would you be interested in having a date with person j?
   2.5 Would you like to have sex with person j?
   2.6 Would you marry person j?
3. Typically such data will be a rectangular matrix for there are asymmetries in closeness.

Moh's hardness scale provides rank orders of hardness

Table: Mohs' scale of mineral hardness. An object is said to be harder than X if it scratches X. Also included are measures of relative hardness using a sclerometer (for the hardest of the planes if there is a ansiotropy or variation between the planes) which shows the non-linearity of the Mohs scale (Burchard, 2004).

Mohs Hardness	Mineral	Scratch hardness
1	Talc	.59
2	Gypsum	.61
3	Calcite	3.44
4	Fluorite	3.05
5	Apaptite	5.2
6	Orthoclase Feldspar	37.2
7	Quartz	100
8	Topaz	121
9	Corundum	949
10	Diamond	85,300

Ordering based upon external measures

Table: The Beaufort scale of wind intensity is an early example of a scale with roughly equal units that is observationally based. Although the units are roughly in equal steps of wind speed in nautical miles/hour (knots), the force of the wind is not linear with this scale, but rather varies as the square of the velocity.

Force	Wind (Knots)	WMO Classification	Appearance of Wind Effects
0	Less than 1	Calm	Sea surface smooth and mirror-like
1	1-3	Light Air	Scaly ripples, no foam crests
2	4-6	Light Breeze	Small wavelets, crests glassy, no breaking
3	7-10	Gentle Breeze	Large wavelets, crests begin to break, scattered whitecaps
4	11-16	Moderate Breeze	Small waves 1-4 ft. becoming longer, numerous whitecaps
5	17-21	Fresh Breeze	Moderate waves 4-8 ft taking longer form, many whitecaps, some spray
6	22-27	Strong Breeze	Larger waves 8-13 ft, whitecaps common more spray
7	28-33	Near Gale	Sea heaps up, waves 13-20 ft, white foam streaks off breakers
8	34-40	Gale Moderately	high (13-20 ft) waves of greater length, edges of crests begin to break into spindrift, foam blown in streaks
9	41-47	Strong Gale	High waves (20 ft), sea begins to roll, dense streaks of foam, spray may reduce visibility
10	48-55	Storm	Very high waves (20-30 ft) with overhanging crests, sea white with densely blown foam, heavy rolling, lowered visibility
11	56-63	Violent Storm	Exceptionally high (30-45 ft) waves, foam patches cover sea, visibility more reduced
12	64+	Hurricane	Air filled with foam, waves over 45 ft, sea completely white with driving spray, visibility greatly reduced

Models of scaling objects

1. Assume each object $(a, b, ...z)$ has a scale value $(A, B, ...Z)$ with some noise for each measurement.
2. Probability of $A > B$ increases with difference between a and b
3. $P(A > B) = f(a - b)$
4. Can we find a function, f, such that equal differences in the latent variable (a, b, c) lead to equal differences in the observed variable?
5. Several alternatives
   - Direct scaling on some attribute dimension (simple but flawed)
   - Indirect scaling by paired comparisons (more complicated but probably better)

Scaling of Objects: O x O comparisons

1. Typical object scaling is concerned with order or location of objects
2. Subjects are assumed to be random replicates of each other, differing only as a source of noise
3. Absolute scaling techniques
   - Grant Proposals: 1 to 5
   - "On a scale from 1 to 10" this [object] is a X?
   - If A is 1 and B is 10, then what is C?
   - College rankings based upon selectivity
   - College rankings based upon "yield"
   - Zagat ratings of restaurants
   - A - F grading of papers

Absolute scaling: difficulties

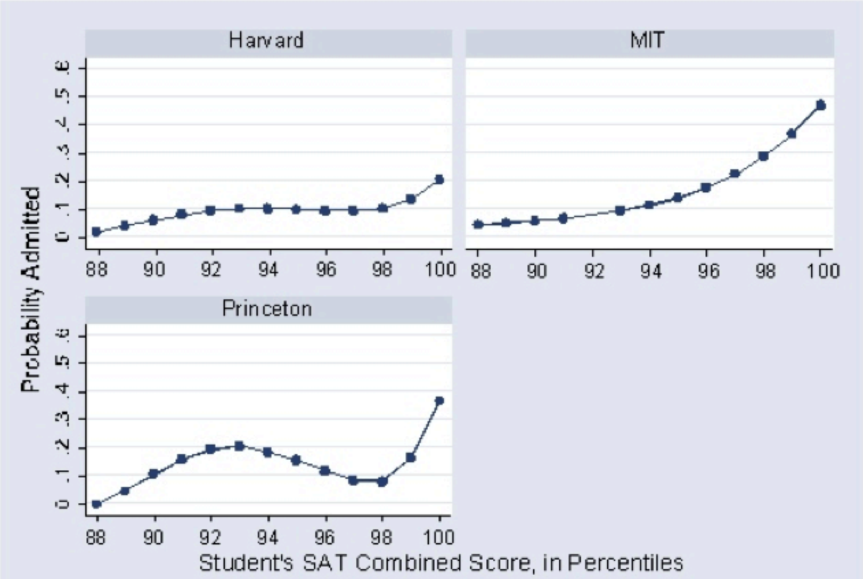
1. "On a scale from 1 to 10" this [object] is a X?
    - sensitive to context effects
    - what if a new object appears?
    - Need unbounded scale
2. If A is 1 and B is 10, then what is C?
    - results will depend upon A, B

Absolute scaling: artifacts

1. College rankings based upon selectivity
   - accept/applied
   - encourage less able to apply
2. College rankings based upon "yield"
   - matriculate/accepted
   - early admissions guarantee matriculation
   - don't accept students who will not attend
3. Proposed solution: college choice as a tournament
   - Consider all schools that accept a student
   - Which school does he/she choose?

Avery, Glickman, Hoxby & Metrick (2013)

A revealed preference ordering Avery et al. (2013)



A revealed preference ordering Avery et al. (2013)

TABLE III

A REVEALED PREFERENCE RANKING OF COLLEGES BASED ON MATRICULATION DECISIONS

Rank Based on Matriculation (with Covariates)	College Name	Theta	Implied Prob. of "Winning" vs. College Listed... 1 Row Below	Implied Prob. of "Winning" vs. College Listed... 10 Rows Below	Rank Based on Matriculation (no Covariates)
1	Harvard University	9.13	0.59	0.93	1
2	Caltech	8.77	0.56	0.92	3
3	Yale University	8.52	0.59	0.92	2
4	MIT	8.16	0.51	0.89	5
5	Stanford University	8.11	0.52	0.90	4
6	Princeton University	8.02	0.73	0.90	6
7	Brown University	7.01	0.56	0.78	7
8	Columbia University	6.77	0.54	0.73	8
9	Amherst College	6.61	0.51	0.71	9
10	Dartmouth	6.57	0.52	0.72	10
11	Wellesley College	6.51	0.53	0.71	12
12	University of Pennsylvania	6.39	0.56	0.71	11

Weber-Fechner Law and non-linearity of scales

1. Early studies of psychophysics by Weber (1834b,a) and subsequently Fechner (1860) demonstrated that the human perceptual system does not perceive stimulus intensity as a linear function of the physical input.
2. The basic paradigm was to compare one weight with another that differed by amount Δ, e.g., compare a 10 gram weight with an 11, 12, and 13 gram weight, or a 10 kg weight with a 11, 12, or 13 kg weight.
3. What was the Δ that was just detectable? The finding was that the perceived intensity follows a logarithmic function.
4. Examining the magnitude of the "*just noticeable difference*" or *JND*, Weber (1834b) found that

$$JND = \frac{\Delta Intensity}{Intensity} = constant. \tag{1}$$

Weber-Fechner Law and non-linearity of scales

1. An example of a logarithmic scale of intensity is the decibel measure of sound intensity.
2. Sound Pressure Level expressed in decibels (dB) of the root mean square observed sound pressure, P_o (in Pascals) is

$$L_p = 20 Log_{10} \frac{P_o}{P_{ref}} \tag{2}$$

3. where the reference pressure, P_{ref}, in the air is $20 \mu Pa$.
4. Just to make this confusing, the reference pressure for sound measured in the ocean is $1 \mu Pa$. This means that sound intensities in the ocean are expressed in units that are 20 dB higher than those units used on land.

The Just Noticeable Difference in Person perception

1. Although typically thought of as just relevant for the perceptual experiences of physical stimuli, Ozer (1993) suggested that the JND is useful in personality assessment as a way of understanding the accuracy and inter judge agreement of judgments about other people.
2. In addition, Sinn (2003) has argued that the logarithmic nature of the *Weber-Fechner Law* is of evolutionary significance for preference for risk and cites Bernoulli (1738) as suggesting that our general utility function is logarithmic.

Money and non linearity

> *... the utility resulting from any small increase in wealth will be inversely proportionate to the quantity of goods already possessed .... if ... one has a fortune worth a hundred thousand ducats and another one a fortune worth same number of semi-ducats and if the former receives from it a yearly income of five thousand ducats while the latter obtains the same number of semi-ducats, it is quite clear that to the former a ducat has exactly the same significance as a semi-ducat to the latter (Bernoulli, 1738, p 25).*

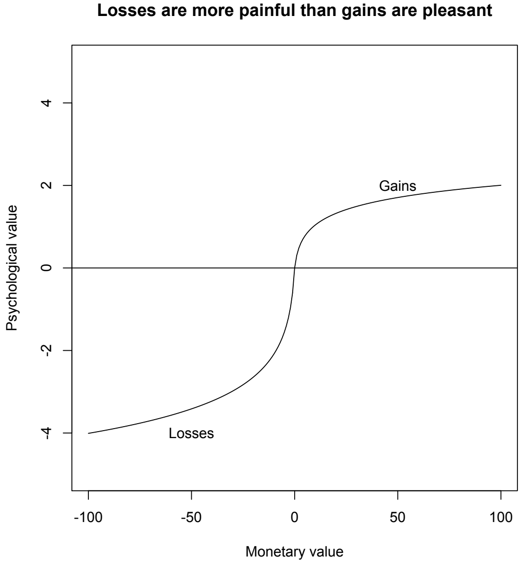
Implies a log function for utility.

Econs and Humans

1. Simple expected value theory $\implies$ value = probability of event x value of event
2. Bernouli theory of expected utility came to dominate choice theory and is fundamental to economics
3. Studied by comparing gambles and showing utility is non linear with value
    - Would you rather have \$80 or a 80% chance of \$100 + 20% of \$10?
    - expected value is 80 versus .8 * 100 + .2 * 10 = 82
4. Bernouli value (from Kahneman, 2011)

Wealth (millions)	1	2	3	4	5	6	7	8	9	10
Utility units	10	30	48	60	70	78	84	90	96	100

Kahneman and Tversky: Prospect Theory



Kahneman & Tversky (1979) Better to skip lunch than be someone's dinner.

Four types of scales and their associated statistics

Table: Four types of scales and their associated statistics (Rossi, 2007; Stevens, 1946) The statistics listed for a scale are invariant for that type of transformation.

Scale	Basic operations	Transformations	Invariant statistic	Examples
Nominal	equality $x_i = x_j$	Permutations	Counts Mode χ^2 and (ϕ) correlation	Detection Species classification Taxons
Ordinal	order $x_i > x_j$	Monotonic (homeomorphic) x' =f(x) f is monotonic	Median Percentiles Spearman correlations*	Mhos Hardness scale Beaufort Wind (intensity) Richter earthquake scale
Interval	differences $(x_i - x_j) > (x_k - x_l)$	Linear (Affine) x' = a + bx	Mean (μ) Standard Deviation (σ) Pearson correlation (r) Regression (β)	Temperature (°F, °C) Beaufort Wind (velocity)
Ratio	ratios $\frac{x_i}{x_j} > \frac{x_k}{x_l}$	Multiplication (Similiarity) x' = bx	Coefficient of variation ($\frac{\sigma}{\mu}$)	Length, mass, time Temperature (°K) Heating degree days

The Beaufort wind speed scale is interval with respect to the velocity of the wind, but only ordinal with respect to the effect of the wind. The Richter scale of earthquake intensity is a logarithmic scale of the energy released but

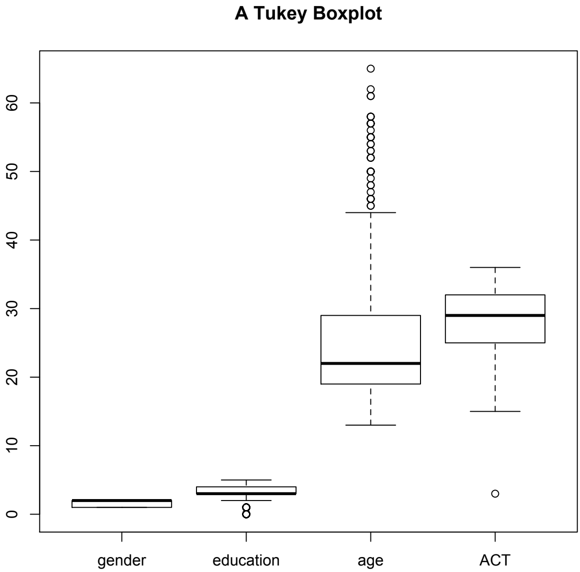
Graphical and tabular summaries of data

1. The Tukey 5 number summary shows the important characteristics of a set of numbers
   - Maximum
   - 75th percentile
   - Median (50th percentile)
   - 25th percentile
   - Minimum
2. Graphically, this is the box plot
   - Variations on the box plot include confidence intervals for the median

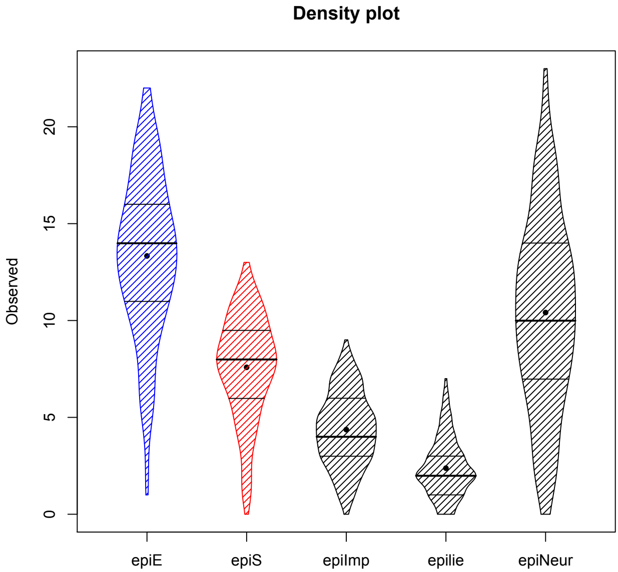
The summary command gives the Tukey 5 numbers

```
>  summary(sat.act)

    gender        education          age             ACT
 Min.   :1.000   Min.   :0.000   Min.   :13.00   Min.   : 3.00   Min.
 1st Qu.:1.000   1st Qu.:3.000   1st Qu.:19.00   1st Qu.:25.00   1st Q
 Median :2.000   Median :3.000   Median :22.00   Median :29.00   Media
 Mean   :1.647   Mean   :3.164   Mean   :25.59   Mean   :28.55   Mean
 3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:29.00   3rd Qu.:32.00   3rd Q
 Max.   :2.000   Max.   :5.000   Max.   :65.00   Max.   :36.00   Max.
```

A box plot of the first 4 sat.act variables



A violin or density plot of the first 5 epi.bfi variables



The `describe` function gives more descriptive statistics

```
> describe(sat.act)
```

```
          vars   n   mean     sd median trimmed    mad min max range  skew kurtosis   se
gender       1 700   1.65   0.48      2    1.68   0.00   1   2     1 -0.61    -1.62 0.02
education    2 700   3.16   1.43      3    3.31   1.48   0   5     5 -0.68    -0.07 0.05
age          3 700  25.59   9.50     22   23.86   5.93  13  65    52  1.64     2.42 0.36
ACT          4 700  28.55   4.82     29   28.84   4.45   3  36    33 -0.66     0.53 0.18
SATV         5 700 612.23 112.90    620  619.45 118.61 200 800   600 -0.64     0.33 4.27
SATQ         6 687 610.22 115.64    620  617.25 118.61 200 800   600 -0.59    -0.02 4.41
```

Multiple measures of central tendency

mode
: The most frequent observation. Not a very stable measure, depends upon grouping. Can be used for categorical data.

median
: The number with 50% above and 50% below. A powerful, if underused, measure. Not sensitive to transforms of the shape of the distribution, nor outliers. Appropriate for ordinal data, and useful for interval data.

mean
: One of at least seven measures that assume interval properties of the data.

Multiple ways to estimate the mean

Arithmetic mean $\bar{X} = X_. = (\sum_{i=1}^{N} X_i)/N$ `mean(x)`

Trimmed mean throws away the top and bottom t% of observations. This follows the principle that all data are normal at the middle. `mean(x,trim=.1)`

Winsorized mean Find the arithmetic mean after replacing the n lowest observations with the nth value, and the N largest values with the Nth largest. `winsor(x,trim=.2)`

Geometric Mean $\bar{X}_{geometric} = \sqrt[N]{\prod_{i=1}^{N} X_i} = e^{\Sigma(ln(x))/N}$ (The anti-log of the mean log score). `geometric.mean(x)`

Harmonic Mean $\bar{X}_{harmonic} = \frac{N}{\sum_{i=1}^{N} 1/X_i}$ (The reciprocal of the mean reciprocal). `harmonic.mean(x)`

Circular Mean $\bar{x}_{circular} = tan^{-1}\left(\frac{\sum cos(x)}{\sum sin(x)}\right)$ `circular.mean(x)` (where x is in radians)

circadian.mean `circular.mean(x)` (where x is in hours)

Class size from the students' point of view.

Table: Class size from the students' point of view. Most students are in large classes; the median class size is 200 with a mean of 223.

Class size	Number of classes	number of students
10	12	120
20	4	80
100	2	200
200	1	200
400	1	400

Time in therapy

A psychotherapist is asked what is the average length of time that a patient is in therapy. This seems to be an easy question, for of the 20 patients, 19 have been in therapy for between 6 and 18 months (with a median of 12) and one has just started. Thus, the median client is in therapy for 52 weeks with an average (in weeks) $(1 * 1 + 19 * 52)/20$ or 49.4.

However, a more careful analysis examines the case load over a year and discovers that indeed, 19 patients have a median time in treatment of 52 weeks, but that each week the therapist is also seeing a new client for just one session. That is, over the year, the therapist sees 52 patients for 1 week and 19 for a median of 52 weeks. Thus, the median client is in therapy for 1 week and the average client is in therapy of $(52 * 1 + 19 * 52)/(52+19) =$ 14.6 weeks.

Does teaching effect learning?

1. A leading research team in motivational and educational psychology was interested in the effect that different teaching techniques at various colleges and universities have upon their students. They were particularly interested in the effect upon writing performance of attending a very selective university, a less selective university, or a two year junior college.
2. A writing test was given to the entering students at three institutions in the Boston area. After one year, a similar writing test was given again. Although there was some attrition from each sample, the researchers report data only for those who finished one year. The pre and post test scores as well as the change scores were as shown below:

Types of teaching affect student outcomes?

Table: Three types of teaching and their effect on student outcomes

School	Pretest	Posttest	Change
Junior College	1	5	4
Non-selective university	5	27	22
Selective university	27	73	45

From these data, the researchers concluded that the quality of teaching at the selective university was much better than that of the less selective university or the junior college and that the students learned a great deal more. They proposed to study the techniques used there in order to apply them to the other institutions.

Teaching and math performance

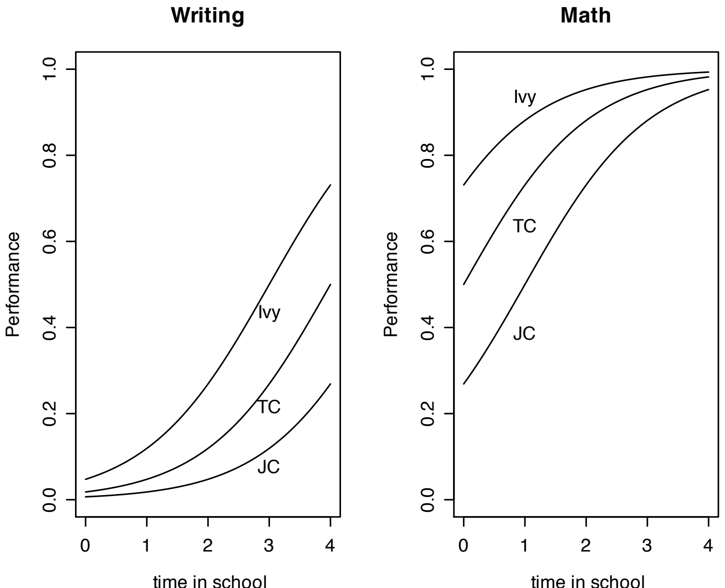
Another research team in motivational and educational psychology was interested in the effect that different teaching at various colleges and universities affect math performance. They used the same schools as the previous example with the same design.

Table: Three types of teaching and their effect on student outcomes

School	Pretest	Posttest	Change
Junior College	27	73	45
Non-selective university	73	95	22
Selective university	95	99	4

They concluded that the teaching at the junior college was far superior to that of the select university. What is wrong with this conclusion?

Effect of teaching, effect of students, or just scaling?

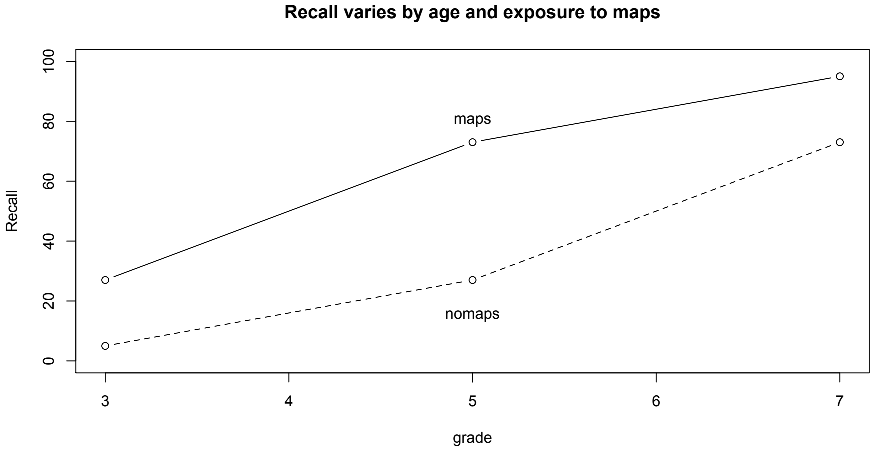


The problem of scaling is ubiquitous

1. A leading cognitive developmentalist believed that there is a critiical stage for learning spatial representations using maps. Children younger than this stage are not helped by maps, nor are children older than this stage.
2. He randomly assigned 3rd, 5th, and 7th grade students into two conditions (nested within grade), control, and map use. Performance was measures on a task of spatial recall (children were shown toys at particular locations in a set of rooms and then asked to find them again later.) Half the children were shown a map of the rooms before doing the task.
3. Their scores were

	No Map	Maps	Effect	
3rd grade	5	27	22	Too young
5th grade	27	73	46	Critical period
7th grade	73	95	22	Too old

Map use is most effective at a particular developmental stage

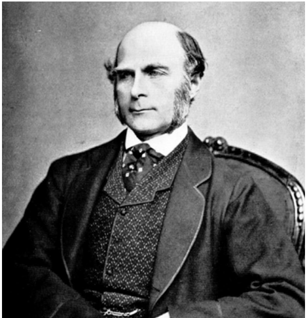


Correlation and Regression

1. Developed in 1886 by Francis Galton
   - Further developments by Karl Pearson and Charles Spearman
2. Correlation/regression are the root concept of psychometrics
   - Other statistics, including factor analysis are ways of partioning correlation matrices
   - Reliability theory is merely an application of factor analysis

Francis Galton 1822-1911

Francis Galton (1822-1911) was among the most influential psychologists of the 19th century. He did pioneering work on the correlation coefficient, behavior genetics and the measurement of individual differences. He introspectively examined the question of free will and introduced the lexical hypothesis to the study of personality and character. In addition to psychology, he did pioneering work in meteorology and introduced the scientific use of fingerprints. Whenever he could, he counted.



Karl Pearson 1857-1936

Carl (Karl) Pearson was among the most influential statisticians of the early 20th century. Founder of the statistics department at University College London. He developed the Pearson Product Moment Correlation Coefficient, its special case the ϕ coefficient, and the tetrachoric correlation. Major behavior geneticist and eugenicist.

Charles Spearman 1863-1945

Charles Spearman (1863-1945) was the leading psychometrician of the early 20th century. His work on the classical test theory, factor analysis, and the g theory of intelligence continues to influence psychometrics, statistics, and the study of intelligence. More than 100 years after their publication, his most influential papers remain two of the most frequently cited articles in psychometrics and intelligence.

Galton's height data

Table: The relationship between the average of both parents (mid parent) and the height of their children. The basic data table is from Galton (1886) who used these data to introduce reversion to the mean (and thus, linear regression). The data are available as part of the **UsingR** or **psych** packages.

```
> library(psych)
> data(galton)
> galton.tab <- table(galton)
> galton.tab[order(rank(rownames(galton.tab)),decreasing=TRUE),] #s
```

parent \ child	61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	73.7
73	0	0	0	0	0	0	0	0	0	0	0	1	3	0
72.5	0	0	0	0	0	0	0	1	2	1	2	7	2	4
71.5	0	0	0	0	1	3	4	3	5	10	4	9	2	2
70.5	1	0	1	0	1	1	3	12	18	14	7	4	3	3
69.5	0	0	1	16	4	17	27	20	33	25	20	11	4	5
68.5	1	0	7	11	16	25	31	34	48	21	18	4	3	0
67.5	0	3	5	14	15	36	38	28	38	19	11	4	0	0
66.5	0	3	3	5	2	17	17	14	13	4	0	0	0	0
65.5	1	0	9	5	7	11	11	7	7	5	2	1	0	0
64.5	1	1	4	4	1	5	5	0	2	0	0	0	0	0
64	1	0	2	4	1	2	2	1	1	0	0	0	0	0

Galton's height data

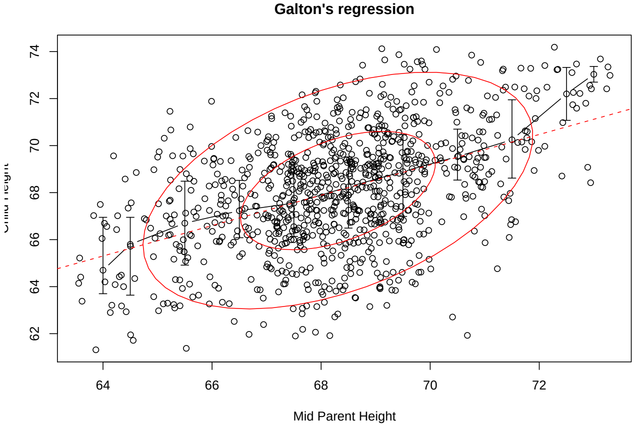


Figure: Galton's data can be plotted to show the relationships between mid parent and child heights. Because the original data are grouped, the data points have been *jittered* to emphasize the density of points along the median. The bars connect the first, 2nd (median) and third quartiles. The dashed line is the best fitting linear fit, the ellipses represent one and two standard deviations from the mean.

Bivariate Regression

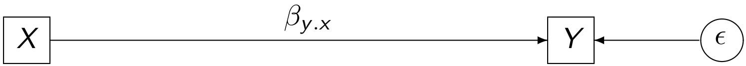
X Y ϵ

$$\hat{y} = \beta_{y.x}x + \epsilon$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

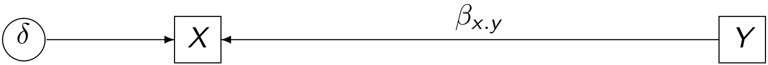
Bivariate Regression

δ X Y ϵ



$\hat{y} = \beta_{y.x}x + \epsilon$

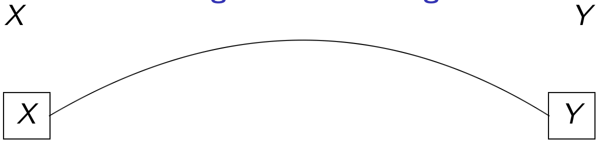
$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$



$\hat{x} = \beta_{x.y}y + \delta$

$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$

Bivariate Correlation is the geometric average of the two regressions



$\hat{x} = \beta_{x.y} y + \delta$

$\hat{y} = \beta_{y.x} x + \epsilon$

$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$

$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$

$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

The variance and the variance of a composite

1. If $\mathbf{x_1}$ and $\mathbf{x_2}$ are vectors of N observations centered around their mean (that is, deviation scores) their variances are $V_{x1} = \sum x_{i1}^2/(N-1)$ and $V_{x2} = \sum x_{i2}^2/(N-1)$, or, in matrix terms $V_{\mathbf{x}1} = \mathbf{x_1'x_1}/(N-1)$ and $V_{\mathbf{x}2} = \mathbf{x_2'x_2}/(N-1)$.
2. The variance of the composite made up of the sum of the corresponding scores, $\mathbf{x}+\mathbf{y}$ is just

$$V_{(\mathbf{x1}+\mathbf{x2})} = \frac{\sum(x_i+y_i)^2}{N-1} = \frac{\sum x_i^2 + \sum y_i^2 + 2\sum x_i y_i}{N-1} = \frac{(\mathbf{x}+\mathbf{y})'(\mathbf{x}+\mathbf{y})}{N-1}. \quad (3)$$

Or, more generally,

$$\mathbf{S} = \begin{pmatrix} V_{x1} & C_{x1x2} & \cdots & C_{x1xn} \\ C_{x1x2} & V_{x2} & & C_{x2xn} \\ \vdots & & \ddots & \vdots \\ C_{x1xn} & C_{x2xn} & \cdots & V_{xn} \end{pmatrix}$$

Sums as matrix products

$$V_{\mathbf{X}} = \sum \frac{\mathbf{X}'\mathbf{X}}{N-1} = \frac{\mathbf{1}'(\mathbf{X}'\mathbf{X})\mathbf{1}}{N-1}.$$

$$V_{\mathbf{Y}} = \sum \frac{\mathbf{Y}'\mathbf{Y}}{N-1} = \frac{\mathbf{1}'(\mathbf{Y}'\mathbf{Y})\mathbf{1}}{N-1}$$

and

$$C_{\mathbf{XY}} = \sum \frac{\mathbf{X}'\mathbf{Y}}{N-1} = \frac{\mathbf{1}'(\mathbf{X}'\mathbf{Y})\mathbf{1}}{N-1}$$

Use R

Get the data from a remote data source

A nice feature of R is that you can read from remote data sets. The example dataset is on the personality-project.org server. Get it and describe it.

R code

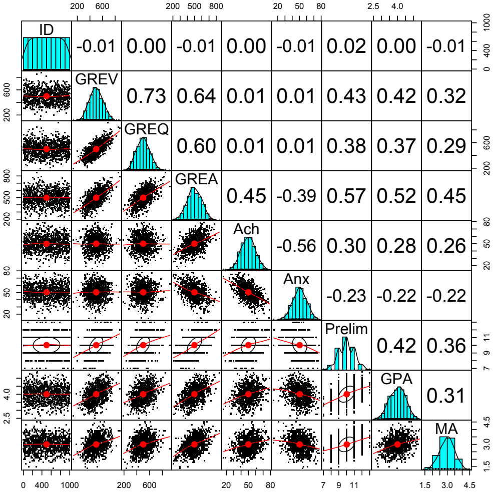
```
>  datafilename="http://personality-project.org/r/datasets/psychometri
>  mydata =read.table(datafilename,header=TRUE)  #read the data file
> describe(mydata,skew=FALSE)
```

```
       var    n   mean     sd median trimmed    mad   min     max  range   se
ID       1 1000 500.50 288.82 500.50  500.50 370.65   1.0 1000.00 999.00 9.13
GREV     2 1000 499.77 106.11 497.50  498.75 106.01 138.0  873.00 735.00 3.36
GREQ     3 1000 500.53 103.85 498.00  498.51 105.26 191.0  914.00 723.00 3.28
GREA     4 1000 498.13 100.45 495.00  498.67  99.33 207.0  848.00 641.00 3.18
Ach      5 1000  49.93   9.84  50.00   49.88  10.38  16.0   79.00  63.00 0.31
Anx      6 1000  50.32   9.91  50.00   50.43  10.38  14.0   78.00  64.00 0.31
Prelim   7 1000  10.03   1.06  10.00   10.02   1.48   7.0   13.00   6.00 0.03
GPA      8 1000   4.00   0.50   4.02    4.01   0.53   2.5    5.38   2.88 0.02
MA       9 1000   3.00   0.49   3.00    3.00   0.44   1.4    4.50   3.10 0.02
```

Plot it using the `pairs.panels` function.

Use the pairs.panels function to show a splom plot (use gap=0 and pch='.').

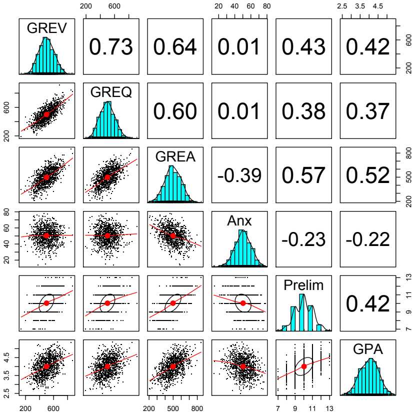
>pairs.panels(mydata,pch=".",gap=0) #pch='.' makes for a cleaner plot



Plot a subset of the data using the c() function (concatenate).

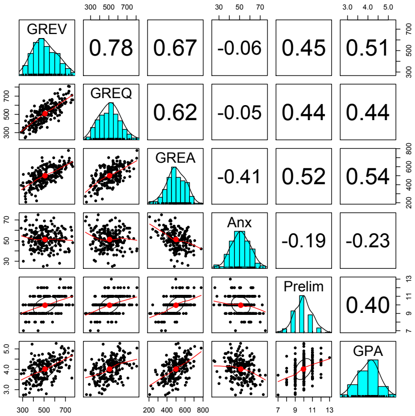
Use the pairs.panels function to show a splom plot. Select a subset of variables using the c() function.

```
>pairs.panels(mydata[c(2:4,6:8)],pch='.')
```



Do this for the first 200 subjects

> pairs.panels(mydata[mydata$ID < 200,c(2:4,6:8)])



0 center the data

In order to do interaction terms in regressions, it is necessary to 0 center the data. We need to turn the result into a data.frame in order to use it in the regression function.

```
> cent <- data.frame(scale(mydata,scale=FALSE))
> describe(cent,skew=FALSE)
       var    n mean     sd median trimmed    mad     min    max  range   se
ID       1 1000    0 288.82   0.00    0.00 370.65 -499.50 499.50 999.00 9.13
GREV     2 1000    0 106.11  -2.27   -1.02 106.01 -361.77 373.23 735.00 3.36
GREQ     3 1000    0 103.85  -2.53   -2.02 105.26 -309.53 413.47 723.00 3.28
GREA     4 1000    0 100.45  -3.13    0.54  99.33 -291.13 349.87 641.00 3.18
Ach      5 1000    0   9.84   0.07   -0.05  10.38  -33.93  29.07  63.00 0.31
Anx      6 1000    0   9.91  -0.32    0.11  10.38  -36.32  27.68  64.00 0.31
Prelim   7 1000    0   1.06  -0.03    0.00   1.48   -3.03   2.97   6.00 0.03
GPA      8 1000    0   0.50   0.02    0.00   0.53   -1.50   1.38   2.88 0.02
MA       9 1000    0   0.49   0.00    0.00   0.44   -1.60   1.50   3.10 0.02
```

The standard deviations and ranges have not changed. However, the means are all 0. We use the `scale` function with the scale=FALSE option.

The standardized data

Alternatively, we could standardize it.

```
> z.data <- data.frame(scale(my.data))
>  describe(z.data)
       var    n mean sd median trimmed  mad   min  max range  skew kurtosis   se
ID       1 1000    0  1   0.00    0.00 1.28 -1.73 1.73  3.46  0.00    -1.20 0.03
GREV     2 1000    0  1  -0.02   -0.01 1.00 -3.41 3.52  6.93  0.09    -0.07 0.03
GREQ     3 1000    0  1  -0.02   -0.02 1.01 -2.98 3.98  6.96  0.22     0.08 0.03
GREA     4 1000    0  1  -0.03    0.01 0.99 -2.90 3.48  6.38 -0.02    -0.06 0.03
Ach      5 1000    0  1   0.01   -0.01 1.05 -3.45 2.95  6.40  0.00     0.02 0.03
Anx      6 1000    0  1  -0.03    0.01 1.05 -3.67 2.79  6.46 -0.14     0.14 0.03
Prelim   7 1000    0  1  -0.02    0.00 1.40 -2.86 2.81  5.67 -0.02    -0.01 0.03
GPA      8 1000    0  1   0.03    0.01 1.06 -3.00 2.74  5.74 -0.07    -0.29 0.03
MA       9 1000    0  1   0.01    0.01 0.90 -3.23 3.04  6.27 -0.07    -0.09 0.03
```

Or, we can standardize it by dividing though by the standard deviation. We use the `scale` function to do this for us.

Show how the correlations do not change with standardization

Find the correlations using the `lowerCor` function. This, by default, uses pairwise Pearson correlations and rounds to two decimals. Compare with the standard `cor` function.

```
> lowerCor(my.data)
       ID    GREV  GREQ  GREA  Ach   Anx   Prelm GPA   MA
ID      1.00
GREV   -0.01  1.00
GREQ    0.00  0.73  1.00
GREA   -0.01  0.64  0.60  1.00
Ach     0.00  0.01  0.01  0.45  1.00
Anx    -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim  0.02  0.43  0.38  0.57  0.30 -0.23  1.00
GPA     0.00  0.42  0.37  0.52  0.28 -0.22  0.42  1.00
MA     -0.01  0.32  0.29  0.45  0.26 -0.22  0.36  0.31  1.00
```

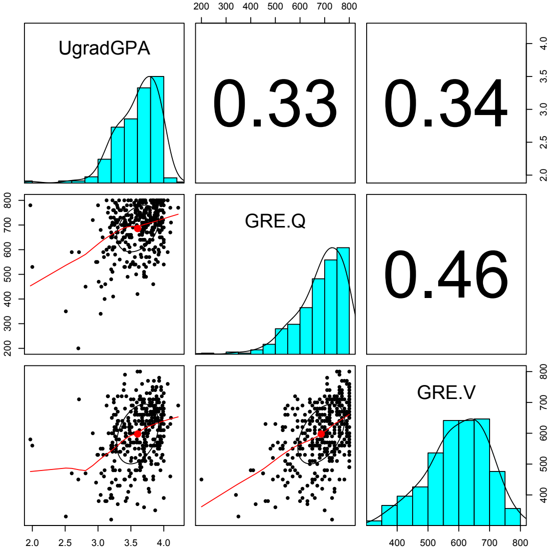
```
> lowerCor(z.data)
       ID    GREV  GREQ  GREA  Ach   Anx   Prelm
ID      1.00
GREV   -0.01  1.00
GREQ    0.00  0.73  1.00
GREA   -0.01  0.64  0.60  1.00
Ach     0.00  0.01  0.01  0.45  1.00
Anx    -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim  0.02  0.43  0.38  0.57  0.30 -0.23  1.00
GPA     0.00  0.42  0.37  0.52  0.28 -0.22  0.42
MA     -0.01  0.32  0.29  0.45  0.26 -0.22  0.36
```

Show that the two matrices do not differ using the `lowerUpper` function

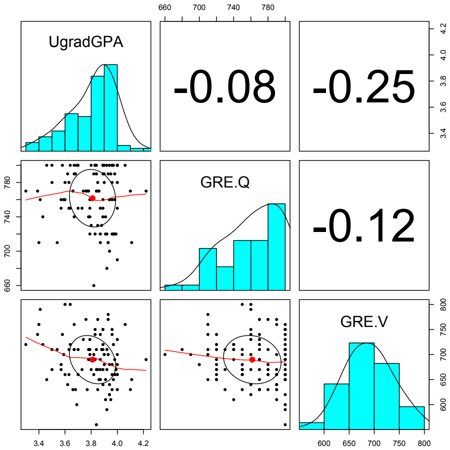
```
r <- lowerCor(my.data)  #find the original correlations
z <- lowerCor(z.data)   #find the z transformed correlations
lu <- lowerUpper(r,z,diff=TRUE) #combine into one matrix and take t

round(lu,2)
```

```
          ID GREV GREQ  GREA   Ach   Anx Prelim  GPA MA
ID        NA 0.00 0.00  0.00  0.00  0.00   0.00 0.00  0
GREV   -0.01   NA 0.00  0.00  0.00  0.00   0.00 0.00  0
GREQ    0.00 0.73   NA  0.00  0.00  0.00   0.00 0.00  0
GREA   -0.01 0.64 0.60    NA  0.00  0.00   0.00 0.00  0
Ach     0.00 0.01 0.01  0.45    NA  0.00   0.00 0.00  0
Anx    -0.01 0.01 0.01 -0.39 -0.56    NA   0.00 0.00  0
Prelim  0.02 0.43 0.38  0.57  0.30 -0.23     NA 0.00  0
GPA     0.00 0.42 0.37  0.52  0.28 -0.22   0.42   NA  0
MA     -0.01 0.32 0.29  0.45  0.26 -0.22   0.36 0.31 NA
```

Scatter Plot Matrix showing correlation and LOESS regression

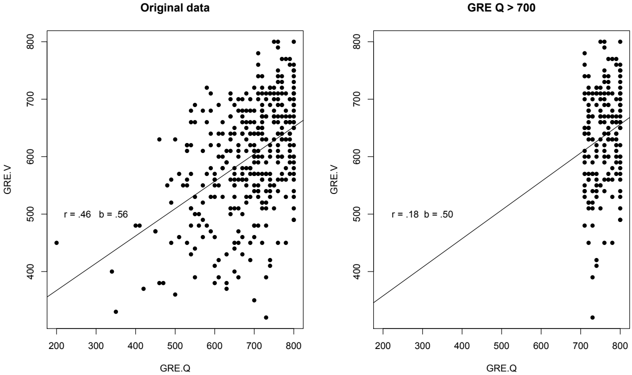


The effect of selection on the correlation



- Consider what happens if we select a subset
  - The "Oregon" model
  - (GPA + (V+Q)/200) > 11.6
- The range is truncated, but even more important, by using a compensatory selection model, we have changed the sign of the correlations.

Regression and restriction of range

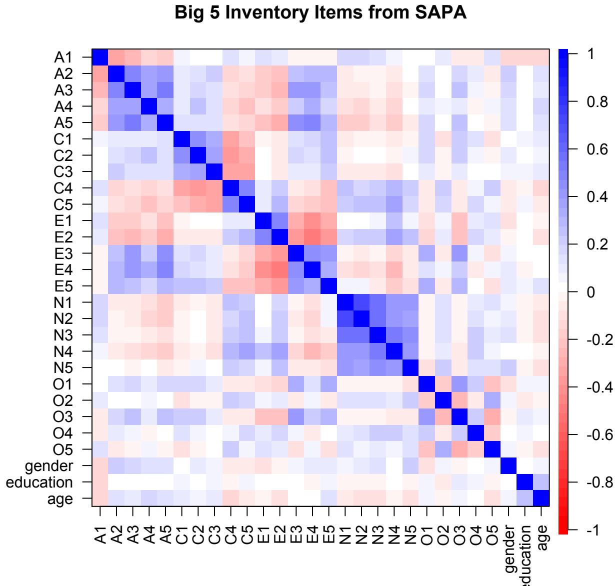


Although the correlation is very sensitive, regression slopes are relatively insensitive to restriction of range.

R code for regression figures

```
gradq <- subset(gradf,gradf[2]>700) #choose the subset
with(gradq,lm(GRE.V ~ GRE.Q)) #do the regression
Call:
lm(formula = GRE.V ~ GRE.Q)
Coefficients:
(Intercept)        GRE.Q
   258.1549       0.4977

#show the graphic
op <- par(mfrow=c(1,2))  #two panel graph
 with(gradf,{
 plot(GRE.V ~ GRE.Q,xlim=c(200,800),main='Original data', pch=16)
 abline(lm(GRE.V ~ GRE.Q))
 })
 text(300,500,'r = .46   b = .56')
 with(gradq,{
 plot(GRE.V ~ GRE.Q,xlim=c(200,800),main='GRE Q > 700',pch=16)
 abline(lm(GRE.V ~ GRE.Q))
 })
  text(300,500,'r = .18  b = .50')
 op <- par(mfrow=c(1,1)) #switch back to one panel
```

Show many correlations with a heat map using `cor.plot`.



Alternative versions of the correlation coefficient

Table: A number of correlations are Pearson r in different forms, or with particular assumptions. If $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, then depending upon the type of data being analyzed, a variety of correlations are found.

Coefficient	symbol	X	Y	Assumptions
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi-serial	r_{bis}	dichotomous	continuous	normality
Tetrachoric	r_{tet}	dichotomous	dichotomous	bivariate normality
Polychoric	r_{pc}	categorical	categorical	bivariate normality

The ϕ coefficient is just a Pearson r on dichotomous data

Table: The basic table for a phi, ϕ coefficient, expressed in raw frequencies in a four fold table is taken from Pearson & Heron (1913)

	Success	Failure	Total
Accept	A	B	R_1 = A + B
Reject	C	D	R_2 = C + D
Total	C_1 = A + C	C_2=B + D	n = A + B + C + D

In terms of the raw data coded 0 or 1, the *phi coefficient* can be derived directly by direct substitution, recognizing that the only non zero product is found in the A cell

$$n\sum X_i Y_i - \sum X_i \sum Y_i = nA - R_1 C_1$$

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \qquad (4)$$

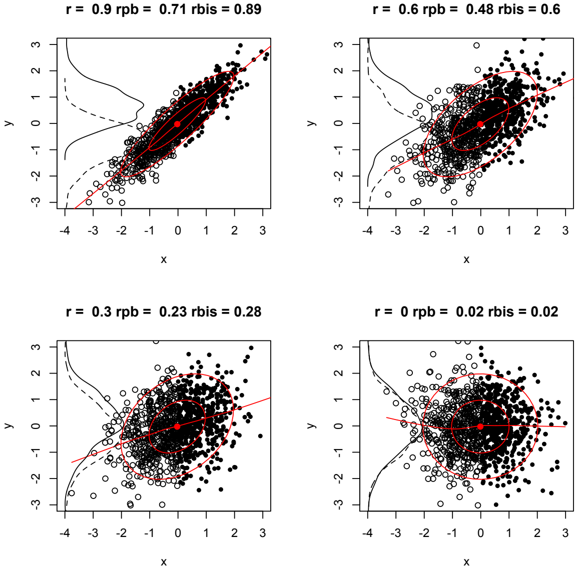
Correlation size $\neq$ causal importance

Table: The relationship between sex and pregnancy (hypothetical data)

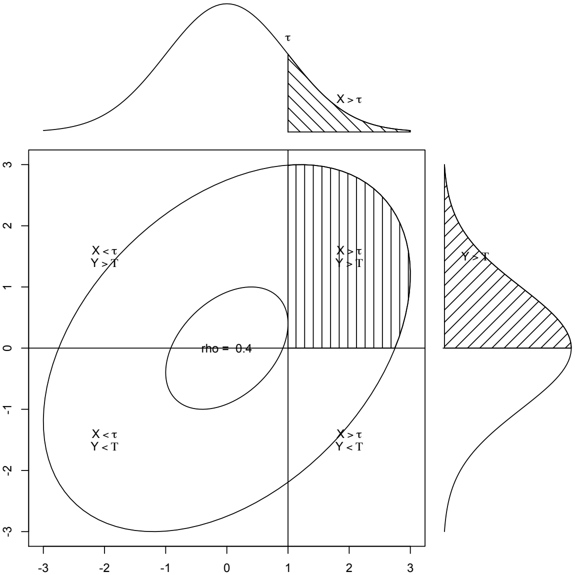
	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300
Phi	.04		

```
> sex <- c(2, 1041,0,6257)
> phi(sex)
[1] 0.04
```

The biserial correlation estimates the latent correlation



The tetrachoric correlation estimates the latent correlation



Correlation size $\neq$ causal importance – tetrachoric correlation

Table: The relationship between sex and pregnancy (hypothetical data)

	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300
Phi	.04	ρ_{tet}	.95

```
> sex <- c(2, 1041,0,6257)
> phi(sex)
[1] 0.04
> tetrachoric(sex,correct=FALSE)
Call: tetrachoric(x = sex, correct = FALSE)
tetrachoric correlation
[1] 0.95

 with tau of
[1] -3.5 -1.1
```

Pearson r versus tetrachoric correlation on dichotomous ability data

```
> tet <- tetrachoric(ability)
Loading required package: mvtnorm
Loading required package: parallel
> per <- lowerCor(ability)
> per.tet <- lowerUpper(tet$rho,per)
> per.tet.diff <- lowerUpper(tet$rho,per,diff=TRUE)
> round(per.tet[1:8,1:8],2)
```

```
          reason.4 reason.16 reason.17 reason.19 letter.7 letter.33 letter.34 letter.58
reason.4        NA      0.28      0.40      0.30     0.28      0.23      0.29      0.29
reason.16     0.45        NA      0.32      0.25     0.27      0.20      0.26      0.21
reason.17     0.61      0.51        NA      0.34     0.29      0.26      0.29      0.29
reason.19     0.46      0.40      0.53        NA     0.25      0.25      0.27      0.25
letter.7      0.45      0.43      0.47      0.40       NA      0.34      0.40      0.33
letter.33     0.37      0.32      0.42      0.39     0.52        NA      0.37      0.28
letter.34     0.46      0.41      0.47      0.43     0.60      0.56        NA      0.32
letter.58     0.47      0.35      0.48      0.40     0.51      0.43      0.50        NA
> round(per.tet.diff[1:8,1:8],2)
          reason.4 reason.16 reason.17 reason.19 letter.7 letter.33 letter.34 letter.58
reason.4        NA      0.17      0.21      0.17     0.16      0.14      0.17      0.18
reason.16     0.45        NA      0.19      0.15     0.16      0.13      0.16      0.14
reason.17     0.61      0.51        NA      0.19     0.18      0.16      0.18      0.19
reason.19     0.46      0.40      0.53        NA     0.14      0.14      0.15      0.15
letter.7      0.45      0.43      0.47      0.40       NA      0.18      0.20      0.18
letter.33     0.37      0.32      0.42      0.39     0.52        NA      0.19      0.15
letter.34     0.46      0.41      0.47      0.43     0.60      0.56        NA      0.18
letter.58     0.47      0.35      0.48      0.40     0.51      0.43      0.50        NA
```

Pearson r versus polychoric correlation on 6 alternative BFI data

```
> poly <- polychoric(bfi[1:10])
> pearson <- cor(bfi[1:10],use="pairwise")
> poly.pear <- lowerUpper(poly$rho,pearson)
> poly.pear.diff <- lowerUpper(poly$rho,pearson,diff=TRUE)
> poly.pear
```

```
> round(poly.pear,2)
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.34 -0.27 -0.15 -0.18  0.03  0.02 -0.02  0.13  0.05
A2 -0.41    NA  0.49  0.34  0.39  0.09  0.14  0.19 -0.15 -0.12
A3 -0.32  0.56    NA  0.36  0.50  0.10  0.14  0.13 -0.12 -0.16
A4 -0.18  0.39  0.41    NA  0.31  0.09  0.23  0.13 -0.15 -0.24
A5 -0.23  0.45  0.57  0.36    NA  0.12  0.11  0.13 -0.13 -0.17
C1  0.00  0.12  0.12  0.11  0.16    NA  0.43  0.31 -0.34 -0.25
C2  0.01  0.16  0.16  0.27  0.14  0.48    NA  0.36 -0.38 -0.30
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40    NA -0.34 -0.34
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38    NA  0.48
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53    NA
> round(poly.pear.diff,2)
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.07 -0.06 -0.03 -0.05 -0.02 -0.01  0.00  0.02  0.01
A2 -0.41    NA  0.07  0.05  0.06  0.02  0.02  0.03 -0.05 -0.03
A3 -0.32  0.56    NA  0.05  0.07  0.03  0.02  0.03 -0.04 -0.03
A4 -0.18  0.39  0.41    NA  0.05  0.02  0.04  0.04 -0.04 -0.04
A5 -0.23  0.45  0.57  0.36    NA  0.04  0.03  0.02 -0.04 -0.03
C1  0.00  0.12  0.12  0.11  0.16    NA  0.06  0.04 -0.06 -0.04
C2  0.01  0.16  0.16  0.27  0.14  0.48    NA  0.04 -0.05 -0.03
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40    NA -0.04 -0.04
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38    NA  0.05
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53    NA
```

Spearman vs. Pearson on BFI data

```
> spear <- cor(bfi[1:10],use="pairwise",method="spearman")
> spear.pear <- lowerUpper(spear,pearson,diff=TRUE)
> round(spear.pear,2)
```

	A1	A2	A3	A4	A5	C1	C2	C3	C4	C5
A1	NA	-0.03	-0.03	-0.01	-0.04	-0.05	-0.03	-0.02	0.02	0.01
A2	-0.37	NA	0.02	0.00	0.01	0.02	0.01	0.01	-0.03	-0.03
A3	-0.30	0.50	NA	0.00	0.03	0.02	0.01	0.02	-0.03	-0.02
A4	-0.16	0.34	0.36	NA	0.01	0.01	0.02	0.02	-0.03	-0.01
A5	-0.22	0.40	0.53	0.31	NA	0.02	0.02	0.01	-0.03	-0.02
C1	-0.02	0.11	0.12	0.10	0.15	NA	0.02	0.01	-0.04	-0.01
C2	-0.01	0.14	0.15	0.25	0.13	0.45	NA	0.01	-0.02	0.00
C3	-0.04	0.21	0.16	0.15	0.14	0.32	0.37	NA	-0.01	-0.01
C4	0.15	-0.18	-0.16	-0.18	-0.16	-0.38	-0.40	-0.35	NA	0.01
C5	0.06	-0.15	-0.18	-0.26	-0.19	-0.26	-0.30	-0.35	0.49	NA

Comments on these alternative correlations

1. The assumption is that there was an underlying bivariate, normal distribution that was somehow artificially dichotomized.
2. But some things are in fact dichotomous, not normally distributed
   - Alive/Dead
   - Vacinated/Not vacinated
3. polychromic and tetrachoric correlations are found by iteratively fitting bivariate normal distributions with varying correlations until the best fit for a n x n table is found.
4. This is done using the `tetrachoric` or `polychoric` functions. They are not fast! (In comparison to Pearson r).

Cautions about correlations–The Anscombe data set

Consider the following 8 variables

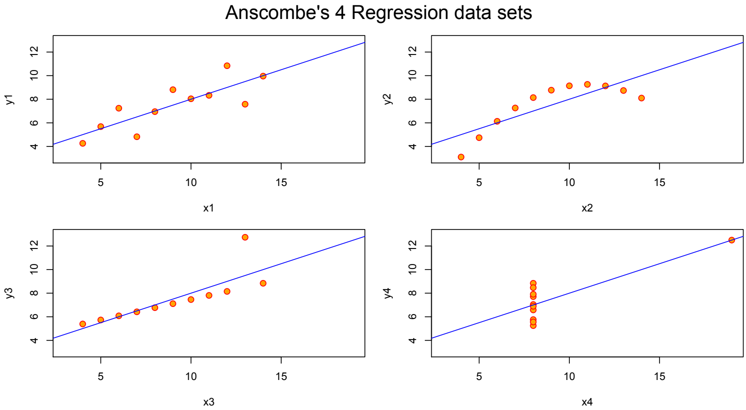
```
   var  n mean   sd median trimmed  mad  min   max range  skew kurtosi
x1   1 11  9.0 3.32   9.00    9.00 4.45 4.00 14.00 10.00  0.00    -1.2
x2   2 11  9.0 3.32   9.00    9.00 4.45 4.00 14.00 10.00  0.00    -1.2
x3   3 11  9.0 3.32   9.00    9.00 4.45 4.00 14.00 10.00  0.00    -1.2
x4   4 11  9.0 3.32   8.00    8.00 0.00 8.00 19.00 11.00  2.47    11.0
y1   5 11  7.5 2.03   7.58    7.49 1.82 4.26 10.84  6.58 -0.05    -0.5
y2   6 11  7.5 2.03   8.14    7.79 1.47 3.10  9.26  6.16 -0.98     0.8
y3   7 11  7.5 2.03   7.11    7.15 1.53 5.39 12.74  7.35  1.38     4.3
y4   8 11  7.5 2.03   7.04    7.20 1.90 5.25 12.50  7.25  1.12     3.1
```

Cautions, Anscombe continued

With regressions of

```
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629

[[2]]
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816

[[3]]
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305

[[4]]
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

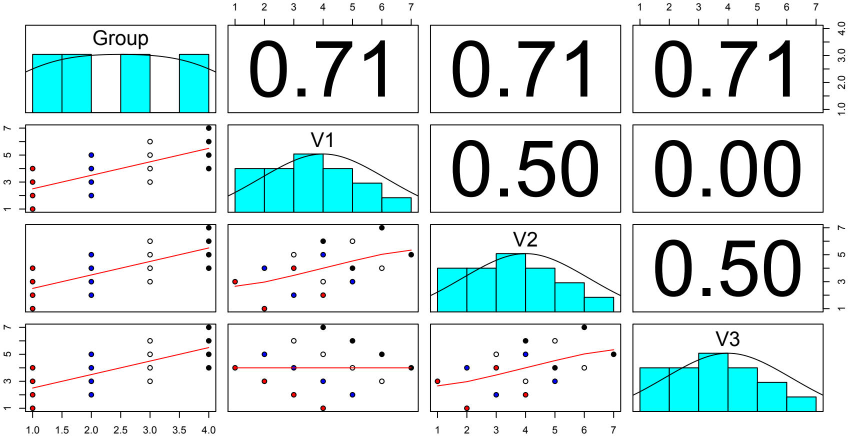
Cautions about correlations: Anscombe data set



Further cautions about correlations–the problem of levels

1. Correlations taken at one level of analysis can be unrelated to those at another level
2. $r_{xy} = \eta_{x_{wg}} * \eta_{y_{wg}} * r_{xy_{wg}} + \eta_{x_{bg}} * \eta_{y_{bg}} * r_{xy_{bg}}$
3. Where η is the correlation of the data with the within group values, or the group means.
4. The within group and between group correlations can even be of different sign!
5. The `withinBetween` data set is an example of this problem.
6. The `statsBy` function will find the within and between group correlations for this kind of multi-level design.

Cautions about correlations: Within versus between groups



Bias, or just Simpson's Paradox?

Table: Hypothetical Admissions data showing sex discrimination

	Admit	Reject	Total
Male	40	10	50
Female	10	40	50
Total	50	50	100

Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)= .60
polychoric rho = .81

Calculate the ϕ and tetrachoric correlations

```
> admit <- c(40,10,10,40)
> phi(admit)
[1] 0.6
> phi2poly(.6,.5,.5)
[1] 0.8090178
> tetrachoric(admit)
Call: tetrachoric(x = admit)
tetrachoric correlation
[1] 0.81

 with tau of
[1] 0 0
```

1. Input the four cell counts
2. Find the ϕ coefficient
3. Convert this to a tetrachoric correlation by specifying the marginals
4. Or, just call tetrachoric with these cell entries

Sex discrimination by department shows opposite effect

Table: Hypothetical Admissions data showing sex discrimination

	Admit	Reject	Total
Male	40	10	50
Female	10	40	50
Total	50	50	100

Table: Males: unselective

	Admit	Reject	Total
Male	40	5	45
Female	5	0	5
Total	45	5	50
ϕ	-.11	ρ	-.95

Table: Females: selective

	Admit	Reject	Total
Male	0	5	5
Female	5	40	45
Total	5	45	50
ϕ	-.11	ρ	-.95

The ubiquitous correlation coefficient

Table: Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Statistic	Estimate	r equivalent	as a function of r
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Regression	$b_{y.x} = \frac{Cxy}{\sigma_x^2}$	$r = b_{y.x} \frac{\sigma_y}{\sigma_x}$	$b_{y.x} = r \frac{\sigma_x}{\sigma_y}$
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g = \frac{2r\sqrt{df/N}}{\sqrt{1-r^2}}$
t - test	$t = \frac{d\sqrt{df}}{2}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = \frac{d^2 df}{4}$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

Correlation as the average of regressions

Galton's insight was that if both x and y were on the same scale with equal variability, then the slope of the line was the same for both predictors and was measure of the strength of their relationship. Galton (1886) converted all deviations to the same metric by dividing through by half the interquartile range, and Pearson (1896) modified this by converting the numbers to standard scores (i.e., dividing the deviations by the standard deviation). Alternatively, the geometric mean of the two slopes ($b_x y$ and $b_y x$) leads to the same outcome:

$$r_{xy} = \sqrt{b_{xy} b_{yx}} = \sqrt{\frac{(Cov_{xy} Cov_{yx}}{\sigma_x^2 \sigma_y^2}} = \frac{Cov_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \qquad (5)$$

which is the same as the covariance of the standardized scores of X and Y.

$$r_{xy} = Cov_{z_x z_y} = Cov_{\frac{x}{\sigma_x} \frac{y}{\sigma_y}} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \qquad (6)$$

Error of correlation

The slope $b_{y.x}$ was found so that it minimizes the sum of the squared residual, but what is it? That is, how big is the variance of the residual?

$$V_r = \sum_{i=1}^{n}(y - \hat{y})^2/n = \sum_{i=1}^{n}(y - b_{y.x}x)^2/n$$

$$V_r = \sum_{i=1}^{n}(y^2 + b_{y.x}^2 x^2 - 2b_{y.x}xy)/n$$

$$V_r = V_y + \frac{Cov_{xy}^2}{V_x} - 2\frac{Cov_{xy}^2}{V_x} = V_y - \frac{Cov_{xy}^2}{V_x}$$

$$V_r = V_y - r_{xy}^2 V_y = V_y(1 - r_{xy}^2) \quad (7)$$

That is, the *variance of the residual* in Y or the variance of the error of prediction of Y is the product of the original variance of Y and one minus the squared correlation between X and Y. The squared correlation between x and y is thus an index of the amount

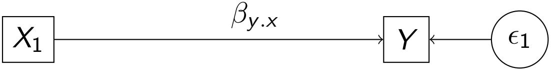
Variance and correlations

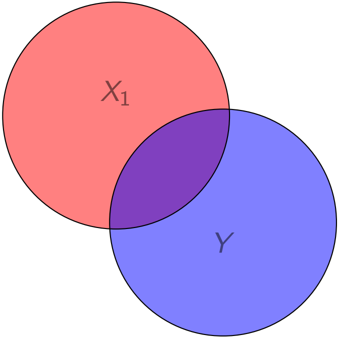
The various relationships between correlations, predicted scores, the variance of the predicted scores, and the variances of the residuals may be seen in the following table (19).

Table: The basic relationships between Variance, Covariance, Correlation and Residuals

	Variance	Covariance with X	Covariance with Y	Correlation with X	Correlation with Y
X	V_x	V_x	C_{xy}	1	r_{xy}
Y	V_y	C_{xy}	V_y	r_{xy}	1
$\hat{Y}$	$r_{xy}^2 V_y$	$C_{xy} = r_{xy}\sigma_x\sigma_y$	$r_{xy} V_y$	1	r_{xy}
$Y_r = Y - \hat{Y}$	$(1 - r_{xy}^2)V_y$	0	$(1 - r_{xy}^2)V_y$	0	$\sqrt{1 - r^2}$

Set theoretic approach: Partitioning the variance in Y





$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$
$$\hat{y} = \beta_{y.x}x$$
$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}}$$
$$V_r = V_y + \frac{Cov_{xy}^2}{V_x} - 2\frac{Cov_{xy}^2}{V_x}$$
$$V_r = V_y - \frac{Cov_{xy}^2}{V_x}$$
$$V_r = V_y - r_{xy}^2 V_y$$
$$V_r = V_y(1 - r_{xy}^2)$$

Variance in Y predicted by $X = r_{xy}^2\sigma_y^2$

Distance in the observational space

Because X and Y are vectors in the space defined by the observations, the covariance between them may be thought of in terms of the average squared distance between the two vectors in that same space. That is, following Pythagorus, the *distance*, d, is simply the square root of the sum of the squared distances in each dimension (for each pair of observations), or, if we find the average distance, we can find the square root of the sum of the squared distances divided by n:

$$d^2_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2.$$

which is the same as

$$d^2_{xy} = V_x + V_y - 2C_{xy}$$

$$d_{xy} = \sqrt{2 * (1 - r_{xy})}. \tag{8}$$

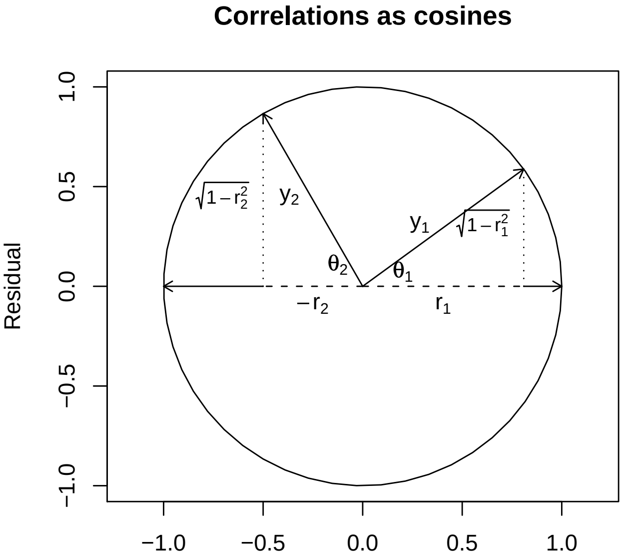
Distance, correlations, and the law of cosines

Compare this to the trigonometric *law of cosines*,

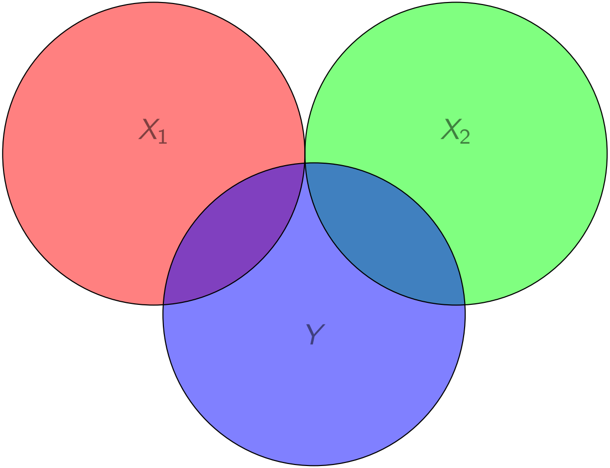
$$c^2 = a^2 + b^2 - 2ab \cdot cos(ab),$$

and we see that the distance between two vectors is the sum of their variances minus twice the product of their standard deviations times the cosine of the angle between them. That is, the correlation is the cosine of the angle between the two vectors. The next figure shows these relationships for two Y vectors. The correlation, r_1, of X with Y_1 is the cosine of $\theta_1 =$ the ratio of the projection of Y_1 onto X. From the *Pythagorean Theorem*, the length of the residual Y with X removed ($Y.x$) is $\sigma_y \sqrt{1 - r^2}$.

A geometric version of correlation

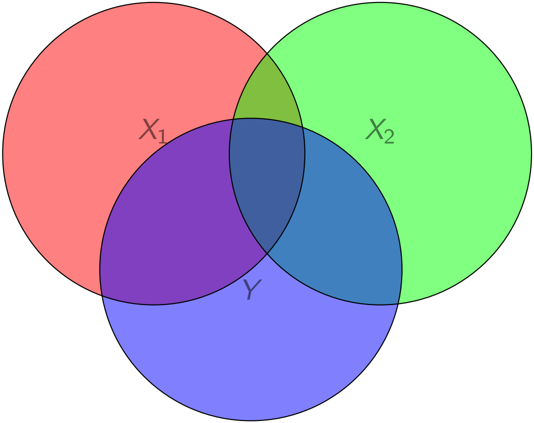


The Ideal model of predicting Y from X_1 and X_2



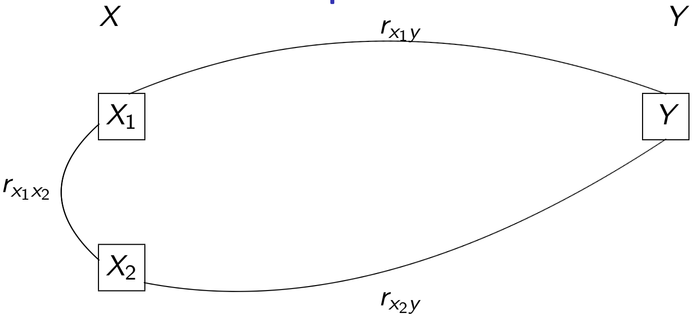
Variance in Y predicted by X_1 and X_2 if X_1 and X_2 are independent. $\hat{V}_y = V_y r^2_{x_1 y} + V_y r^2_{x_2 y}$

The usual case of predicting Y from X_1 and X_2

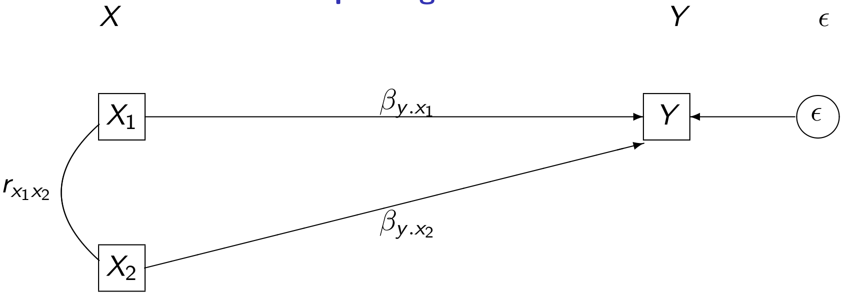


Variance in Y predicted by X_1 and X_2 if X_1 and X_2 - overlapping predictions $\hat{V}_y = V_y r^2_{x_1 y} + V_y r^2_{x_2 y} -$ overlap
But what is the overlap?

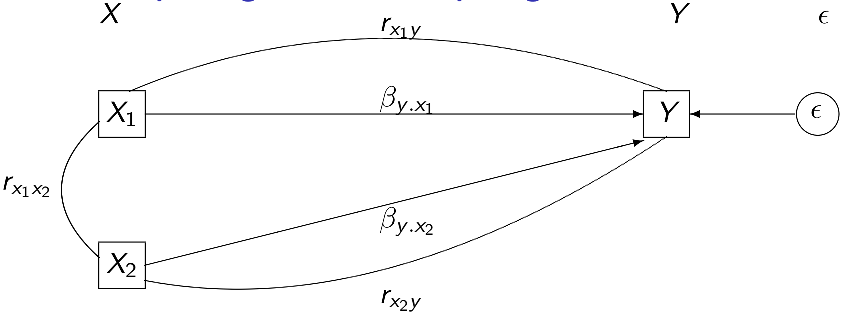
Multiple correlations



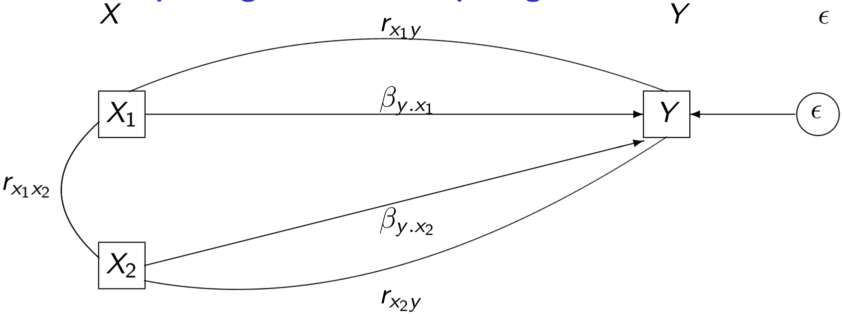
Multiple Regression



Multiple Regression: decomposing correlations



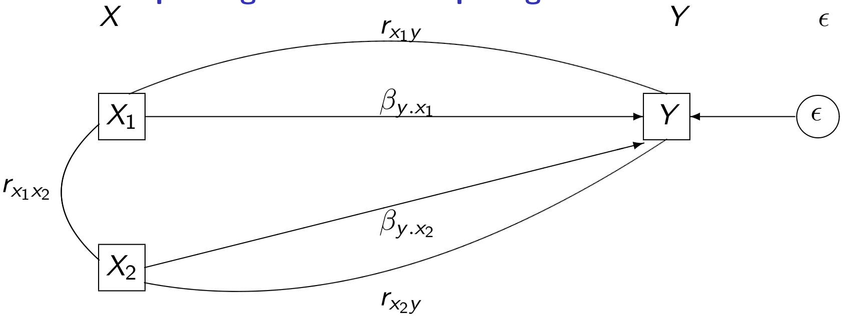
Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect}$$

$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect}$$

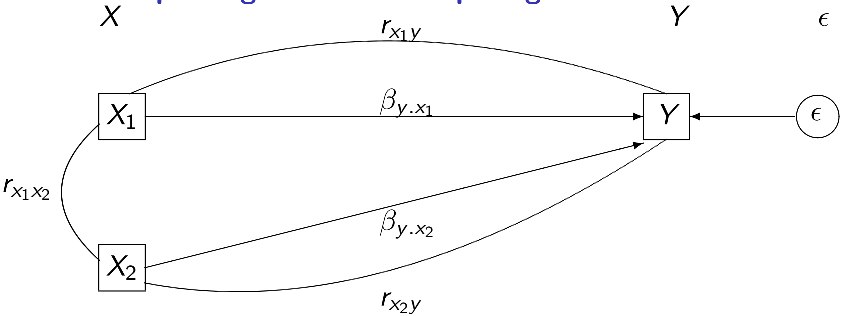
Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect} \qquad \beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2} r_{x_2y}}{1 - r^2_{x_1x_2}}$$

$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect} \qquad \beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2} r_{x_1y}}{1 - r^2_{x_1x_2}}$$

Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{\text{direct}} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{\text{indirect}}$$

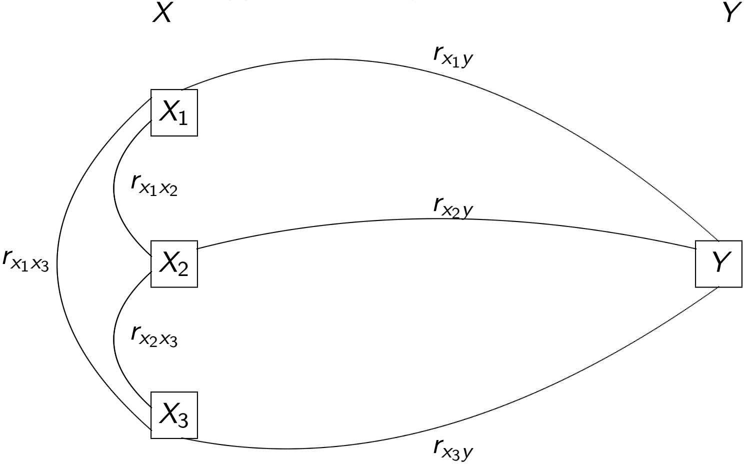
$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{\text{direct}} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{\text{indirect}}$$

$$\beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r^2_{x_1x_2}}$$

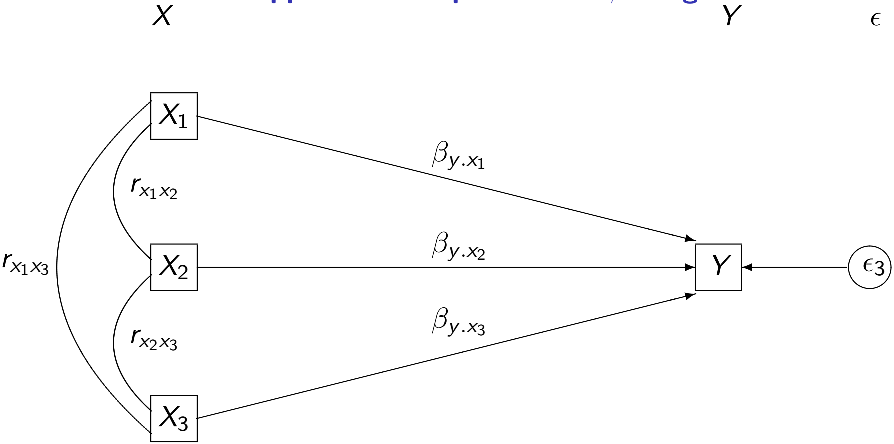
$$\beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r^2_{x_1x_2}}$$

$$R^2 = r_{x_1y}\beta_{y.x_1} + r_{x_2y}\beta_{y.x_2}$$

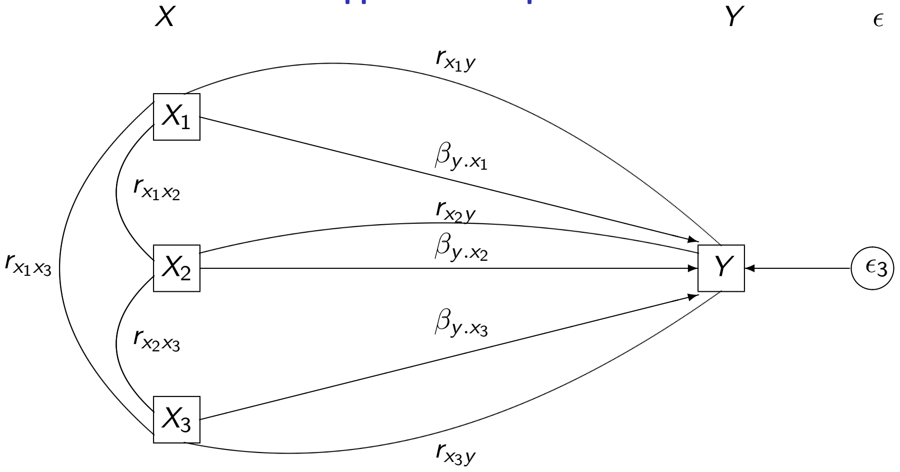
What happens with 3 predictors? The correlations



What happens with 3 predictors? β weights



What happens with 3 predictors?



$$r_{x_1 y} = \underbrace{\beta_{y.x_1}}_{direct} + \underbrace{r_{x_1 x_2}\beta_{y.x_1} + r_{x_1 x_3}\beta_{y.x_3}}_{indirect} \quad r_{x_2 y} = \ldots \quad r_{x_3 y} = \ldots$$

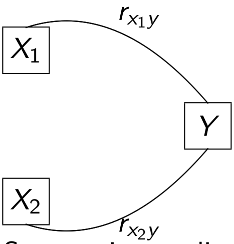
The math gets tedious

Multiple regression and linear algebra

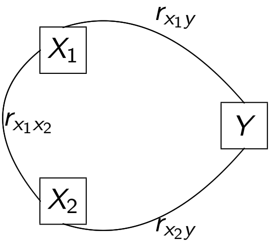
- Multiple regression requires solving multiple, simultaneous equations to estimate the direct and indirect effects.
    - Each equation is expressed as a $r_{x_i y}$ in terms of direct and indirect effects.
    - Direct effect is $\beta_{y.x_i}$
    - Indirect effect is $\sum_{j \neq i} beta_{y.x_j} r_{x_j y}$
- How to solve these equations?
- Tediously, or just use linear algebra.

3 special cases of regression

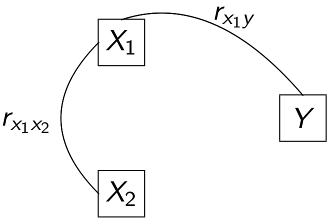
Orthogonal predictors



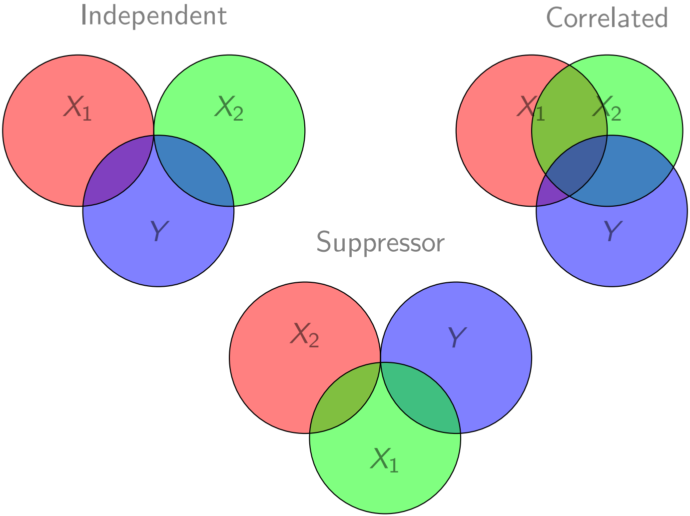
Correlated predictors



Suppressive predictors

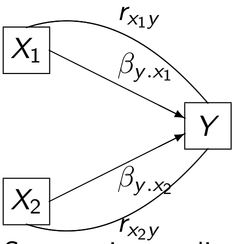


Three basic cases

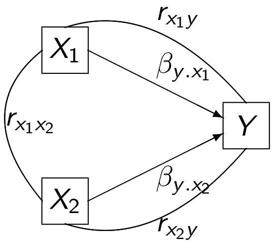


3 special cases of regression

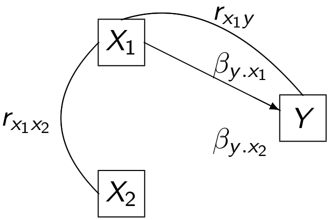
Orthogonal predictors



Correlated predictors



Suppressive predictors

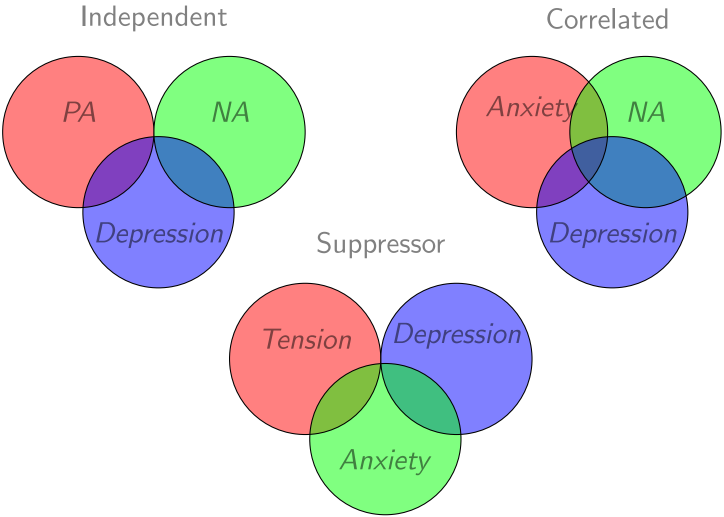


$$\beta_{y.x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r_{x_1 x_2}^2}$$

$$\beta_{y.x_2} = \frac{r_{x_2 y} - r_{x_1 x_2} r_{x_1 y}}{1 - r_{x_1 x_2}^2}$$

$$R^2 = r_{x_1 y}\beta_{y.x_1} + r_{x_2 y}\beta_{y.x_2}$$

Three basic cases: Theoretical examples



Find the regression of rated Prelim score on GREV

```
> mod1 <- lm(GPA~GREV,data=mydata)
> summary(mod1)

Call:
lm(formula = GPA ~ GREV, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45807 -0.32322  0.00107  0.32811  1.44850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.0117292  0.0694343   43.38   <2e-16 ***
GREV        0.0019839  0.0001359   14.60   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.4558 on 998 degrees of freedom
Multiple R-squared: 0.176,       Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Regression on z transformed data

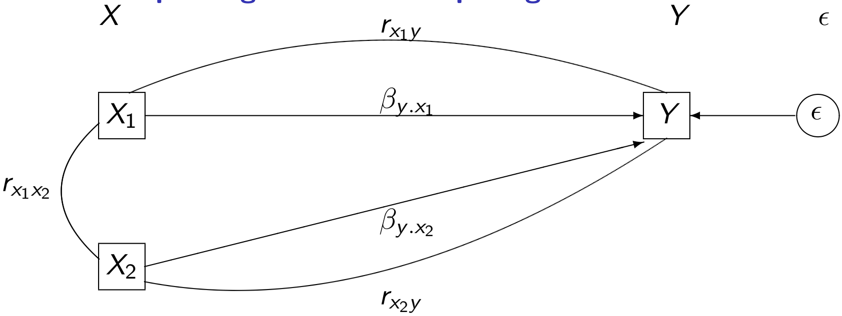
```
> mod2 <- lm(GPA~GREV,data=z.data)
> summary(mod2)

Call:
lm(formula = GPA ~ GREV, data = z.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90526 -0.64404  0.00213  0.65377  2.88619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.888e-17  2.872e-02    0.00        1
GREV        4.195e-01  2.873e-02   14.60   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.9082 on 998 degrees of freedom
Multiple R-squared: 0.176,        Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Note that the slope is the same as the correlation.

```
> mod3 <- lm(GPA~GREV,data=cent)
> summary(mod3)

Call:
lm(formula = GPA ~ GREV, data = cent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45807 -0.32322  0.00107  0.32811  1.44850

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.332e-17  1.441e-02    0.00        1
GREV         1.984e-03  1.359e-04   14.60   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.4558 on 998 degrees of freedom
Multiple R-squared: 0.176,       Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Note that the slope of the centered data is in the same units as the raw data, just the intercept has changed.

Multiple Regression: decomposing correlations



$$r_{x_1 y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1 x_2}\beta_{y.x_2}}^{indirect}$$

$$r_{x_2 y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1 x_2}\beta_{y.x_1}}_{indirect}$$

$$\beta_{y.x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r_{x_1 x_2}^2}$$

$$\beta_{y.x_2} = \frac{r_{x_2 y} - r_{x_1 x_2} r_{x_1 y}}{1 - r_{x_1 x_2}^2}$$

$$R^2 = r_{x_1 y}\beta_{y.x_1} + r_{x_2 y}\beta_{y.x_2}$$

2 predictors

```
> summary(lm(GPA ~ GREV + GREQ , data= cent))

Call:
lm(formula = GPA ~ GREV + GREQ, data = cent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42442 -0.33228  0.00616  0.32465  1.43765

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.651e-17  1.435e-02   0.000  1.00000
GREV         1.534e-03  1.976e-04   7.760 2.10e-14 ***
GREQ         6.314e-04  2.019e-04   3.127  0.00182 **
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.4538 on 997 degrees of freedom
Multiple R-squared: 0.184,       Adjusted R-squared: 0.1823
F-statistic: 112.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

Multiple R with z transformed data

Do the same regression, but on the z transformed data. The units are now in correlation units.

```
> z.data <- data.frame(scale(my.data))
>  summary(lm(GPA ~ GREV + GREQ , data= z.data))

Call:
lm(formula = GPA ~ GREV + GREQ, data = z.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.83821 -0.66208  0.01228  0.64688  2.86457 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.205e-17  2.860e-02   0.000  1.00000    
GREV        3.242e-01  4.179e-02   7.760 2.10e-14 ***
GREQ        1.306e-01  4.179e-02   3.127  0.00182 ** 
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1 

Residual standard error: 0.9043 on 997 degrees of freedom
Multiple R-squared: 0.184,	Adjusted R-squared: 0.1823 
F-statistic: 112.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

The 3 correlations produce the beta weights

```
> R.small <- cor(my.data[c(2,3,8)])
> round(R.small,2)
     GREV GREQ  GPA
GREV 1.00 0.73 0.42
GREQ 0.73 1.00 0.37
GPA  0.42 0.37 1.00
> solve(R.small[1:2,1:2])
          GREV      GREQ
GREV  2.133188 -1.554768
GREQ -1.554768  2.133188
> beta <- solve(R.small[1:2,1:2],
    R.small[3,1:2])
> beta
     GREV      GREQ
0.3242492 0.1306439
> beta.1 <- (.42 - .73*.37)/(1-.73^2)
> beta.1
[1] 0.3209163
> beta.2 <- (.37 - .73 * .42)/(1-.73^2)
> beta.2
[1] 0.1357311
```

- Find the correlation matrix
- Display it to two decimals
- Find the inverse of GREV and GREQ correlations
- Show them
- Find the beta weights by solving the matrix equation
- show them
- Find the beta weights by using the formula
- Show them

3 predictors, no interactions

Use three predictors, but print it with only 2 decimals

```
> print(summary(lm(GPA ~ GREV + GREQ + GREA , data= cent)),digits=3)

Call:
lm(formula = GPA ~ GREV + GREQ + GREA, data = cent)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2668 -0.3038  0.0073  0.3051  1.3022 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.89e-17   1.35e-02    0.00  1.00000    
GREV         6.66e-04   2.00e-04    3.32  0.00092 ***
GREQ         7.75e-05   1.96e-04    0.40  0.69233    
GREA         2.08e-03   1.81e-04   11.52  < 2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.427 on 996 degrees of freedom
Multiple R-squared: 0.28,       Adjusted R-squared: 0.278 
F-statistic:  129 on 3 and 996 DF,  p-value: <2e-16
```

3 predictors, no interactions

Use three predictors, but just the middle 200 subjects

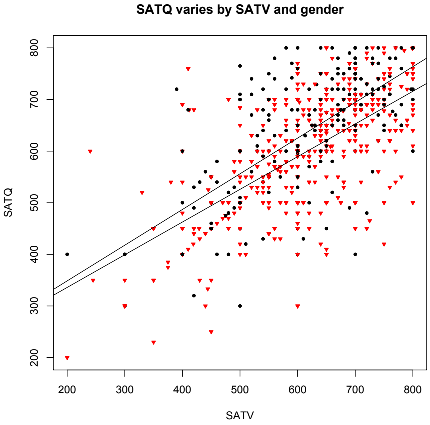
```
> mod4 <- lm(GPA ~ GREV + GREQ + GREA , data= cent[400:600,])
> summary(mod4)

Call:
lm(formula = GPA ~ GREV + GREQ + GREA, data = cent[400:600, ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.03553 -0.30799 -0.00889  0.29320  1.20228

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0397399  0.0310412   1.280    0.202
GREV        0.0004706  0.0004530   1.039    0.300
GREQ        0.0005236  0.0004515   1.160    0.248
GREA        0.0017904  0.0004360   4.107 5.88e-05 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.4394 on 197 degrees of freedom
Multiple R-squared: 0.2259,	Adjusted R-squared: 0.2141
F-statistic: 19.16 on 3 and 197 DF,  p-value: 6.051e-11
```

Interaction terms are just products in regression

- To interpret all effects, the data need to be 0 centered.
  - This makes the main effects orthogonal to the interaction term.
  - Otherwise, need to compare model with and without interactions
- Graph the results in non-standardized form
- Consider a real data set of SAT V, SAT Q and Gender

```
> data(sat.act)
>  colors=c("black","red")          #choose some nice colors
>  symb=c(19,25)
>  colors=c("black","red")          #choose some nice colors
> with(sat.act,plot(SATQ~SATV,pch=symb[gender], col=colors[gender],
  bg=colors[gender],cex=.6,main="SATQ varies by SATV and gender"))
>  by(sat.act,sat.act$gender,function(x)
              abline(lm(SATQ~SATV,data=x)))
```

An example of an interaction plot




```
>  data(sat.act)
>   c.sat <- data.frame(scale(sat.act,scale=FALS
>   summary(lm(SATQ~SATV * gender,data=c.sat))

Call:
lm(formula = SATQ ~ SATV * gender, data = c.sat)

Residuals:
     Min       1Q   Median       3Q      Max
-294.423  -49.876    5.577   53.210  291.100

Coefficients:
             Estimate Std. Error t value Pr(>|t|
(Intercept)  -0.26696    3.31211  -0.081    0.93
SATV          0.65398    0.02926  22.350  < 2e-1
gender      -36.71820    6.91495  -5.310 1.48e-0
SATV:gender  -0.05835    0.06086  -0.959    0.33
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05

Residual standard error: 86.79 on 683 degrees of
  (13 observations deleted due to missingness)
Multiple R-squared: 0.4391,        Adjusted R-sq
F-statistic: 178.3 on 3 and 683 DF,  p-value: <
```

Interaction of Anxiety with Verbal

```
> mod5 <- lm(GPA ~ GREV * Anx,data=cent)
> summary(mod5)

Call:
lm(formula = GPA ~ GREV * Anx, data = cent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49677 -0.31527 -0.00054  0.31223  1.32156

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.375e-04  1.395e-02  -0.017    0.986
GREV         1.996e-03  1.316e-04  15.167  < 2e-16 ***
Anx         -1.131e-02  1.414e-03  -7.997 3.51e-15 ***
GREV:Anx     2.219e-05  1.377e-05   1.612    0.107
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.4412 on 996 degrees of freedom
Multiple R-squared: 0.2294,      Adjusted R-squared: 0.227
F-statistic: 98.81 on 3 and 996 DF,  p-value: < 2.2e-16
```

Testing for the significance of correlations

```
> corr.test(sat.act)

Call:corr.test(x = sat.act)
Correlation matrix
          gender education   age   ACT  SATV  SATQ
gender      1.00      0.09 -0.02 -0.04 -0.02 -0.17
education   0.09      1.00  0.55  0.15  0.05  0.03
age        -0.02      0.55  1.00  0.11 -0.04 -0.03
ACT        -0.04      0.15  0.11  1.00  0.56  0.59
SATV       -0.02      0.05 -0.04  0.56  1.00  0.64
SATQ       -0.17      0.03 -0.03  0.59  0.64  1.00
Sample Size
          gender education age ACT SATV SATQ
gender       700       700 700 700  700  687
education    700       700 700 700  700  687
age          700       700 700 700  700  687
ACT          700       700 700 700  700  687
SATV         700       700 700 700  700  687
SATQ         687       687 687 687  687  687
Probability values (Entries above the diagonal are adjusted for multip
          gender education  age  ACT SATV SATQ
gender      0.00      0.17 1.00 1.00    1    0
education   0.02      0.00 0.00 0.00    1    1
age         0.58      0.00 0.00 0.03    1    1
```

Various tests of significance

1. Is the correlation different from 0? `cor.test`, `corr.test` (for more than two variables)
2. Does a correlation differ from another correlation, `r.test` with or without a third variable.
3. Does a correlation matrix differ from an Identity matrix? `cortest`
4. Bootstrapping confidence intervals for correlations `cor.ci`

The correlation coefficient

1. Perhaps the most powerful and useful statistic ever developed
2. Special cases of the correlation are used throughout statistics.
3. The basic concepts of correlation are very straight forward
4. Many ways to be misled with correlations.

Avery, C. N., Glickman, M. E., Hoxby, C. M., & Metrick, A. (2013). A revealed preference ranking of u.s. colleges and universities. *The Quarterly Journal of Economics*, *128*(1), 425–467.

Bernoulli, D. (1954/1738). Exposition of a new theory on the measurement of risk ("Specimen theoriae novae de mensura sortis," Commentarii Academiae Scientiarum Imperialis Petropolitanae 5, St. Petersburg 175-92.) translated by Louise C. Sommer. *Econometrica*, *22*(1), 23–36.

Burchard, U. (2004). The sclerometer and the determination of the hardness of minerals. *Mineralogical Record*, *35*, 109–120.

Coombs, C. (1964). *A Theory of Data*. New York: John Wiley.

Fechner, Gustav Theodor (H.E. Adler, T. (1966/1860). *Elemente der Psychophysik (Elements of psychophysics)*. Leipzig: Breitkopf & Hartel.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Ozer, D. J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments of personality. *Journal of Personality*, *61*(4), 739–767.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philisopical Transactions of the Royal Society of London. Series A*, *187*, 254–318.

Pearson, K. & Heron, D. (1913). On theories of association. *Biometrika*, *9*(1/2), 159–315.

Rossi, G. B. (2007). Measurability. *Measurement*, *40*(6), 545 – 562.

Sinn, H. W. (2003). Weber's law and the biological evolution of risk preferences: The selective dominance of the logarithmic utility function, 2002 geneva risk lecture. *Geneva Papers on Risk and Insurance Theory*, *28*(2), 87–100.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Weber, E. H. (1834b). *De pulsu, resorptione, auditu et tactu. Annotationes anatomicae et physiologicae*. Leipzig: Kohler.

Weber, E. H. (1948/1834a). Concerning touch, 1834. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 155–156). East Norwalk, CT: Appleton-Century-Crofts Appleton-Century-Crofts Print.