Measuring Association and Dimensionality

Wednesday, September 6, 2006

Statistics for Psychosocial Research

Lecture 2

Statistical topics for you to review on your own….

- Means, medians, proportions
- Confidence intervals
- T-tests, Z-tests
- Chi-square tests
- ANOVA
- Linear and logistic regression

Today’s Topics

- Brief discussion about measuring L.V.
- Measuring Association:
  - Covariance
  - Pearson correlation
  - Spearman correlation
  - Measuring associations with non-linear data
    - tetrachoric / polychoric correlation
    - Odds ratios
  - Association matrices
  - Other commonly used measures of association and “dis-association.”
- Dimensionality
  - Of items
  - Of constructs

Critical Ideas on Measurement

- Measurement of “latent” variables
- Latent variable ≈ construct ≈ factor ≈ domain
- Measurement = rules for assigning symbols to objects to numerically represent quantities of attributes
- Measuring attributes of a person
- Abstract nature
- Focus tends to be on constructs (social-psychological) that are based on strong theoretical framework
- REQUIRES MULTIPLE ITEMS PER CONSTRUCT

Critical Ideas on Measurement

- REQUIRES MULTIPLE ITEMS PER CONSTRUCT
- A part of scale development, we assess things like reliability and validity
- The way we do that is using the associations between the items
- Makes sense: items used to measure the same construct should be related!

Measuring Associations

- Our specific goal: Evaluate associations between pairs of variables being used to measure a construct of interest
- Examples of associations in latent constructs:
  - depression: sleeping problems ~ guilt?
  - disability: time to walk 10 m ~ speed to walk up 10 steps?
  - schizophrenia: delusions ~ hallucinations?
  - SES: education ~ income?

Associations in Psychosocial Research

- Crucial to the process of defining a construct
  - (1) "too" associated (i.e. redundant)?
  - (2) not associated at all?
    - not appropriately describing "construct"
    - measuring different dimensions of "construct" (e.g. positive versus negative symptoms of schizophrenia)

Associations between variables affect….

- Reliability
- Validity
- Factor Analysis
- Latent Class Analysis
- Structural Equation Models

→ Measurement of Associations is VERY important!

Variance and Covariance

- (Sample) Variance: Measures variability in one variable, X.

$$s_x^{\ 2} = \tfrac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \cong \sigma_x^{\ 2}$$

- (Sample) Covariance: Measures how two variables, X and Y, covary.

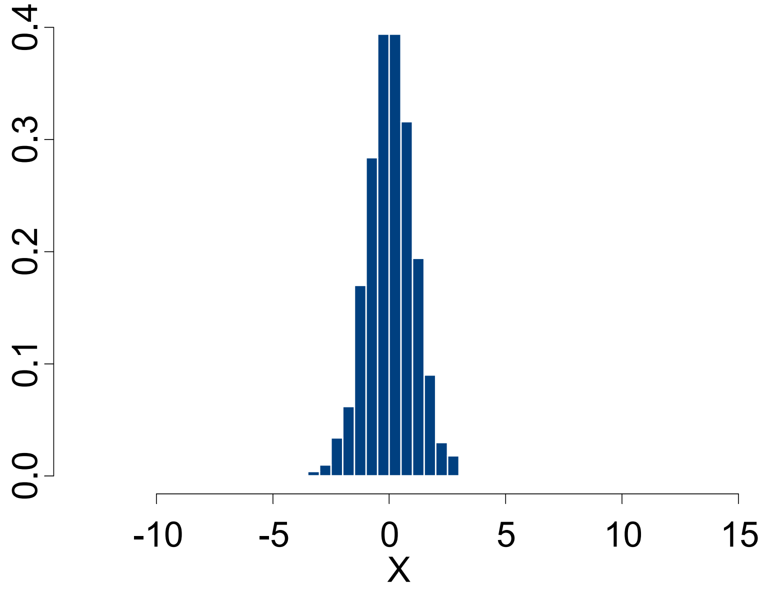
$$s_{xy} = \tfrac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \cong \sigma_{xy}$$

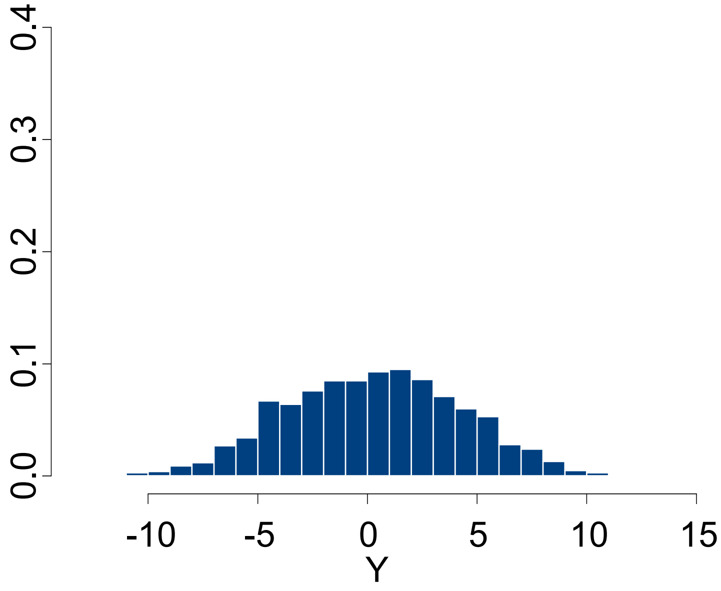
Covariances and Variances

$$Cov(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

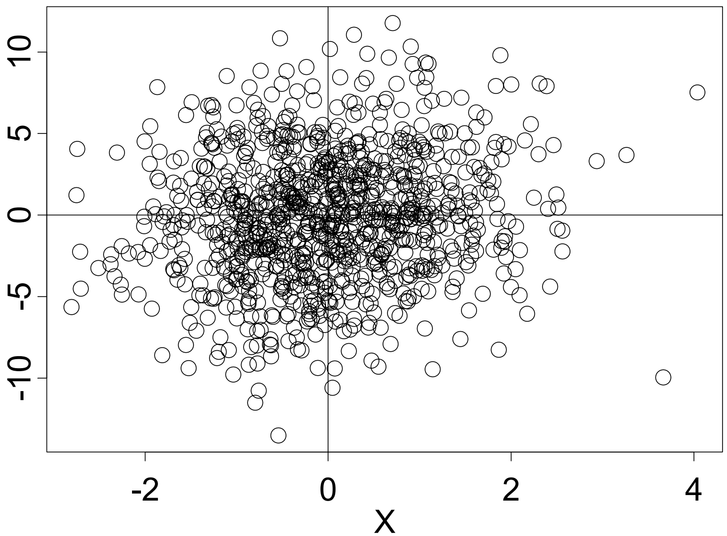
$$\begin{aligned}
Cov(X+Y) &= E(X+Y)^2 - [E(X+Y)]^2 \\
&= E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2 \\
&= E(X^2) + 2E(XY) + E(Y^2) - [E(X)]^2 - 2E(X)E(Y) - [E(Y)]^2 \\
&= \{E(X^2) - [E(X)]^2\} + \{E(Y^2) - [E(Y)]^2\} + 2\{E(XY) - E(X)E(Y)\} \\
&= Var(X) + Var(Y) + 2Cov(X,Y)
\end{aligned}$$

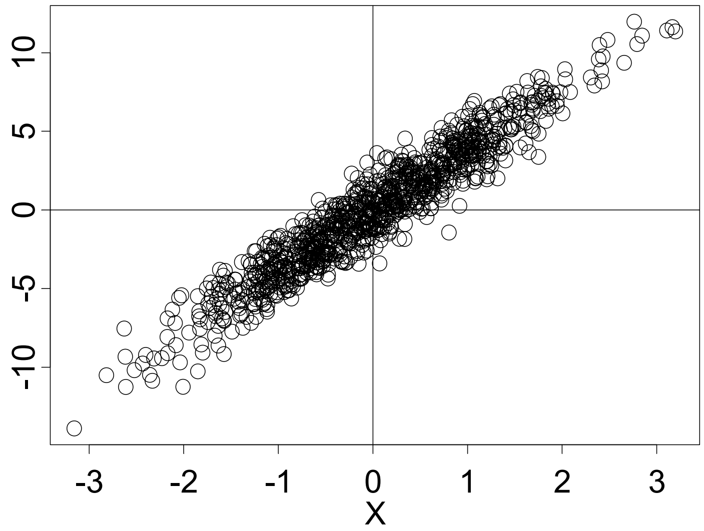
Examples of Variance





Examples of Covariance





Correlation, *r*

Correlation (i.e. “Pearson” correlation) is a scaled version of covariance

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

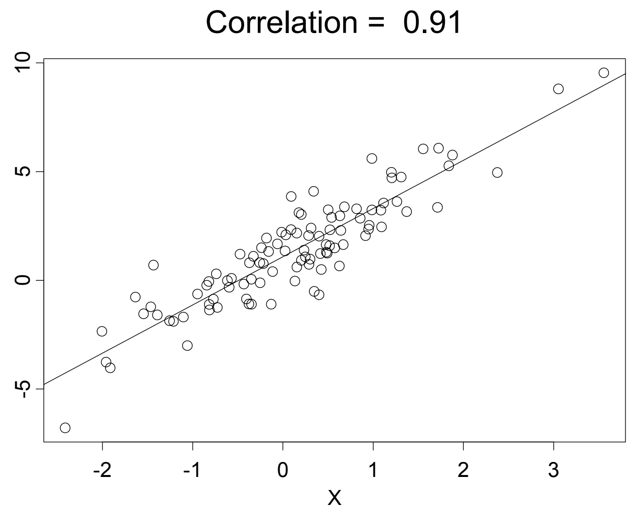
$-1 \leq r \leq 1$

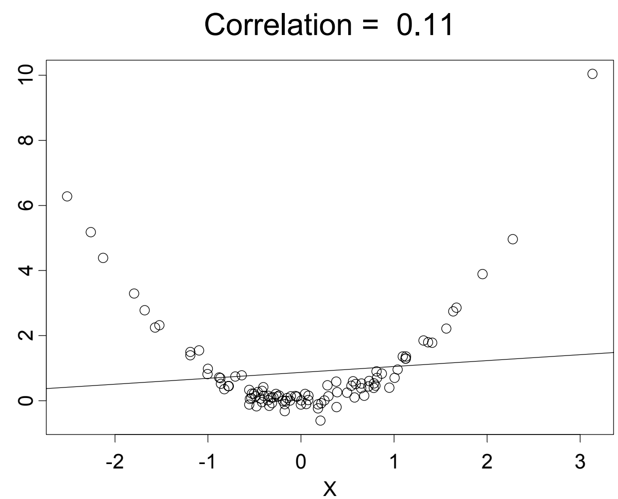
$r = 1$	perfect positive correlation
$r = -1$	perfect negative correlation
$r = 0$	uncorrelated

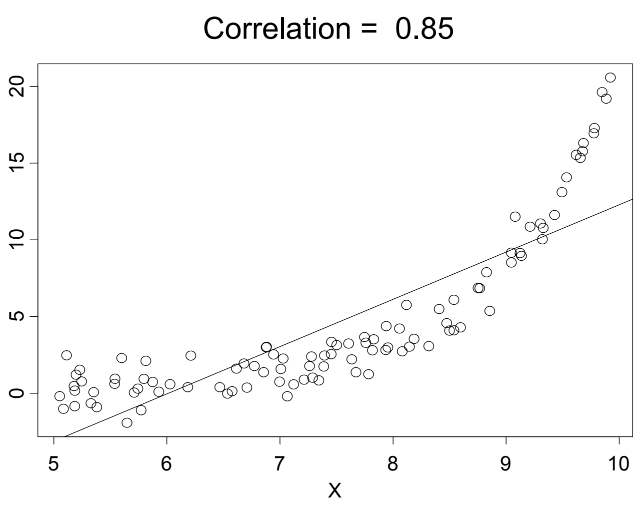
Covariance and Correlation

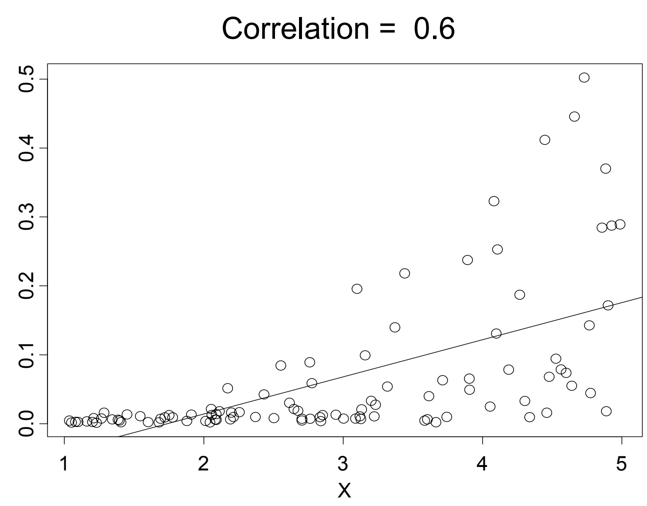
- Consider four different correlations:
  - 0.11
  - 0.60
  - 0.85
  - 0.91
- Which of the above indicates strongest association between two variables?

SCATTERPLOTS:
Importance of Looking at Your Data!

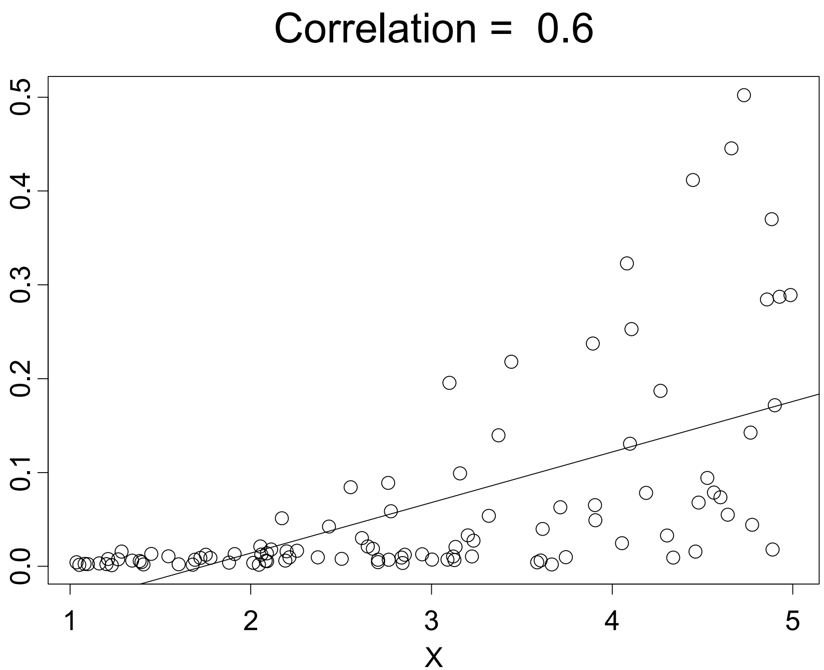






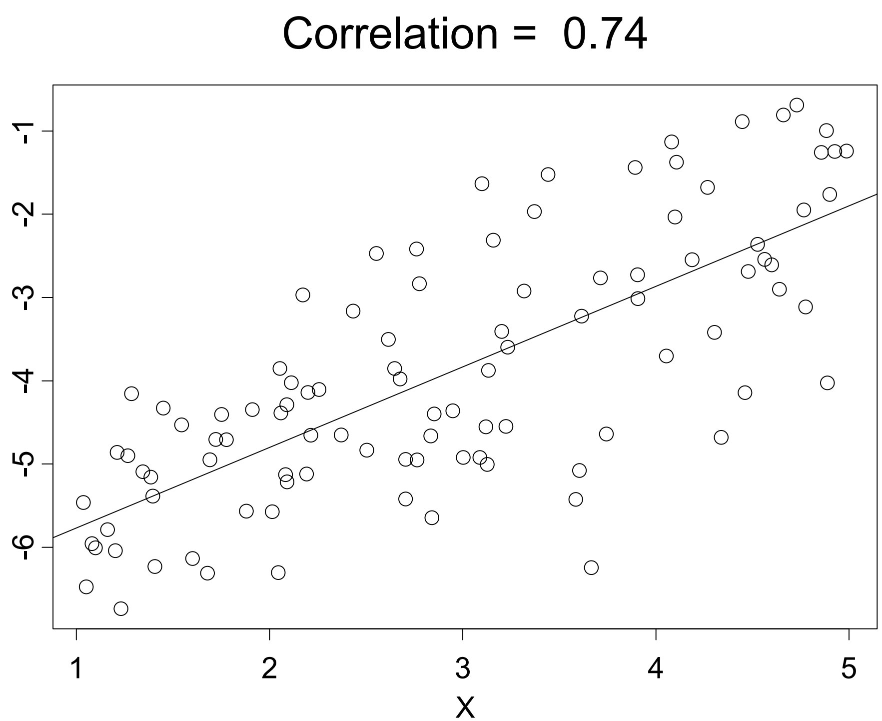


Transforming Variables Can Assist in Obtaining Linear Relationship

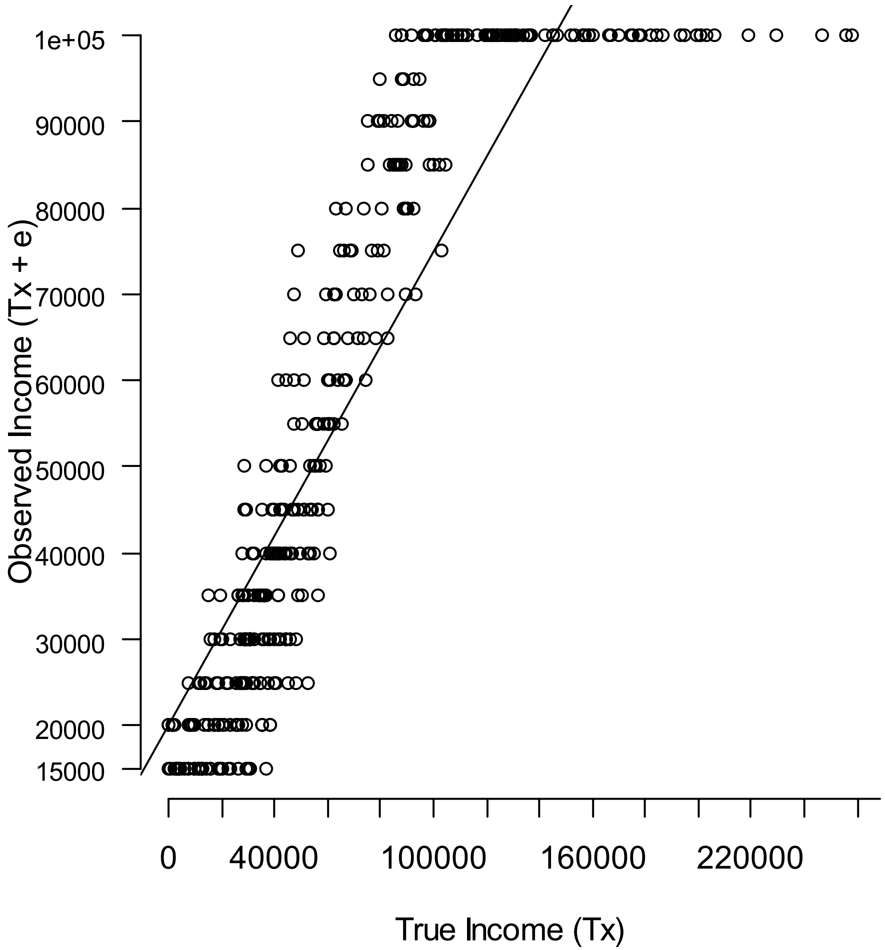


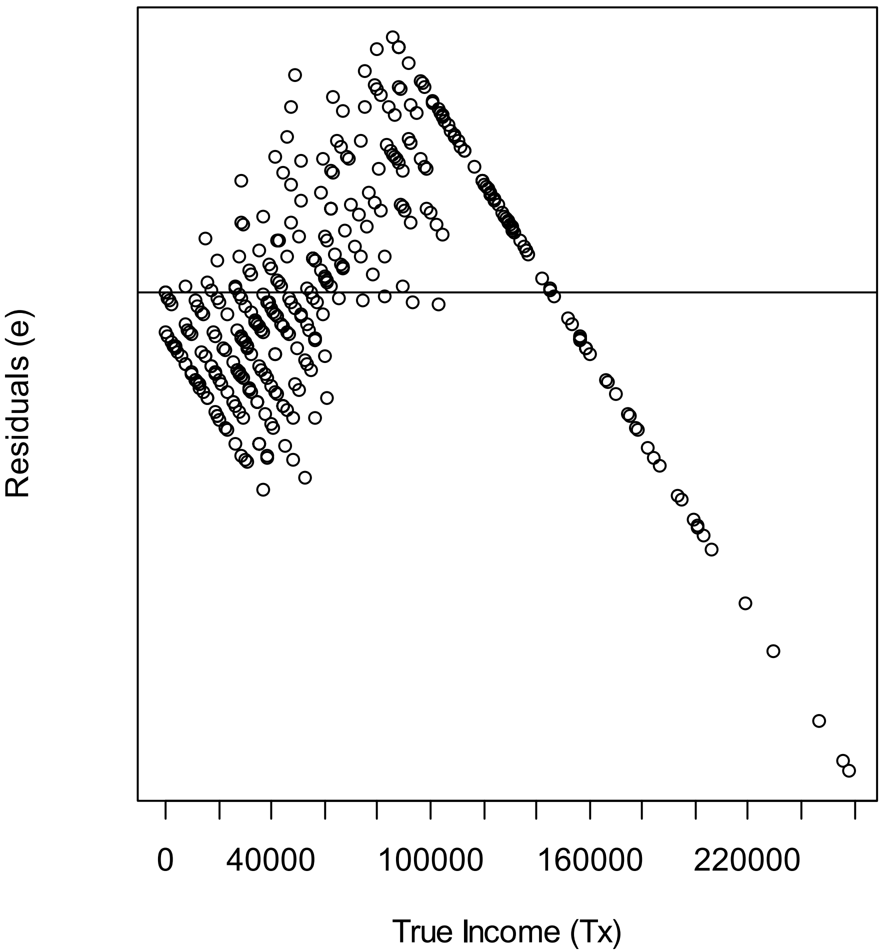
Y ~ X

log(Y) ~ X



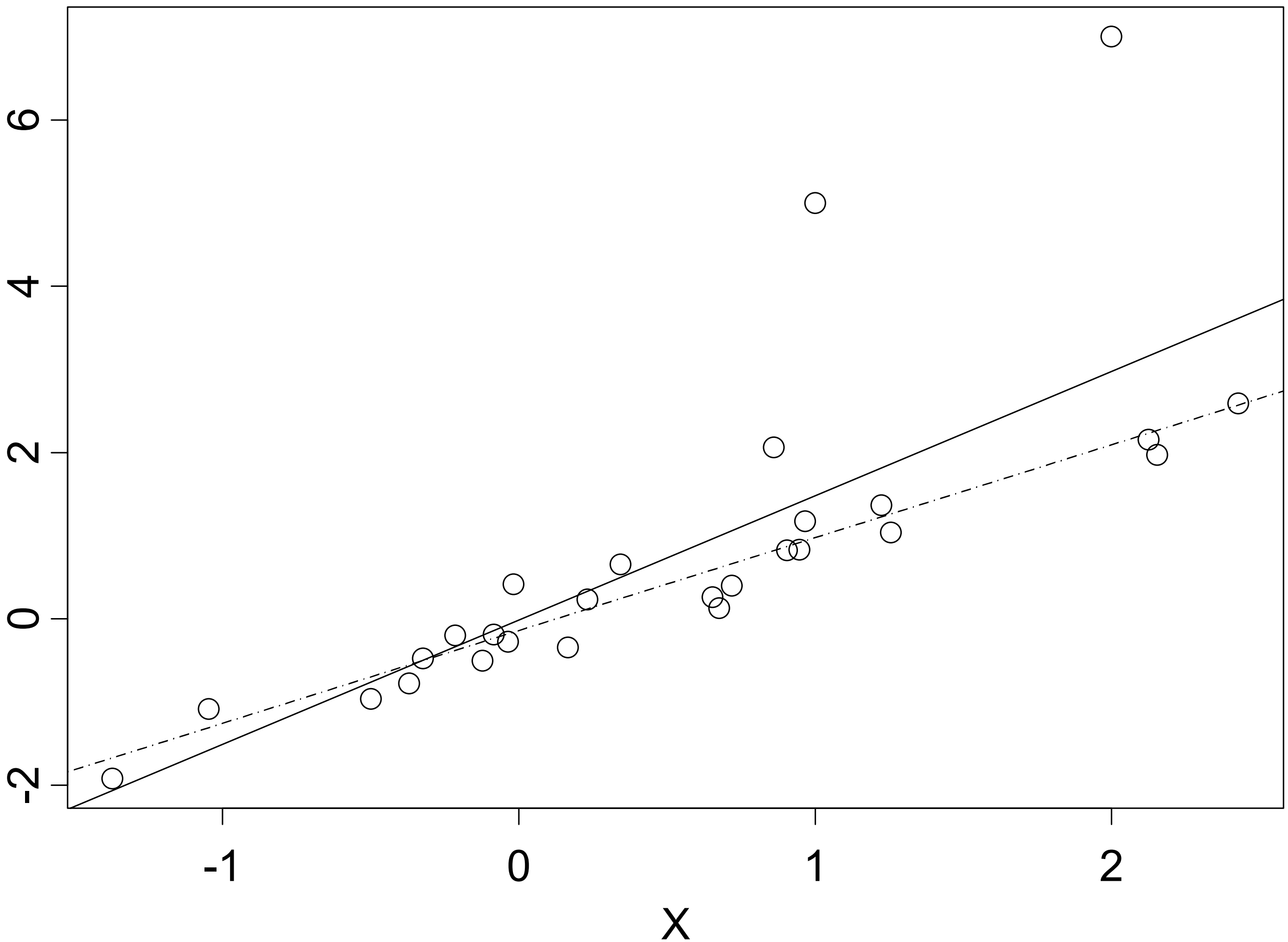
Floor and Ceiling Effects





Outliers

Corr = 0.77 : Corr = 0.95



Covariance and Correlation

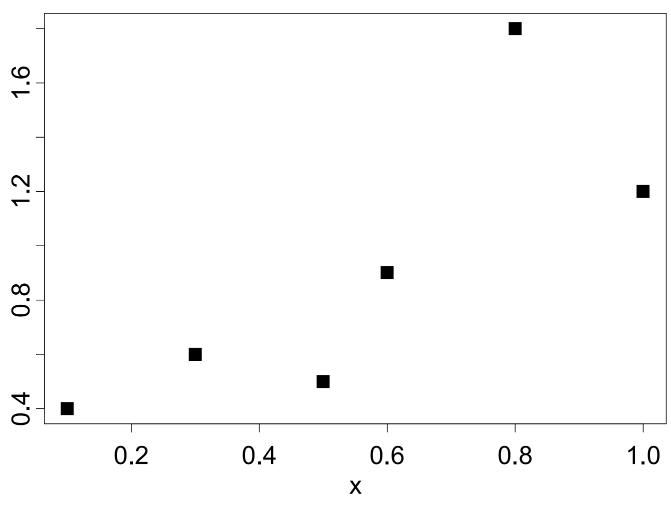
- When are they appropriate measures of association?
- What type of association do they describe?
- What is a drawback of transforming variable so that relationship is linear?

Spearman Correlation

- Use when:
  - skewed data
  - outliers
  - sparse data
- Effect:
  - downweights outliers
  - smoothes a curve to a straight line
- If relationship is *already* linear, then what?

Spearman Correlation

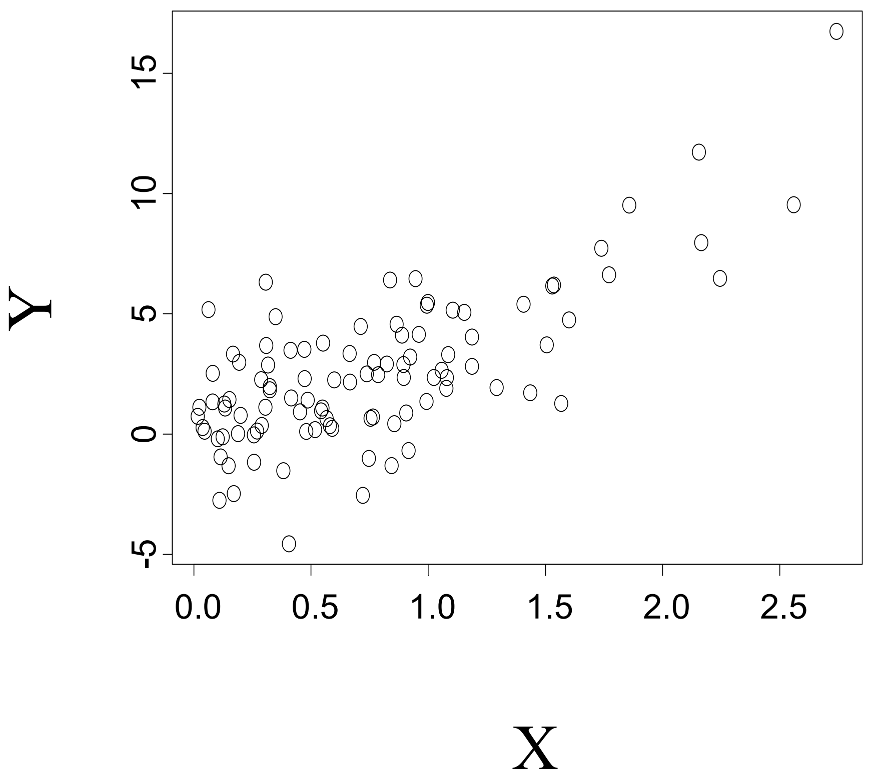
- Method:
  - sort x and y
  - replace data with ranks
  - calculate pearson correlation on ranks.



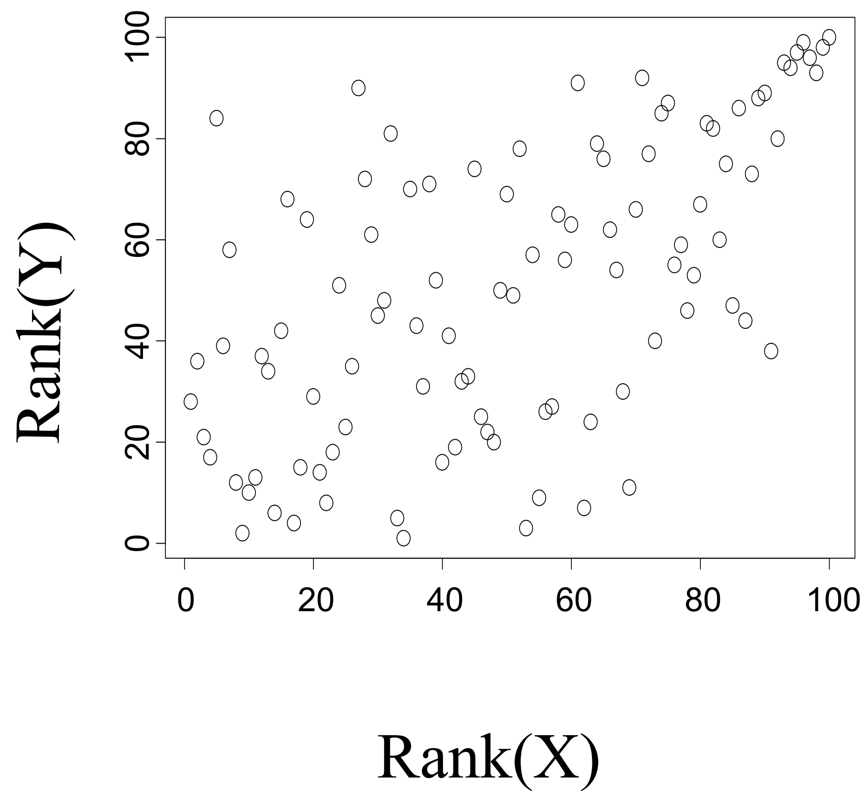
data		ranks	
x	y	x^*	y^*
0.1	0.4	1	1
0.3	0.6	2	3
0.5	0.5	3	2
0.6	0.9	4	4
0.8	1.8	5	6
1.0	1.2	6	5
r=0.79		r=0.89	

Spearman Correlation

Pearson r = 0.72



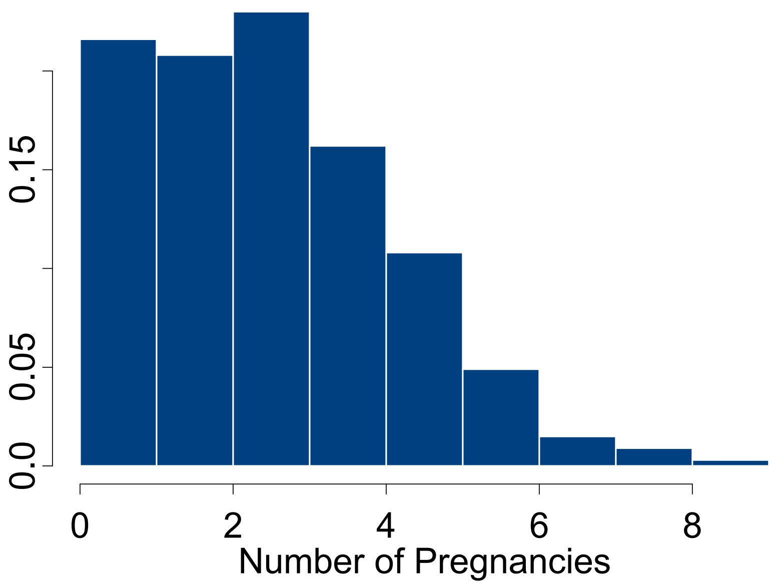
Spearman r = 0.59

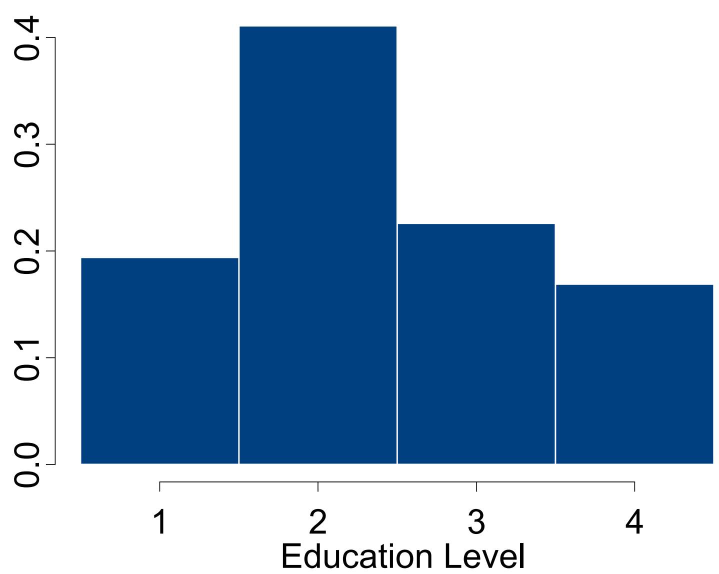


Problems with Correlation/Covariance between variables

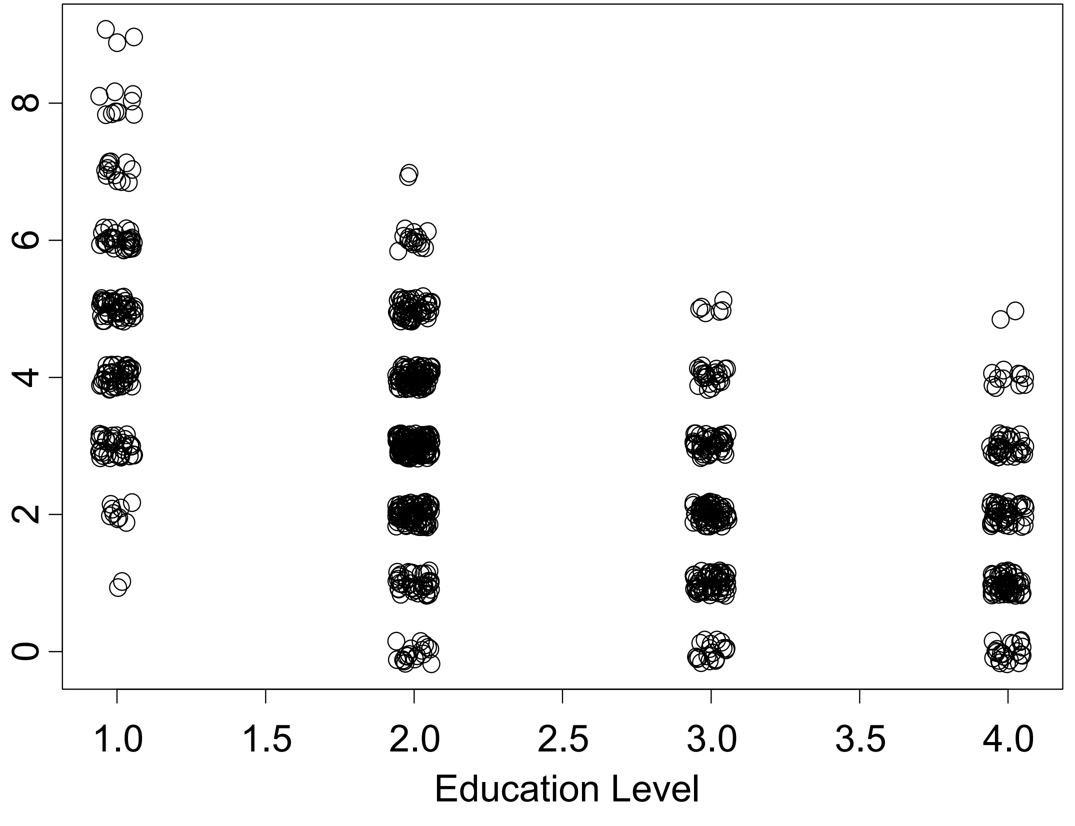
What if one (or both) variable(s) is (are) not really continuous?

e.g. number of pregnancies and education level



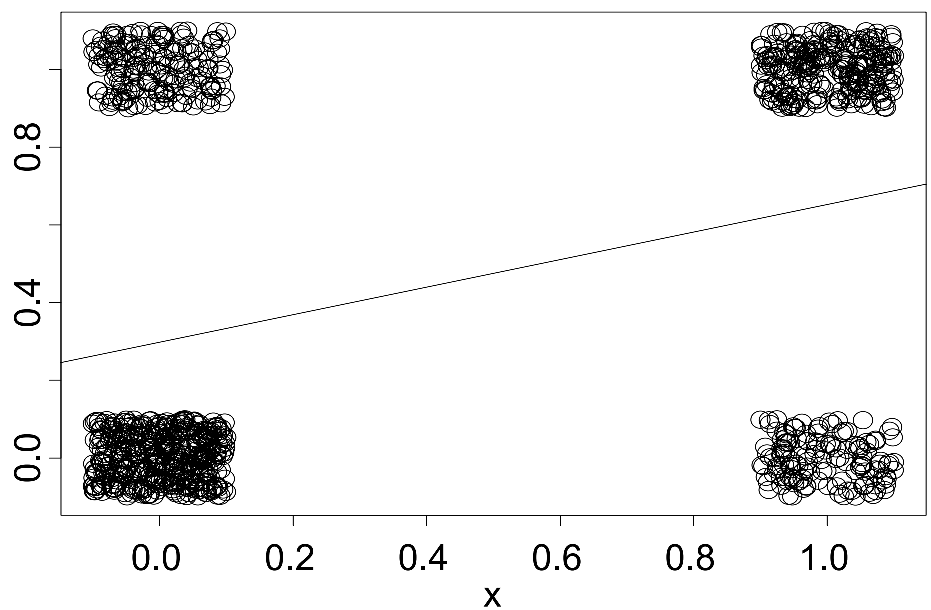


Is correlation appropriate?



Other issues

- Correlation assumes continuous variables
- Ordinal: Takes finite number of values
  - e.g. on a scale of 1 to 5
- Binary: $r = 0.35$



Binary Example: Disability

- Two types of association
  - redundancy: b and c cells are close to 0 (i.e. disagreement is small).
  - hierarchy: either b OR c is close to 0, but other is not.
  - Pearson correlation mixes up association and similarity of "marginal"distribution

		Difficulty Walking 1/4 Mile: No	Difficulty Walking 1/4 Mile: Yes	
Difficulty Walking 1 mile	No	40	0	40
	Yes	40	20	60
		80	20	100

- Consequences: If hierarchy is relevant, you get low reliability, consistency, and misleading internal validity by using pearson correlation.

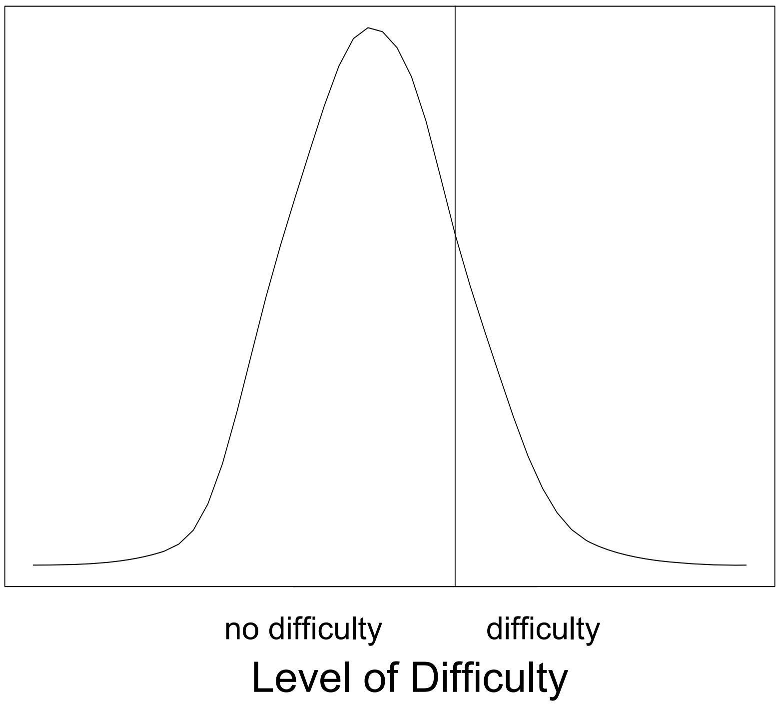
Some Alternative Measures

- Tetrachoric Correlation
  - binary variables
- Polychoric Correlation
  - ordinal variables
- Odds Ratio
  - binary variables

Tetrachoric Correlation

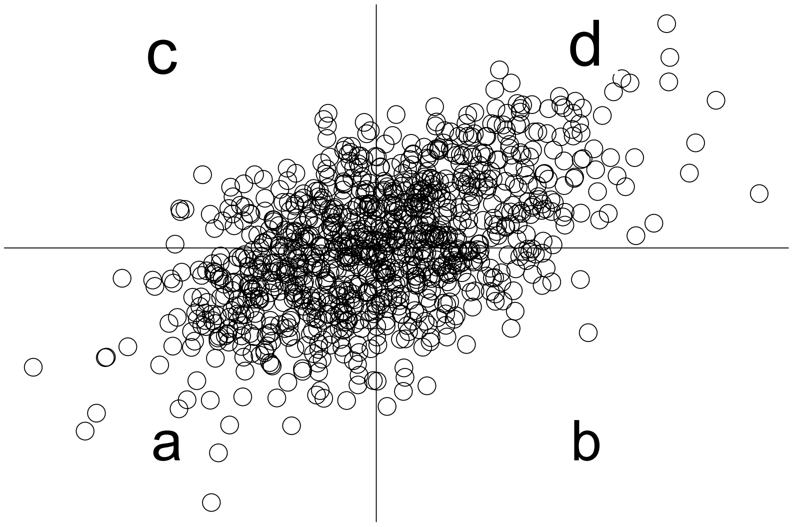
- Estimates what the correlation between two binary variables *would be* if you could measure variables on a continuous scale.
- Example: difficulty walking up 10 steps and difficulty lifting 10 lbs.

Difficulty Walking Up 10 Steps



Tetrachoric Correlation

- Assumes that both "traits" are normally distributed
- Correlation, *r*, measures how narrow the ellipse is.
- a, b, c, d are the proportions in each quadrant



Tetrachoric Correlation

For φ = ad/bc,

Approximation 1:

$$Q = \frac{\varphi - 1}{\varphi + 1}$$

Approximation 2 (Digby):

$$Q = \frac{\varphi^{3/4} - 1}{\varphi^{3/4} + 1}$$

Tetrachoric Correlation

- Example:

  Tetrachoric correlation = 0.61

  Pearson correlation = 0.41
- TC Interpretation?
  - Same as Pearson correlation.
- As good as Pearson correlation?
  - **Makes assumptions that can't be tested**
  - **Assumes threshold is the same across people**
  - **Strong assumption that underlying quantity of interest is truly continuous**

		Difficulty Walking Up 10 Steps: No	Difficulty Walking Up 10 Steps: Yes	
Difficulty Lifting 10 lb.	No	40	10	50
	Yes	20	30	50
		60	40	100

Odds Ratio

- Measure of association between two binary variables
- Risk associated with x given y.
- Example:

  odds of difficulty walking up 10 steps to the odds of difficulty lifting 10 lb:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

$$\cong \frac{ad}{bc}$$

$$= \frac{(40)(30)}{(20)(10)} = 6$$

Odds Ratio

		Difficulty Walking 1/4 Mile		
		No	Yes	
Difficulty Walking 1 mile	No	40	0	40
	Yes	40	20	60
		80	20	100

$$\frac{ad}{bc} = \frac{(40)(20)}{(40)(0)} = \infty$$

Other option:

- continuity corrections.

Problem with continuity correction:

- somewhat arbitrary what value to use for correction.

Pros and Cons

- Tetrachoric correlation
  - same interpretation as Spearman and Pearson correlations
  - "difficult" to calculate exactly
  - makes (strong) assumptions
- Odds Ratio
  - easy to understand, but no "perfect" association that is manageable (i.e. $\{\infty, 0\}$)
  - easy to calculate
  - not comparable to correlations
- May give you different results/inference!

Association Matrices: Age, income, education

- Covariance Matrix

	grade	income	age
grade	6.61	28.18	-5.77
income	28.18	592.69	-29.10
age	-5.77	-29.10	81.23

- Correlation Matrix

	grade	income	age
grade	1.00	0.45	-0.25
income	0.45	1.00	-0.13
age	-0.25	-0.13	1.00

Association Matrices: depressed mood, sleep problems, fatigue

- Odds Ratio Matrix

	depress	sleep	fatigue
depress	---	8.17	10.91
sleep	8.17	---	16.12
fatigue	10.91	16.12	---

Other measures of association

- "Distances"
  - Euclidean
  - Canberra
  - Manhattan
  - Mahalanobis
- Often used for classification and clustering methods
- Our focus:
  - Cronbach's alpha
  - Factor analysis

  ⇨ Both usually use Pearson correlation (most of the time)

Issues in Dimensionality of Constructs

- **Dimensionality**: concerned with the *homogeneity of items* used to measure a construct
- **Unidimensional** construct: items underlie a single factor
- **Multidimensional** construct: items "tap into" more than one factor
- **Reliability and validity assessment DEPEND STRONGLY on unidimensionality assumption!**

Examples

- Schizophrenia has two (or more) domains of symptoms:
  - Negative symptoms: e.g. lack of energy, social withdrawal
  - Positive symptoms: e.g. hallucinations, thought disorder
- Intelligence has 8 domains (by some definitions)

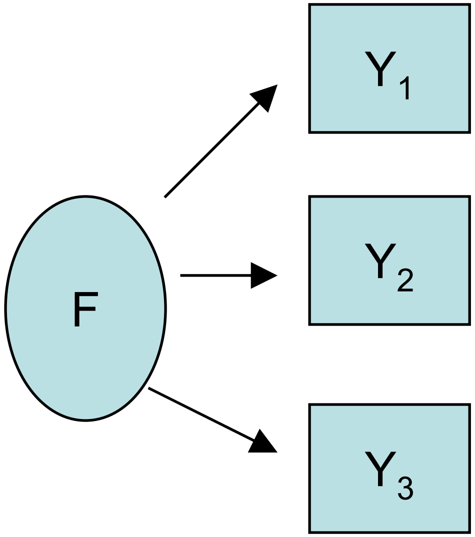
– Linguistic	Logical
– Musical	Spatial
– Kinesthetic	Intrapersonal
– Interpersonal	Naturalistic

Unidimensionality

- F is the latent variable
- Y's are the items measuring F

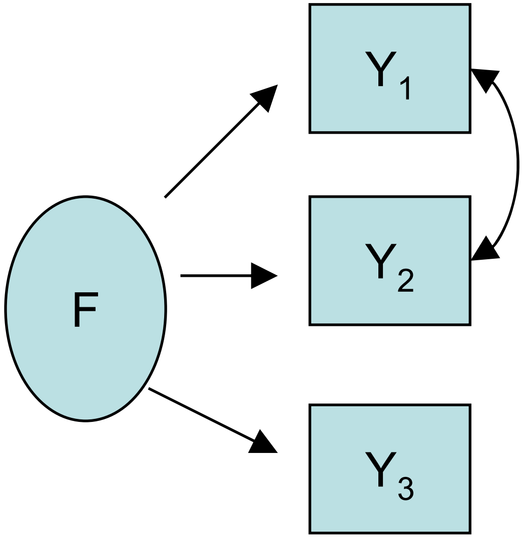
- Unidimensionality of items: F (factor) is responsible for ALL of the associations between the Y variables.

- A set of items is unidimensional if the correlations among them can be accounted for by a single factor



NOT Unidimensional

- Arrows between variables (either straight and uni-directional, or curved and bi-directional) imply 'associations exist'
- Here, Y_1 and Y_2 are associated even after accounting for F
- This set of items is NOT unidimensional
- Implications: there is something else going on….perhaps another factor?



Mathematically

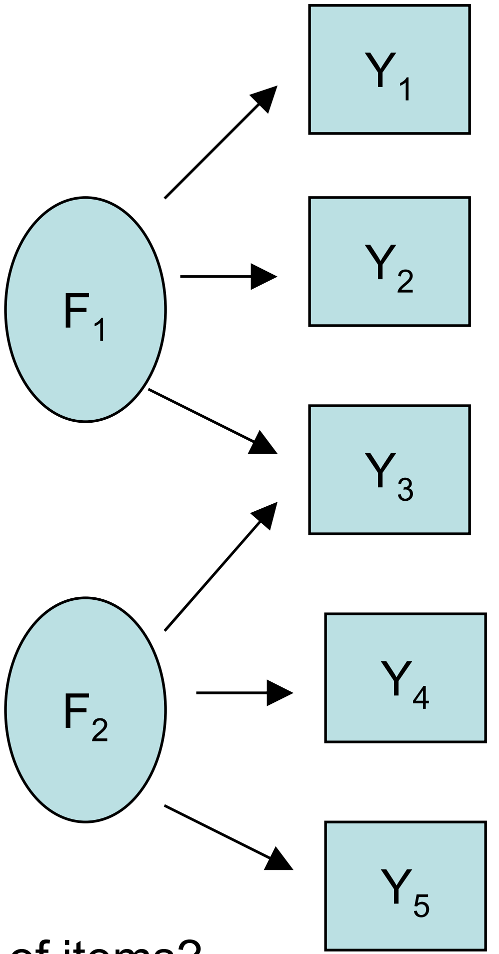
- Partial correlation:

$$corr(x_1, x_2 | F) = \frac{r_{12} - r_{1F} r_{2F}}{\sqrt{(1 - r_{1F}^2)(1 - r_{2F}^2)}}$$

- If the partial correlation between each pair of items is equal to zero, then the set of items are unidimensional.

That sounds "easy", but….

- A ***set of items*** is unidimensional if the correlations among them are accounted for by a single factor
- An ***item*** is unidimensional if it is measure of only a single construct.



Are Y_1, Y_1, and Y_3 a unidimensional set of items?

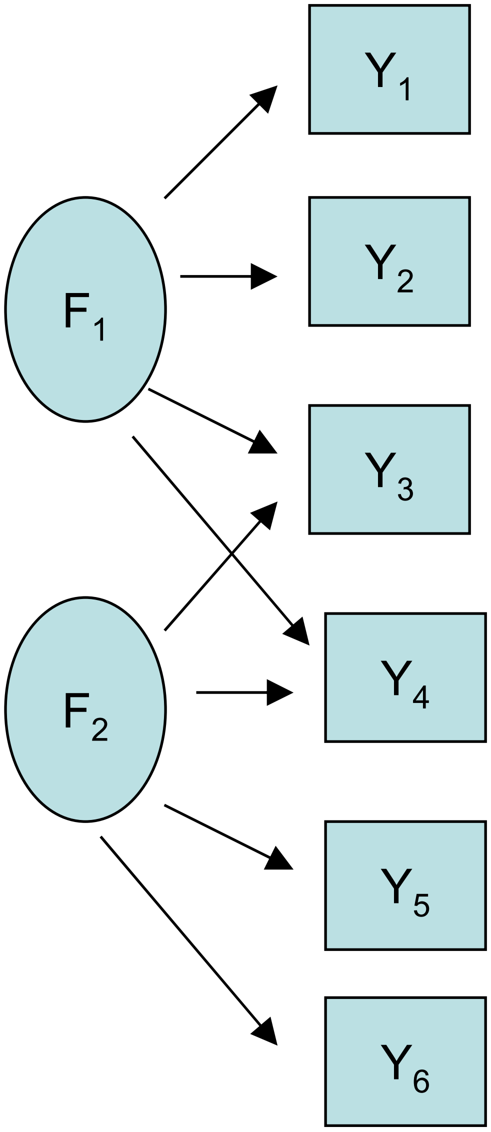
Is Y_3 a unidimensional item?

What about dimensionality of a construct?

- If multiple sets of *n* items from the domain of the construct are taken and the partial correlations among each set are zero, then the ***construct*** is unidimensional.
- BUT, it is possible that a set of items is unidimensional, but the construct or the individual indicators are not unidimensional

What about dimensionality of a construct?

- F_1 is NOT unidimensional
- Why? All sets of items from the domain are not unidimensional.
- The association between Y_3 and Y_4 cannot be accounted for by F_1 alone



Relevance

- Relevance
  - It is generally preferable to have unidimensional items
  - This means that the items load strongly on only one factor
- Testing for dimensionality
  - Factor analysis (exploratory and confirmatory)
  - NOT alpha-coefficient!