Correlational designs

Issues of Reliability, Validity, and Causality
Alternative Explanatory Variables

• A developmental psychologist has noticed that children with bigger feet tend to have greater vocabularies than children with smaller feet?

• This is an example of a simple correlational design. Can you think of a powerful alternative explanation?
Types of data and correlational designs

• Varieties of data
  – Direct
    • Self report of personality/attitudes/values
    • Peer/supervisor/subordinate ratings of performance
    • Ability scales
  – Indirect
    • Implicit measures (e.g., of attitude)
    • Unobtrusive measures
      – Historical, archival
Theory and Theory Testing I: Theory

Construct 1 → Construct 2
Theory and Theory Testing II: Experimental manipulation

Construct 1 -> Construct 2

Manipulation 1

Observation 1
Theory and Theory Testing III: Correlational inference

Construct 1

Observation 1

? \( r_{mo} \)

? ?

Construct 2

Observation 2
Theory and Theory Testing IV: Correlational inference

Construct 1

Construct X

Construct 2

Observation 1

Observation 2

$r_{mo}$
Theory and Theory Testing V: Alternative Explanations

Construct 1

Observation 1

Construct 2

Observation 2
Theory and Theory Testing VI: Eliminate Alternative Explanations

Construct 1

Construct 2

Observation 1

Observation 2
Steps in correlational inference

• Estimate the reliability of the variables
  – Magnitude of correlation is influenced by the reliability of the correlation
  – Varieties of reliability
• Estimate the construct validity of the measures
  – Convergent, Discriminant, Incremental validity
• Consider alternative explanatory variables
Classic Reliability Theory: How well do we measure what ever we are measuring

\[ X_1 \rightarrow L_1 \]

\[ X_2 \rightarrow L_1 \]

\[ X_3 \rightarrow L_1 \]
Classic Reliability Theory:
How well do we measure what ever we are measuring and what is the relationships between latent variables
Observed = True + Error
Observed = True + Error
Estimating true score: regression artifacts

• Consider the effect of reward and punishment upon pilot training:
  – From 100 pilots, reward the top 50 flyers, punish the worst 50.
  – Observation: praise does not work, blame does!
  – Explanation?
Reliability of measurement
(how well does an observation reflect the construct)

\[ r_{12} = r_{c1} \times r_{c2} \]

Assume \( r_{c1} = r_{c2} \) then
\[ r_{c1} = \sqrt{r_{12}} \]

Observed Variance \( _1 = \) Variance \( C_1 + \) Variance \( E_1 \)
\( (V_x = V_c + V_e) \)
\[ r_{12} = \frac{V_c}{V_x} \]
Reliability of measurement
(how well does an observation reflect the construct)
Domain Sampling theory

• Consider a domain (D) of k items relevant to a construct. (E.g., English vocabulary items, expressions of impulsivity). Let Di represent the number of items in D which the ith subject can pass (or endorse in the keyed direction) given all D items. Call this the domain score for subject I. What is the correlation (across subjects) of scores on an item j with the domain scores?
The effect of test length on internal consistency

<table>
<thead>
<tr>
<th>Number of items</th>
<th>Average r</th>
<th>Average r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>2</td>
<td>.33</td>
<td>.18</td>
</tr>
<tr>
<td>4</td>
<td>.50</td>
<td>.31</td>
</tr>
<tr>
<td>8</td>
<td>.67</td>
<td>.47</td>
</tr>
<tr>
<td>16</td>
<td>.80</td>
<td>.64</td>
</tr>
<tr>
<td>32</td>
<td>.89</td>
<td>.78</td>
</tr>
<tr>
<td>64</td>
<td>.94</td>
<td>.88</td>
</tr>
<tr>
<td>128</td>
<td>.97</td>
<td>.93</td>
</tr>
</tbody>
</table>
Find Alpha from correlations

<table>
<thead>
<tr>
<th></th>
<th>q_262</th>
<th>q_1480</th>
<th>q_819</th>
<th>q_1180</th>
<th>1742</th>
</tr>
</thead>
<tbody>
<tr>
<td>q_262</td>
<td>1</td>
<td>0.26</td>
<td>0.41</td>
<td>0.51</td>
<td>0.48</td>
</tr>
<tr>
<td>q_1480</td>
<td>0.26</td>
<td>1</td>
<td>0.66</td>
<td>0.52</td>
<td>0.47</td>
</tr>
<tr>
<td>q_819</td>
<td>0.41</td>
<td>0.66</td>
<td>1</td>
<td>0.41</td>
<td>0.65</td>
</tr>
<tr>
<td>q_1180</td>
<td>0.51</td>
<td>0.52</td>
<td>0.41</td>
<td>1</td>
<td>0.49</td>
</tr>
<tr>
<td>q_1742</td>
<td>0.48</td>
<td>0.47</td>
<td>0.65</td>
<td>0.49</td>
<td>1</td>
</tr>
</tbody>
</table>
Alpha from correlations

- Total variance = sum of all item correlations
  \[ = 14.74570 \]

- total covariances = \( \Sigma \) item variance
  \[ = 9.74570 \]

- average covariance =
  \[ = (\Sigma \text{item variance})/(nvar *(nvar-1)) = .66 \]

- alpha = \( ((\Sigma \text{item variance})/\Sigma \text{Vt})*(nvar *(nvar-1)) \)
  \[ = \text{alpha} = .83 \]
## Facets of reliability

<table>
<thead>
<tr>
<th>Across items</th>
<th>Domain sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Internal consistency</td>
</tr>
<tr>
<td>Across time</td>
<td>Temporal stability</td>
</tr>
<tr>
<td>Across forms</td>
<td>Alternate form reliability</td>
</tr>
<tr>
<td>Across raters</td>
<td>Inter-rater agreement</td>
</tr>
<tr>
<td>Across situations</td>
<td>Situational stability</td>
</tr>
<tr>
<td>Across “tests” (facets unspecified)</td>
<td>Parallel test reliability</td>
</tr>
</tbody>
</table>
Types of reliability

- Items in a test of positive affect correlate .3 with other items of positive affect given at the same time.
- An ability test given in 5th grade correlates .6 with an ability test given in college.
- Baseball batting averages correlate .35 from year to year.
Reliability- Correction for attenuation

\[ r_{xt_x} = \sqrt{r_{xx}} \]

\[ r_{yty} = \sqrt{r_{yy}} \]

\[ \text{Rho} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}} \]
Effect of preschool

• A team of educational psychologists examined the effect of early reading in preschool upon later academic attainment. They randomly selected 20 preschools in Evanston and gave a measure of reading skill to 200 children (mean = 5.0, sd =1.0). They followed the progress of the top 50 students (mean score = 6 ) for a year. At the end of the year they compared their sample students to the mean and found the group was no different from the average. They concluded that preschool hurt these students.
Effect of preschool

- A team of educational psychologists examined the effect of early reading in preschool upon later academic attainment. They randomly selected 20 preschools in Evanston and gave a measure of reading skill to 200 children (mean = 5.0, sd = 1.0). They followed the progress of the bottom 50 students (mean score = 4) for a year. At the end of the year they compared their sample students to the mean and found the group was no different from the average. They concluded that preschool helped these students.
Classic reliability - limitation

All of the conventional approaches are concerned with generalizing about individual differences (in response to an item, time, form, rater, or situation) between people. Thus, the emphasis is upon consistency of rank orders. Classical reliability is a function of large between subject variability and small within subject variability. It is unable to estimate the within subject precision for a single person.
The New Psychometrics- Item Response Theory

• Classical theory estimates the correlation of item responses (and sums of item responses, i.e., tests) with domains.

• Classical theory treats items as random replicates but ignores the specific difficulty of the item, nor attempts to estimate the probability of endorsing (passing) a particular item.
Item Response Theory

• Consider the person’s value on an attribute dimension ($\theta_i$).
• Consider an item as having a difficulty $\delta_j$
• Then the probability of endorsing (passing) an item $j$ for person $i$ = $f(\theta_i, \delta_j)$
• $p(\text{correct} \mid \theta_i, \delta_j) = f(\theta_i, \delta_j)$
• What is an appropriate function?
• Should reflect $\delta_j - \theta_i$ and yet be bounded 0,1.
Item Response Theory

- \( p(\text{correct} \mid \theta_i, \delta_j) = f(\theta_i, \delta_j) = f(\delta_j - \theta_i) \)

- Two logical functions:
  - Cumulative normal (see, e.g., Thurstonian scaling)
  - Logistic \( = \frac{1}{1 + \exp(\delta_j - \theta_i)} \) (the Rasch model)
  - Logistic with weight of 1.7
    - \( \frac{1}{1 + \exp(1.7*(\delta_j - \theta_i))} \) approximates cumulative normal
Logistic and cumulative normal

![Graph showing the logistic and cumulative normal distributions. The x-axis represents the latent variable, ranging from -3 to 3, and the y-axis represents the observed probability, ranging from 0.0 to 1.0. The graph demonstrates the characteristic S-shaped curve of the logistic function and the smooth cumulative normal distribution.](image-url)
Item difficulty and ability

- Consider the probability of endorsing an item for different levels of ability and for items of different difficulty.
  - Easy items ($\delta_j = -1$)
  - Moderate items ($\delta_j = 0$)
  - Difficulty items ($\delta_j = 1$)
IRT of three item difficulties
item difficulties = -2, -1, 0, 1, 2
Estimation of ability for a particular person for known item difficulty

• The probability of any pattern of responses \((x_1, x_2, x_3, \ldots, X_n)\) is the product of the probabilities of each response \(\prod(p(x_i))\).

• Consider the odds ratio of a response

\[
\frac{p}{1-p} = \frac{1/(1+\exp(1.7*(\delta_j - \theta_i)))}{1/(1+\exp(1.7*(\delta_j - \theta_i)))} =
\]

\[
\frac{p}{1-p} = \exp(1.7*(\delta_j - \theta_i))
\]

and therefore:

\[
\ln(\text{odds}) = 1.7*(\delta_j - \theta_i) \quad \text{and}
\]

\[
\ln(\text{odds of a pattern}) = 1.7\sum(\delta_j - \theta_i) \quad \text{for known difficulty}
\]
Unknown difficulty

- Initial estimate of ability for each subject (based upon total score)
- Initial estimate of difficulty for each item (based upon percent passing)
- Iterative solution to estimate ability and difficulty (with at least one item difficulty fixed.)
Classical versus the “new”

- Ability estimates are logistic transform of total score and are thus highly correlated with total scores, so why bother?
- IRT allows for more efficient testing, because items can be tailored to the subject.
- Maximally informative items have $p$(passing given ability and difficulty) of .5
- With tailored tests, each person can be given items of difficulty appropriate for them.
Computerized adaptive testing

- CAT allows for equal precision at all levels of ability
- CAT/IRT allows for individual confidence intervals for individuals
- Can have more precision at specific cut points (people close to the passing grade for an exam can be measured more precisely than those far (above or below) the passing point.)
Psychological (non-psychometric) problems with CAT

- CAT items have difficulty level tailored to individual so that each person passes about 50% of the items.
- This increases the subjective feeling of failure and interacts with test anxiety.
- Anxious people quit after failing and try harder after success -- their pattern on CAT is to do progressively worse as test progresses (Gershon, 199x, in preparation).
Generalizations of IRT to 2 and 3 item parameters

- Item difficulty
- Item discrimination (roughly equivalent to correlation of item with total score)
- Guessing (a problem with multiple choice tests)
- 2 and 3 parameter models are harder to get consistent estimates and results do not necessarily have monotonic relationship with total score
3 parameter IRT
slope, location, guessing
Item Response Theory

- Can be seen as a generalization of classical test theory, for it is possible to estimate the correlations between items given assumptions about the distribution of individuals taking the test.
- Allows for expressing scores in terms of probability of passing rather than merely rank orders (or even standard scores). Thus, a 1 sigma difference between groups might be seen as more or less important when we know how this reflects chances of success on an item.
- Emphasizes non-linear nature of response scores.
Varieties of Validity

- Face
- Concurrent
- Predictive
- Construct
  - Convergent
  - Discriminant
Face (Faith Validity)

- Representative content
- Seeming relevance
Concurrent Validity

- Does a measure correlate with the criterion?
- Need to define the criterion.
- Assumes that what correlates now will have predictive value.
Predictive Validity

- Does a measure correlate with the criterion?
- Need to define the criterion.
- Requires waiting for time to pass.
Predictive and Concurrent Validity and Decision Making

Hit Rate = Valid Positive + False Negative

Selection Ratio = Valid Positive + False Positive

\[ \text{Phi} = \frac{\text{VP} - \text{HR} \times \text{SR}}{\sqrt{\text{HR} \times (1 - \text{HR}) \times \text{SR} \times (1 - \text{SR})}} \]
Validity as decision making

VP
FP
FN
VN
Validity as decision making
Validity as decision making
Decision Theory and Signal Detection

Probability VP

Sensitivity (correlation)

Probability FP
Construct Validity: Convergent, Discriminant, Incremental

Diagram:
- X1, X2, X3 → L1
- X4, X5, X6 → L2
- L1 → Y
- L2 → Y
- Y → Y1, Y2, Y3
Construct Validity

• Convergent
  – Do alternative measures of the same construct correlate with each other?

• Discriminant
  – Do measures of alternative constructs not correlate with each other

• Incremental
  – Does knowing something about a construct improve the predictability of other constructs more than what you already know?
Alternative Explanatory Variables

• A developmental psychologist has noticed that children with bigger feet tend to have greater vocabularies than children with smaller feet?

• This is an example of a simple correlational design. Can you think of a powerful alternative explanation?
Theory and Theory Testing IV: Correlational inference

Construct 1: Nutrition

Observation 1: Shoe size

Construct 2: Intellectual Ability

Observation 2: Vocabulary

$r_{mo}$
Theory and Theory Testing IV: Correlational inference

Construct 1: Nutrition
- Observation 1: Shoe size

Construct 2: Intellectual Ability
- Observation 2: Vocabulary

Construct X: Age

Relationships:
- $r_{mo}$
Alternative Explanatory Variables

• A Social Psychologist did an archival investigation of the relationship between rum drinking and the salaries of ministers. He has found that as the amount paid in salaries to ministers increases, that the price of Puerto Rican rum increases. He interprets this as an example of the law of supply and demand. What other variables should be included?
Theory and Theory Testing IV: Correlational inference

Construct 1: demand
Observation 1: Salaries

Construct 2: supply
Observation 2: Rum price

$r_{mo}$
Theory and Theory Testing IV: Correlational inference

Construct 1: Demand
  Observation 1: Salaries

Construct X: General inflation

Construct 2: Supply
  Observation 2: Rum price

$r_{mo}$
Alternative Explanatory Variables

Many developmental psychopathologists claim that harsh parenting causes psychopathology in adulthood. A recent study reports evidence in favor of this hypothesis: Depressed college students report that their parents were much harsher in the way they treated them than do non-depressed students.

Consider alternative explanations
A correlational study:
Depression and harsh parenting

Recollections of parenting style

Parenting style

Proband
A correlational study: Depression and harsh parenting

Diagram:
- Parenting style
- Proband
- Sib
A correlational study: Depression and harsh parenting
A correlational study: Depression and harsh parenting

Note that Genes and family environment are confounded
A correlational study: Depression and harsh parenting

Note that Genes and family environment are confounded
Typical Behavior Genetic Design

\[ r_g = 1, .5, 0 \]
\[ r_c = 1, 0 \]

A = additive genetic variance
C = Common family environment
E = Unique environment

\[ r_{g_{s1,s2}} = 1 \text{ for MZ, } .5 \text{ for DZ} \]
\[ r_{c_{s1,s2}} = 1 \text{ for together, } 0 \]
A correlational study: Depression and harsh parenting
A correlational study:
Depression and harsh parenting
Longitudinal environment/genetic study:

Genes

Parenting style

Proband

BioSib

Adopted Sib

Childhood

BioSib apart

Proband

BioSib

Adopted Sib

Now