PMAP 8131 Applied Research Methods II

Instrumentation

Matteo Zullo

Georgia State University & Georgia Institute of Technology

June 6, 2022

Outline

1. Instrumental variables
2. Latent variables
3. Reliability & validity

Introduction

- Instrumental variable
  - Proxy for unobserved feature at *same* semantic level of observed feature
- Latent variable
  - Proxy for unobserved feature at *deeper* semantic level of observed feature(s)

Table of Contents

1. Instrumental variables
2. Latent variables
3. Reliability & validity

Instrumental variables

Instrumental variables: Examples

- Social capital
  - Blood donations (Guiso et al., 2004)
- Intrinsic motivation
  - CEO time in insider meetings (Bandiera et al., 2020)
- Athletic participation
  - Height (Eide & Ronan, 2001)

Instrumental variables

Instrumental variables: Why?

- Feature is *omitted or unmeasured*
  - E.g., Census surveys do not "measure" social capital
- Self-declared measure is *biased*
  - E.g., CEOs surveys overstate intrinsic motivation
- Feature affects outcome through *spurious* pathways
  - E.g., Sports participation "badly" predicts test scores

Instrumental variables

Instrumental variables: How does it look like?

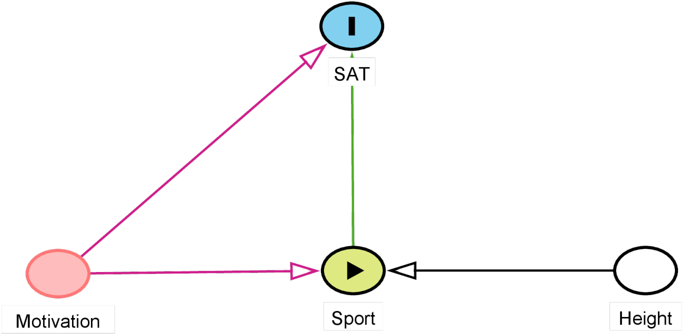

Figure: Effect of athletic participation on SAT test scores

Instrumental variables

Two-Stage Least Squares (2SLS)

- 1st stage: Model confounded regressor using instrument
- 2nd stage: Re-model outcome using first-stage proxy

Instrumental variables

Two-Stage Least Squares (2SLS)

- Latent model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \lambda_i + \epsilon_i$$

- Empirical model

$$Y_i = \beta_0 + \beta_{1LS} X_i + \epsilon_i$$

Instrumental variables

Two-Stage Least Squares (2SLS)

- Latent model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \lambda_i + \epsilon_i$$

- Empirical model

$$Y_i = \beta_0 + \beta_{1LS} X_i + \epsilon_i$$

  - Omitted variable bias: λ_i unobserved and not included
    - Because $\rho(X_i, \lambda_i) \neq 0$, β_{1LS} is biased and inconsistent
    - To prevent, instrument X_i with Z_i, where $\rho(Z_i, \lambda_i) = 0$

Instrumental variables

Two-Stage Least Squares (2SLS)

- 1S: Model "true values" of X_i as function of instrument

$$\mathbf{X_i} = \delta_0 + \delta_1 Z_i + \nu_i$$

  - Note: Z_i must be independent of λ_i and ν_i!

Instrumental variables

Two-Stage Least Squares (2SLS)

- 1S: Model "true values" of X_i as function of instrument

$$\mathbf{X_i} = \delta_0 + \delta_1 Z_i + \nu_i$$

  - Note: Z_i must be independent of λ_i and ν_i!

- 2S: Substitute $X_i = \hat{X}_i + \nu_i$ and rewrite

$$\begin{aligned} Y_i &= \beta_0 + \beta_{2SLS}(\hat{\mathbf{X}}_\mathbf{i} + \nu_i) + \epsilon_i \\ Y_i &= \beta_0 + \beta_{2SLS}\hat{\mathbf{X}}_\mathbf{i} + (\beta_{2SLS}\nu_i + \epsilon_i) \\ Y_i &= \beta_0 + \beta_{2SLS}\hat{\mathbf{X}}_\mathbf{i} + \epsilon_i^* \end{aligned}$$

Instrumental variables

2SLS example: Grades and athletics (Eide & Ronan, 2001)

- Latent model

$$Grade_i = \beta_0 + \beta_1 \mathbf{Sports_i} + \beta_2 Motivation_i + \epsilon_i$$

- 2SLS model

$$\begin{aligned} \mathbf{Sports_i} &= \delta_0 + \delta_1 Height_i + \nu_i & (1) \\ Grade_i &= \beta_0 + \beta_{2SLS}(\hat{\mathbf{Sports_i}} + \nu_i) + \epsilon_i & (2) \\ &= \beta_0 + \beta_{2SLS}\hat{\mathbf{Sports_i}} + (\beta_{2SLS}\nu_i + \epsilon_i) \\ &= \beta_0 + \beta_{2SLS}\hat{\mathbf{Sports_i}} + \epsilon_i^* \end{aligned}$$

Table of Contents

1 Instrumental variables

2 Latent variables

3 Reliability & validity

Latent variables

Latent variables: Examples

- Intelligence
- Motivation
- Agreeableness

Latent variables

Latent variables: Algorithms

- Principal Component Analysis (PCA)
- Factor analysis (FA)
- Item-Response Theory (IRT)

Latent Variables

Principal Component Analysis (PCA)

- Algorithm
  - Find **M linear combinations** of features X $(1, \ldots, K)$
    - Decorrelate any adjacent dimensions (Y_m, Y_{m+1})
    - Maximize variance of each dimension $Var(Y_m)$

$$Y_1 = \sum_{k=1}^{K} W_{1k} X_k$$

$$Y_m = \ldots$$

$$Y_M = \sum_{k=1}^{K} W_{Mk} X_k$$

Latent Variables

Principal Component Analysis (PCA)

- Selecting PCs: Eigenvalue rule
  - Retain all principal components having eigenvalue > 1
- Performing PCA
  - Standardize input features before performing PCA
  - Assume linear relationship between input features
  - Remove outliers in the input data

Latent Variables

Factor Analysis (FA)

- FA of O*NET skill codes (Zullo et al., 2022)

code	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
10	72.5	75.1	78.5	53.8	54.7	78.2	74.5	19.8	79.3	70.6
20	58.7	68.5	66.7	48.8	34.9	68.5	66.7	19.8	68.5	56.0
100	51.4	68.5	64.0	48.8	30.9	56.8	68.5	0.0	66.0	63.3
110	58.1	68.5	70.3	54.1	50.0	68.7	68.5	19.1	66.0	64.6

Figure: Skills codes for Census occupations

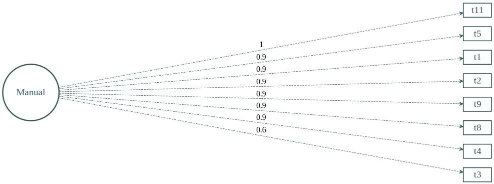

Figure: Factor loadings on “manual” factor

Latent Variables

Item Response Theory (IRT)

- Used to score examinees on a test
- Items are dichotomous $(0, 1)$ or polytomous $(0, 1, 2, \ldots)$
- Widely used in psychometrics (SAT, Minnesota, etc.)

ST012 **<u>How many</u> of these are there at your home?**

(Please select one response in each row.)

		None	*One*	*Two*	*Three or more*
ST012Q01TA	Televisions	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q02TA	Cars	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q03TA	Rooms with a bath or shower	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	$\square_1$	$\square_2$	$\square_3$	$\square_4$

Figure: PISA 2018 questionnaire (Avvisati, 2020)

Latent Variables

Item Response Theory (IRT)

- 3PL (Three parameter Logistic) model

$$P(Y_{ik} = 1|\theta_i, a_k, b_k) = c(1-c)\frac{e^{Da_k(\theta_i - b_k)}}{1 + e^{Da_k(\theta_i - b_k)}}$$

- where:
  - θ_s = <u>ability</u> parameter of individual i
  - a_k = <u>discrimination</u> parameter of item k
  - b_k = <u>difficulty</u> parameter of item k
  - D = scaling factor

Latent Variables

- IRT vs Classical Test Theory: No "ground truth" score
    - Item-dependent scores, sample-dependent item statistics

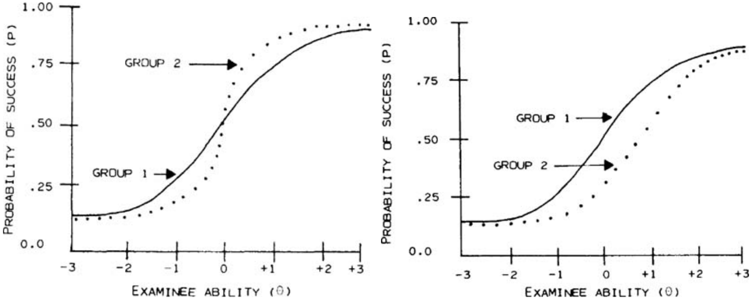

Figure: IR-curves with different discrimination (L) and ability (R) (Osterlind, 1983)

Latent Variables

Putting it all together: PISA's ESCS index

- *ESCS*: Index of Economic, Social and Cultural Status
  - Weighted average of latent dimensions

$$ESCS = \lambda_1 HISEI + \lambda_2 PARED + \lambda_3 HOMEPOS$$

ST012 **<u>How many</u> of these are there at your home?**

(Please select one response in each row.)

		None	*One*	*Two*	*Three or more*
ST012Q01TA	Televisions	□1	□2	□3	□4
ST012Q02TA	Cars	□1	□2	□3	□4
ST012Q03TA	Rooms with a bath or shower	□1	□2	□3	□4
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	□1	□2	□3	□4

Figure: PISA 2018 questionnaire (Avvisati, 2020)

Latent Variables

- *ESCS* index: Latent dimensions scores (IRT)
  - Highest Parents' Occupation (*HISEI*): Continuous
    - Numeric codes proxying for occupational status
  - Highest Parental Education (*PARED*): Categorical
    - None, primary, lower secondary, upper secondary, non-tertiary post-secondary, vocational tertiary, tertiary
  - Home Possessions (*HOMEPOS*): Numeric
    - Summary index from background items
- ESCS index: Weights (PCA)
  - $(\lambda_1, \lambda_2, \lambda_3)$ are loadings on the first principal component from PCA of input features *HISEI*, *PARED*, *HOMEPOS*

Table of Contents

1 Instrumental variables

2 Latent variables

3 Reliability & validity

Latent variables

Reliability

- **Test-retest reliability**: Correlation across time
  - Pearson's correlation (ρ)
- **Internal consistency**: Correlation across items
  - Cronbach's alpha (α)

Latent variables

Validity

- **Face validity**
  - Items are valid at face value
- **Content validity**
  - Items comprehensively cover latent construct
- **Convergent validity**
  - Instrument positive correlation with similar instruments
- **Discriminant validity**
  - Negative correlation with dissimilar instruments

Latent variables

Validity

- **Face validity**
  - Self-reports
- **Content validity**
  - Items comprehensively cover latent construct
- **Convergent validity**
  - Instrument positive correlation with similar instruments
- **Discriminant validity**
  - Negative correlation with dissimilar instruments

Latent variables

- **Face validity**
  - Beware self-reports: Eliciting biased responses!
    - E.g., "Are you satisfied with your relationships?" worse proxy than "How many closed friends do you have?"

Latent Variables

Reliability

- Cronbach's alpha

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_i^2}{\sigma^2}\right)$$

- where:
  - N = number of items
  - σ_i^2 = variance of scores on item i
  - σ^2 = variance of scores on all times

Latent Variables

Cronbach's example: Rosenberg's Self-Esteem Scale (SES)

- Measuring positive and negative self-beliefs
- Using Likert scale (0-3)
  - Strongly disagree (*SD*), Disagree (*D*), Agree (*A*), Strongly Agree (*SA*)
- Achieving often very high Cronbach's alpha (0.75+)

Latent Variables

#	R*	Question	SD	D	A	SA
I1	0	On the whole, I am satisfied with myself	0	1	2	3
I2	1	At times, I think I am no good at all	3	2	1	0
I3	0	I feel that I have a number of good qualities	0	1	2	3
I4	0	I am able to do things as well as most other people	0	1	2	3
I5	1	I feel I do not have much to be proud of	3	2	1	0
I6	1	I certainly feel useless at times	3	2	1	0
I7	0	I feel I'm a person of worth, at least equally to others	0	1	2	3
I8	1	I wish I could have more respect for myself	3	2	1	0
I9	1	All in all, I am inclined to think that I am a failure	3	2	1	0
I10	0	I take a positive attitude towards myself	0	1	2	3

Table: Scoring system for Rosenberg's SES; *(R = 1 if reverse-coded item)

Latent Variables

- Sample responses

Subject	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Score
1	1	1	2	2	1	1	1	2	1	1	13
2	2	2	3	2	2	1	2	2	2	2	20
3	0	0	1	0	1	1	0	1	0	0	4
4	0	0	1	1	0	0	1	1	0	0	4
5	2	3	2	3	3	2	3	2	2	1	23
Var_i	1	1.7	0.7	1.3	1.3	0.5	1.3	0.3	1	0.7	77.7

Latent Variables

- Sample responses

Subject	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Score
1	1	1	2	2	1	1	1	2	1	1	13
2	2	2	3	2	2	1	2	2	2	2	20
3	0	0	1	0	1	1	0	1	0	0	4
4	0	0	1	1	0	0	1	1	0	0	4
5	2	3	2	3	3	2	3	2	2	1	23
Var_i	1	1.7	0.7	1.3	1.3	0.5	1.3	0.3	1	0.7	77.7

- Cronbach's alpha calculations
  - $N = 10$, $\sum_{i=1}^{N} \sigma_i^2 = 1 + \ldots + 0.7 = 9.8$, $\sigma^2 = 77.7$
  - Hence, Cronbach's alpha: $\alpha = \frac{10}{10-1}(1 - \frac{9.8}{77.7}) = 0.971$