An Introduction to Latent Variable Models

Karen Bandeen-Roche
ABACUS Seminar Series

November 28, 2007

LATENT VARIABLES: TRUTH, LIES, AND EVERYTHING BETWEEN

Karen Bandeen-Roche
Department of Biostatistics
Johns Hopkins University

ABACUS Seminar Series
November 28, 2007

Objectives
For you to leave here knowing…

- What is a latent variable?
- What are some common latent variable models?
- What is the role of assumptions in latent variable models?
- Why should I consider using—or decide against using—latent variable models?

ALATENT@

1. Present or potential but not evident or active: latent talent.
2. Pathology. In a dormant or hidden stage: a latent infection.
3. Biology. Undeveloped but capable of normal growth under the proper conditions:a latent bud.
4. Psychology. Present and accessible in the unconscious mind but not consciously expressed.

<u>The American Heritage7 Dictionary of the English Language</u>, Fourth Edition, 2000

Aexisting in hidden or dormant form but usually capable of being brought to light@

<u>Merriam-Webster's Dictionary of Law</u>, 1996

ALATENT@

A... concepts in their purest form....>unobserved= or >unmeasured=... >hypothetical@

Bollen KA, <u>Structural Equations with Latent Variables</u> p. 11, 1989

A... in principle or practice, cannot be observed@

Bartholomew DJ, <u>The Statistical Approach to Social Measurement</u>, p. 12, 1996

AUnderlying: not directly measurable. Existing in hidden form but usually capable of being measured indirectly by observables@

Bandeen-Roche K, Synthesis, 2006

ALATENT VARIABLES@

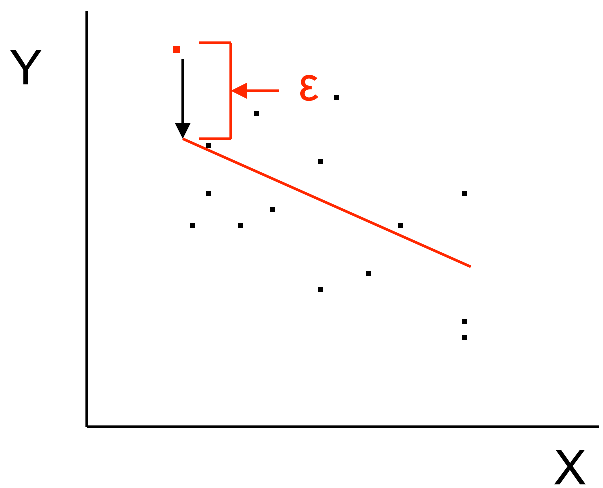
Ordinary linear regression model:

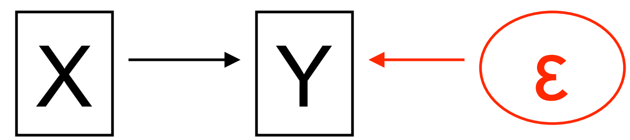
Y_i = outcome (measured)
$\underline{X}_i$ = covariate vector (measured)
ε_i = residual (unobserved)

$$Y_i = \underline{X}_i^T \underline{\beta} + \varepsilon_i$$

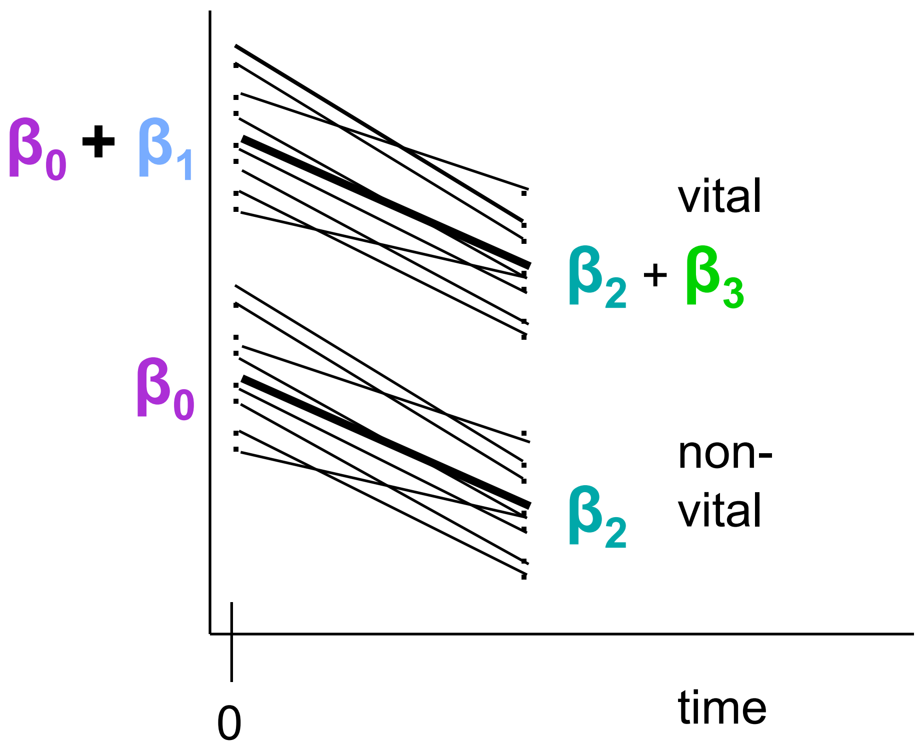
Ordinary Linear Regression

Residual as Latent Variable



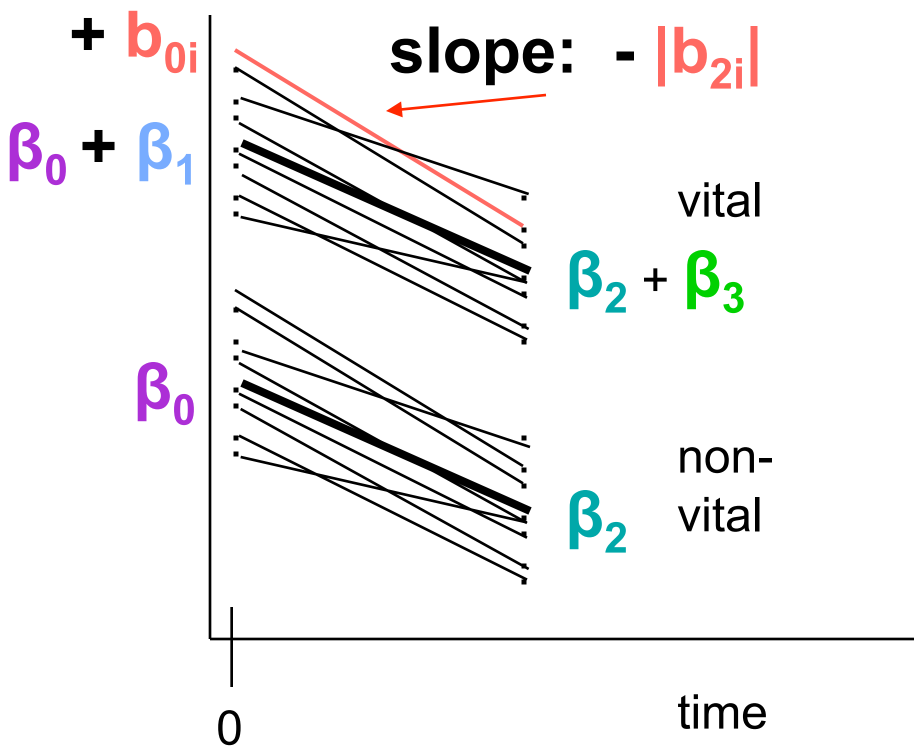


Mixed effect / Multi-level models
Random effects as Latent Variables



$$Y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 t_{ij} + \beta_3 x_i \cdot t_{ij} + e_{ij}$$

Mixed effect / Multi-level models
Random effects as Latent Variables



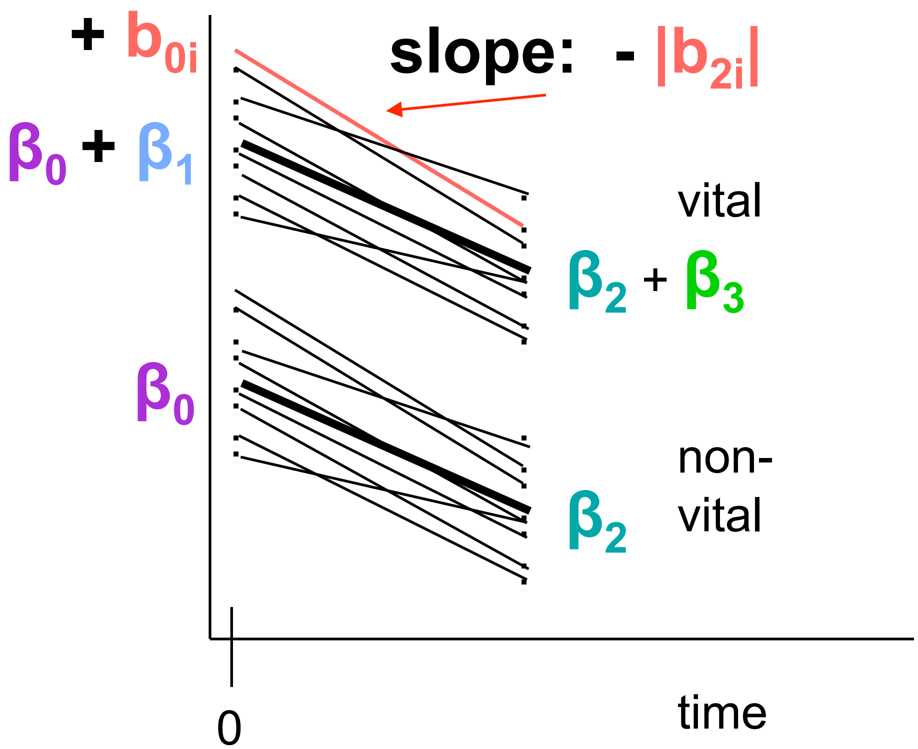
- b_{0i} = random intercept

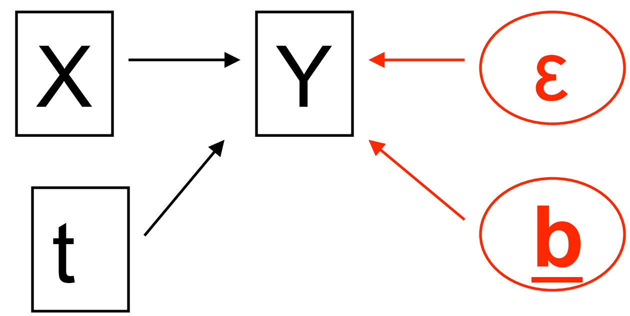
  b_{2i} = random slope

  (could define more)
- Population heterogeneity captured by spread in intercepts, slopes

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 x_i + \beta_2 t_{ij} + b_{2i} t_{ij} + \beta_3 x_i \cdot t_{ij} + e_{ij}$$

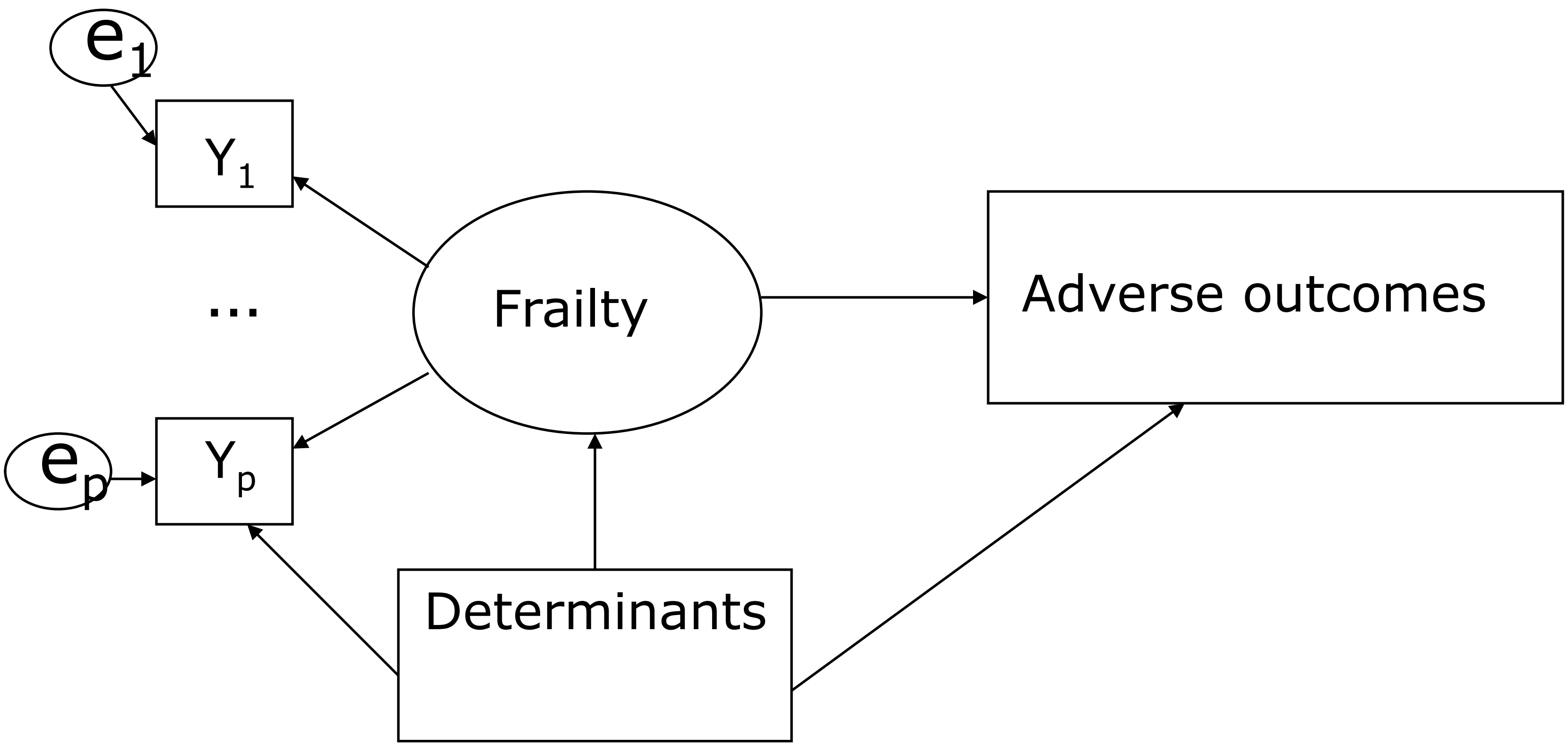
Mixed effect / Multi-level models
Random effects as Latent Variables





$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 x_i + \beta_2 t_{ij} + b_{2i} t_{ij} + \beta_3 x_i \cdot t_{ij} + e_{ij}$$

Frailty
Latent Variable Illustration



ALATENT VARIABLES@

Linear structural equations model with latent variables (LISREL):

Y_{ij} = outcome (j th measurement perAperson@i)
$\underline{x}_{ij}$ = covariate vector (corresponds to j th measurement, person i)
$\underline{\lambda}_j$ = Aloading@(relates LV to j th measurement)
$\underline{\eta}_i$ = latent variable=random coefficient vector, person i
ε_{ij} = observed response residual
$\underline{\varsigma}_i$ = latent response residual vector (specified distribution)

$\mathbf{Y_{ij} = \underline{\lambda}_{ij}^T \underline{\eta}_i + \varepsilon_{ij}}$ (measurement model—here, factor analysis)

$\underline{\boldsymbol{\eta}}_i = \mathbf{B}\underline{\boldsymbol{\eta}}_i + \boldsymbol{\Gamma}\underline{\mathbf{x}}_i + \underline{\varsigma}_i$ (structural model: linear regression
marginal model: $[Y|x] = \square[Y|\eta,x][\eta|x]$)

> <u>My sense</u>: It=s the unknown $\underline{\lambda}_j$ that distinguishes above as a Alatent variable model@in most minds

Why do people use latent variable models?

- The complexity of my problem demands it
- NIH wants me to be sophisticated
- Reveal underlying truth (e.g. “discover” latent types)
- Operationalize and test theory
- Sensitivity analyses
- Acknowledge, study issues with measurement; correct attenuation; etc.

Well-used latent variable models

Latent variable scale	Observed variable scale	
	Continuous	Discrete
Continuous	Factor analysis LISREL	Discrete FA IRT (item response)
Discrete	Latent profile Growth mixture	Latent class analysis, regression

General software: MPlus, Latent Gold, WinBugs (Bayesian), NLMIXED (SAS)

WELL USED LATENT VARIABLE MODELS
FACTOR ANALYSIS / SEM

Linear structural equations model with latent variables (LISREL):

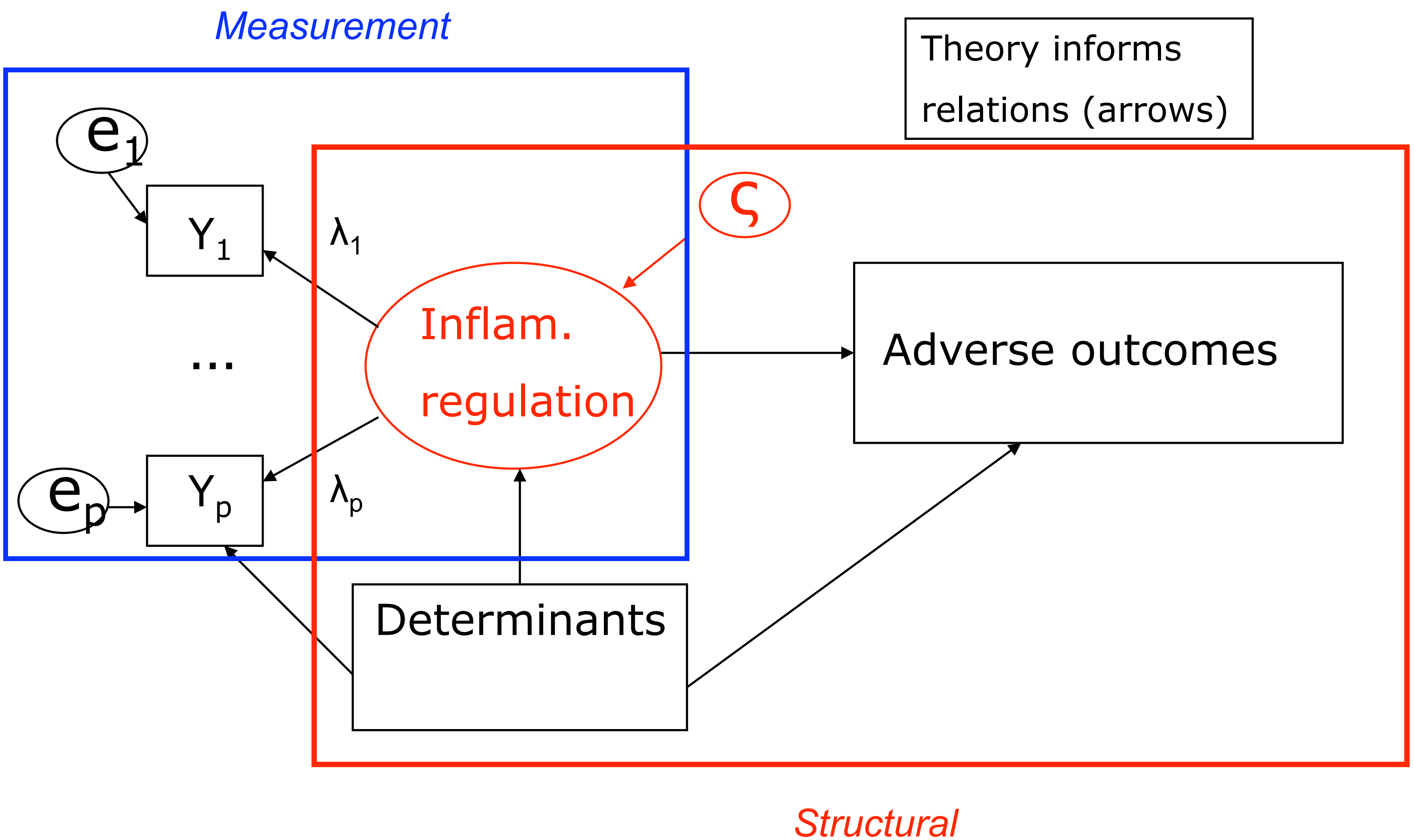
Y_{ij} = outcome (j th measurement per "person" i)
$\underline{x}_{ij}$ = covariate vector (corresponds to j th measurement, person i)
$\underline{\lambda}_j$ = "loading" (corresponds to j th measurement)
$\underline{\eta}_i$ = latent variable = random coefficient vector, person i
ε_{ij} = observed response residual
$\underline{\varsigma}_i$ = latent response residual vector (specified distribution)

$\mathbf{Y_{ij} = \underline{\lambda}_j^T \underline{\eta}_i + \varepsilon_{ij}}$ (measurement model—here, factor analysis)

$\underline{\boldsymbol{\eta}}_i = \mathbf{B}\underline{\boldsymbol{\eta}}_i + \boldsymbol{\Gamma}\underline{\mathbf{x}}_i + \underline{\varsigma}_i$ (structural model: linear regression
marginal model: $[Y|x] = \int [Y|\eta,x][\eta|x]$)

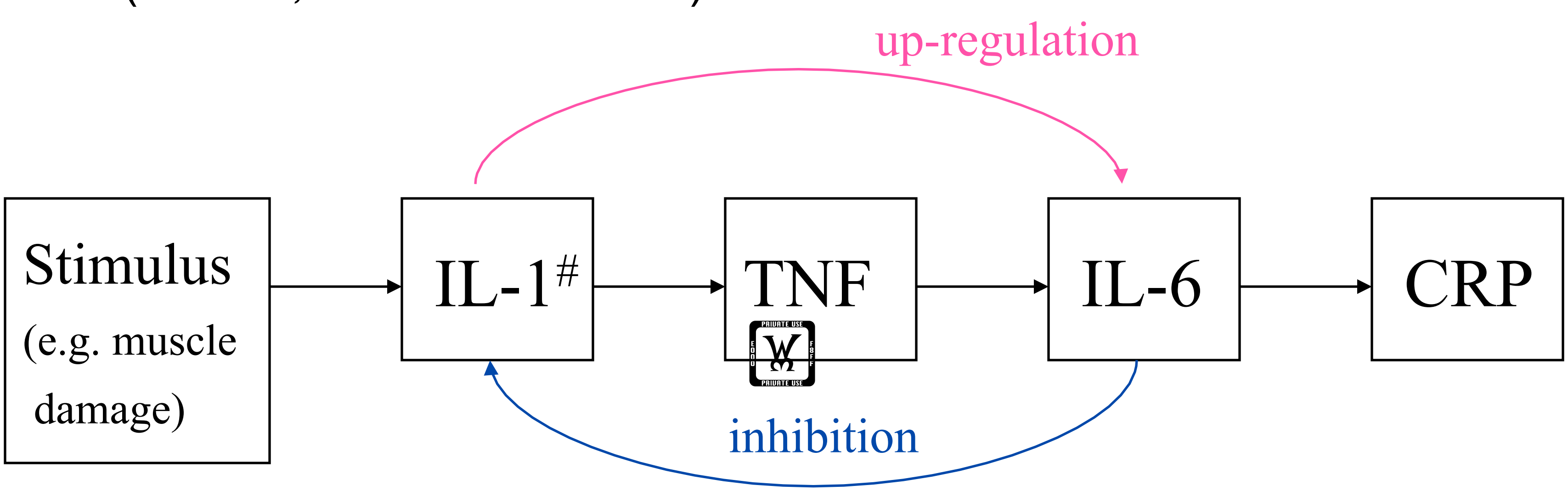
Tailored software: AMOS, LISREL, CALIS (SAS)

Frailty
Latent Variable Illustration



Example: Theory Infusion

- Inflammation: central in cellular repair
- Hypothesis: dysregulation=key in accel. aging
  - Muscle wasting (*Ferrucci et al., JAGS 50:1947-54; Cappola et al, J Clin Endocrinol Metab 88:2019-25*)
  - Receptor inhibition: erythropoetin production / anemia (*Ershler, JAGS 51:S18-21*)

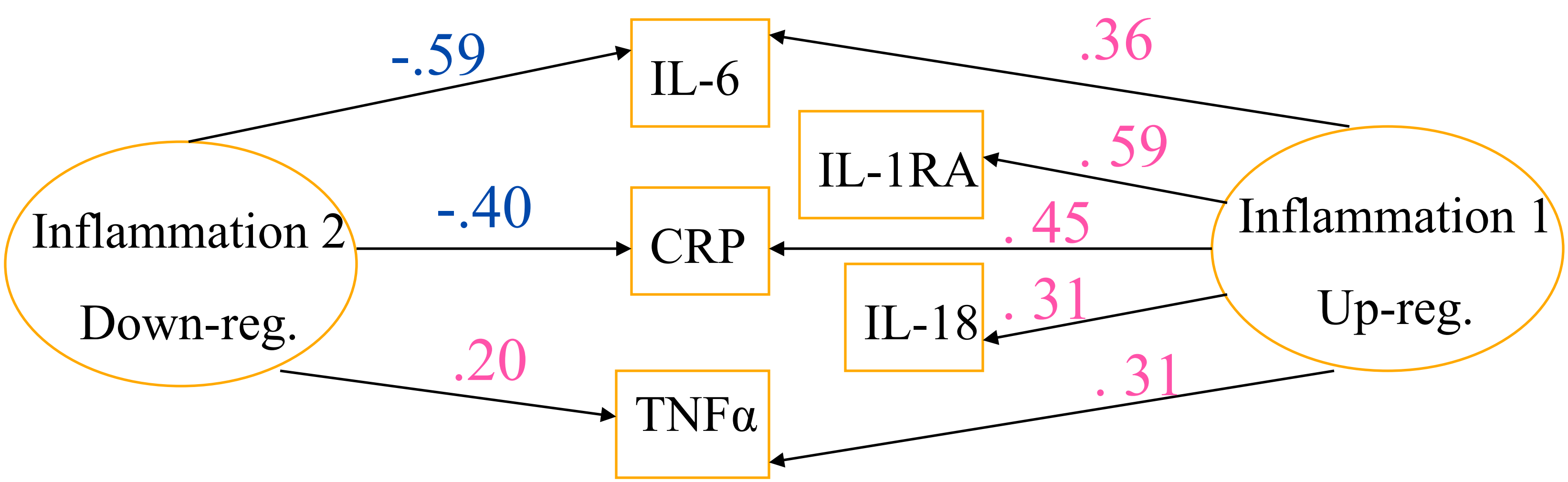


Difficult to measure. IL-1RA = proxy

Theory infusion

InCHIANTI data (*Ferrucci et al., JAGS, 48:1618-25*)

- LV method: factor analysis model
  - two independent underlying variables
  - down-regulation IL-1RA path=0
  - conditional independence

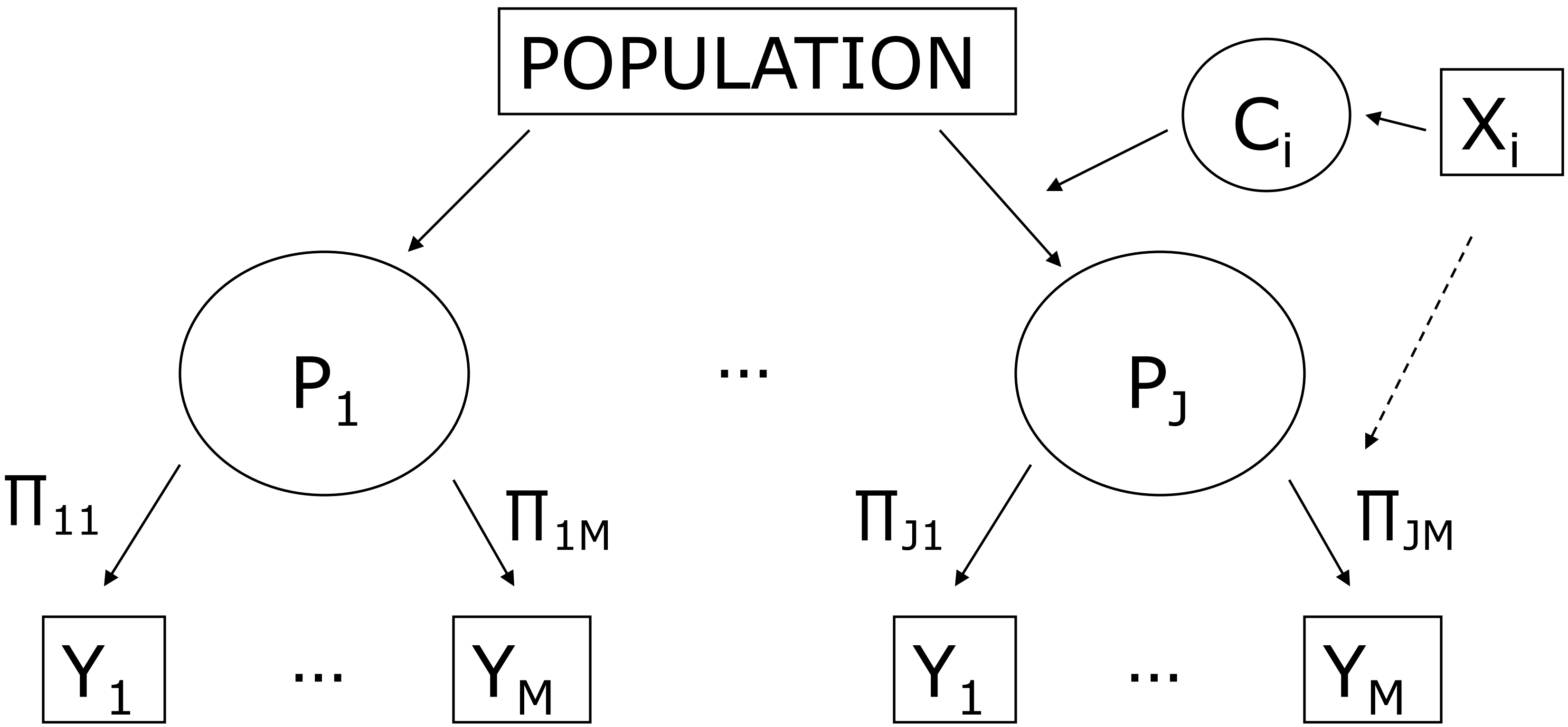


ANOTHER WELL-USED LATENT VARIABLE MODEL

Motivation: Self-reported Visual functioning

- Questionnaires have proliferated
  - This talk: Activities of Daily Vision[5] (ADV)
  - "Far vision" subscale: How much difficulty with *reading signs* (night, day); *seeing steps* (day, dim light); *watching TV* = Y1,...,Y5
- Question of interest: What aspects of vision determine "far vision" function
- One point of view on such "function": Latent subpopulations

Analysis of underlying subpopulations
Latent class analysis / regression



19-Goodman, 1974; 27-McCutcheon, 1987

Analysis of underlying subpopulations Method: Latent class analysis/ regression

- Seeks homogeneous subpopulations
  - **Assumption:** *reporting heterogeneity unrelated to measured or unmeasured characteristics*
  - *conditional independence, non differential measurement by covariates of responses within latent groups :* ***partially determine features***
- Features that characterize latent groups
  - Prevalence in overall population
  - Proportion reporting each symptom
  - Number of them

Latent Class Regression (LCR) Model: Technical Detail

- **Model**:

$$f_{Y|x}(y|x) = \sum_{j=1}^{J} P_j(x,\beta) \boxed{\prod_{m=1}^{M}} \pi_{mj}^{y_m}(1-\pi_{mj})^{1-y_m}$$

no x

- **Structural model assumption** : $[U_i|x_i] = Pr\{U_i=j\,|x_i\} = P_j(x_i,\beta)$
  - $RPR_j = Pr\{U_i = j|x_i\}/Pr\{U_i = J|x_i\}$; $j=1,\dots,J$
  - Latent polytomous logistic regression

- **Measurement assumptions** : $[Y_i|U_i]$
  - conditional independence
  - nondifferential measurement
    - ***reporting heterogeneity unrelated to measured, unmeasured characteristics***

- **Fitting**: Max. likelihood (e.g. *Muthén & Muthén 1998, MPlus*); Bayes

- ***Prediction:*** *Posterior* latent outcome info: $Pr\{U_i=j|Y_i,x_i;\theta=(\pi,\beta)\}$

LCR:
Self-reported Visual functioning

- Study: Salisbury Eye Evaluation (SEE; West et al. 1997[6])
  - Representative of community-dwelling elders
  - n=2520; 1/4 African American
  - This talk: 1643 night drivers

- Analyses control for potential confounders:
  - **Demographic**: age (years), sex **(1**{female**})**, race **(1{**nonwhite**}),** education (years)
  - **Cognition**: Mini-Mental State Exam score (MMSE; 30-0 points)
  - **Depression**: GHQ subscale score (0-6)
  - **Disease burden**: # reported comorbidities

Aspects of vision

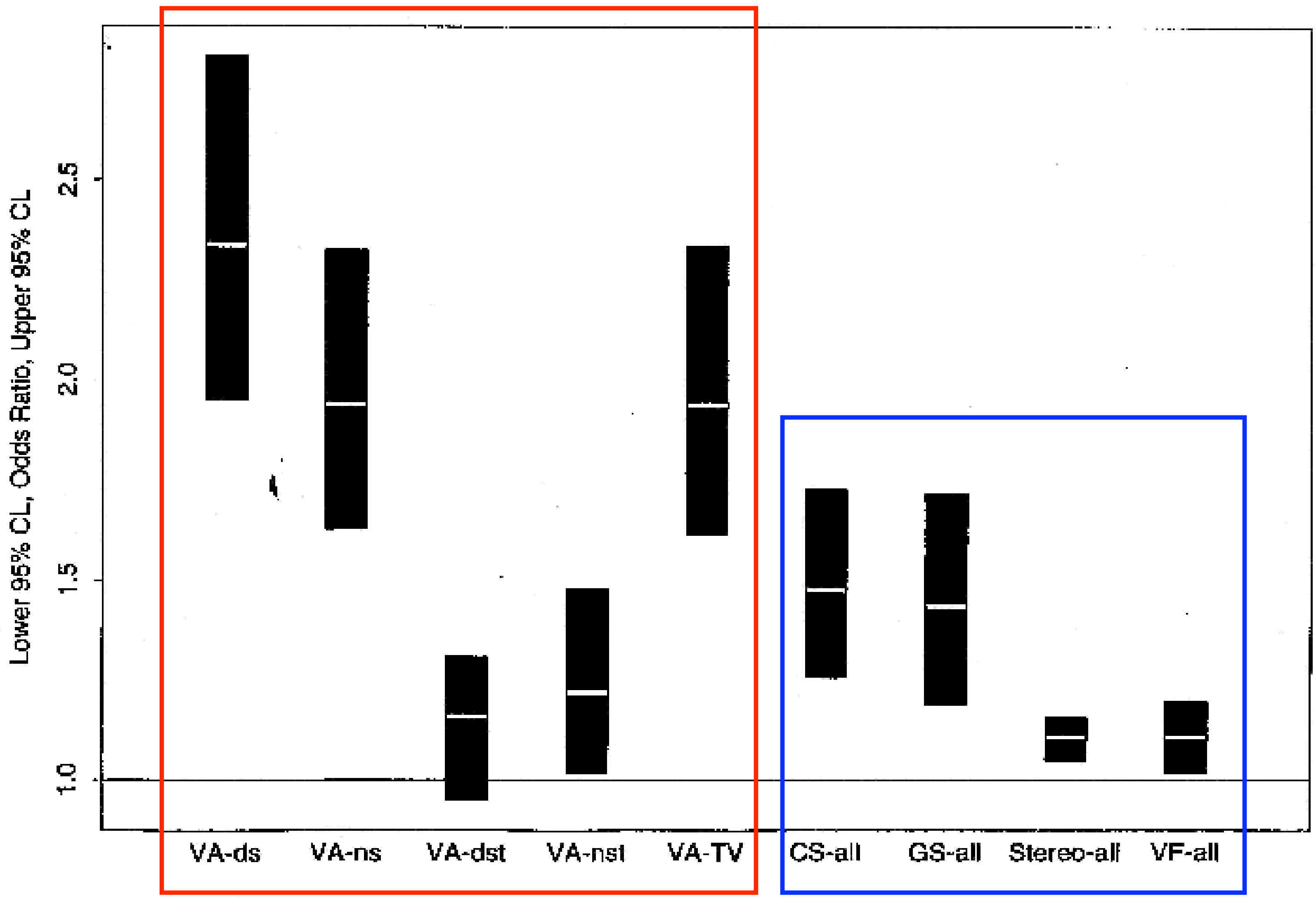
- **Visual acuity**: .3 logMAR (about 2 lines)
- **Contrast sensitivity**: 6 letters
- **Glare sensitivity**: 6 letters
- **Stereoacuity**: .3 log arc seconds
- **Visual field**: root-2 central points missed
  - Latter two: span approximately .60 IQR

ANALYSIS: SUMMARY SCORES
Multiple Regression of ADVS Far Vision Scores on Vision Variables

Vision Variable	Comparison[1]	OR	95% C.I.
Visual Acuity(.3)	Best vs Worst	**2.74**	(2.04, 3.68)
	Mid vs Worst	**1.72**	(1.29, 2.28)
Contrast Sens. (6)	Best vs Worst	**1.69**	(1.23, 2.32)
	Mid vs Worst	**1.46**	(1.06, 2.01)
Glare Sens. (6)	Best vs Worst	1.39	(0.97, 2.00)
	Mid vs Worst	1.07	(0.73, 1.56)
Stereoacuity (.3)	Best vs Worst	**1.25**	(1.13, 1.39)
	Mid vs Worst	**1.23**	(1.10, 1.37)
Visual Field (1.4)	Best vs Worst	1.14	(0.98, 1.33)
	Mid vs Worst	1.03	(0.88, 1.21)

[1] Best = 94-100; Mid = 72-93.99; Worst = < 72

Multiple Regression of ADVS Items on Vision Variables



Odds Ratio for association between items: 7.69

EXAMPLE
Latent Class Regression Measurement Model, ADVS Far Vision

TASK	REPORTING PROBABILITIES (π)				
	CLASS 1 ANONE@	CLASS 2 ANT SIGN@	CLASS 3 ASIGNS@	CLASS 4 ASTEPS@	CLASS 5 ASEVERE@
Signs--Night	.006	1.00	.949	.709	.991
Signs--Day	.005	.055	.955	0.00	.976
Steps--Day	.002	.006	.152	.625	.953
Steps--Dim	.019	.087	.441	.909	1.00
Watch TV	.010	.045	.241	.162	.613
PREVALENCES (mean η)	*.571*	*.221*	*.106*	*.062*	*.041*

Fit statistic (LR chi-square): 7.19 on 3 df

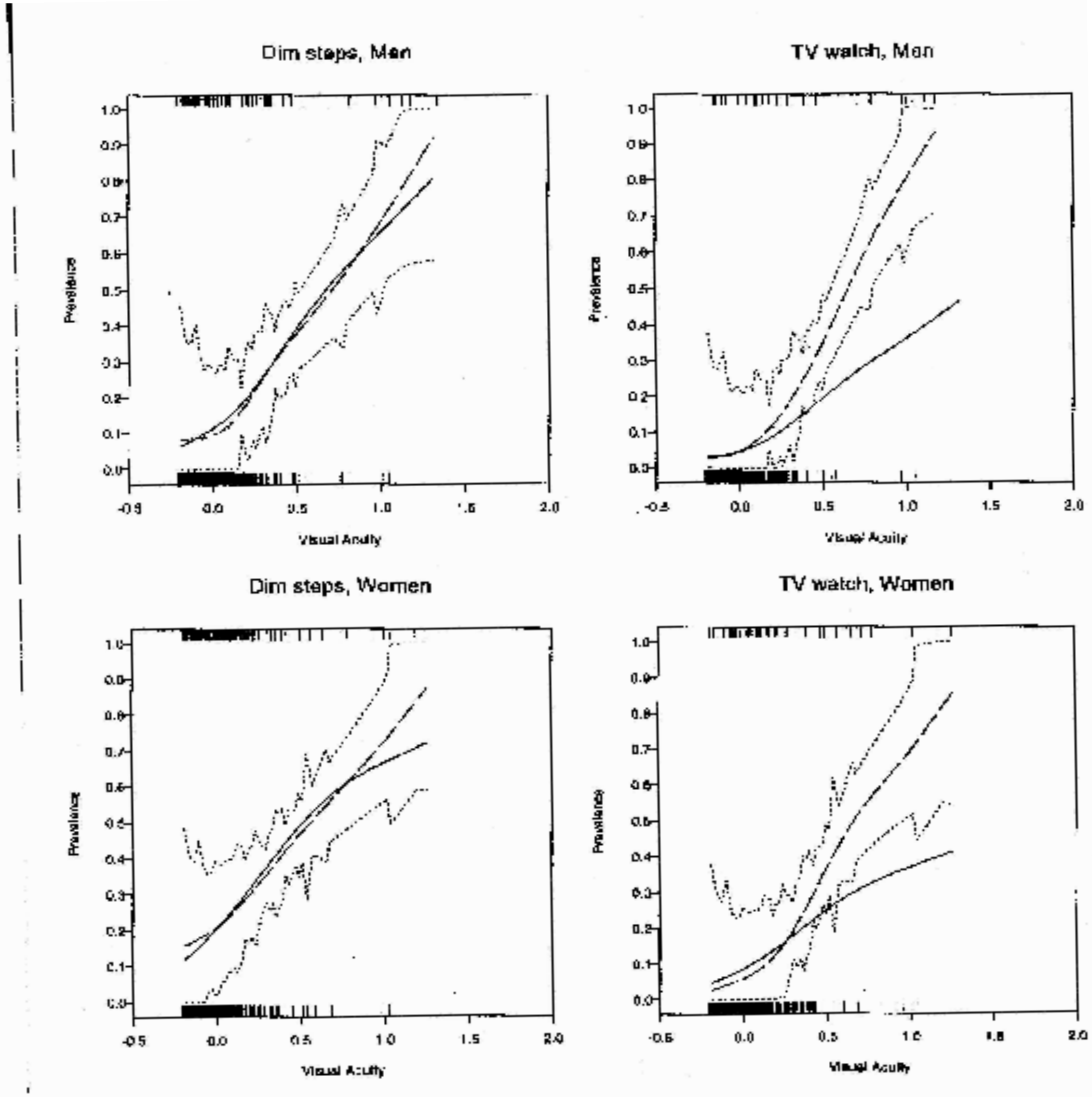
-2 log likelihoods: 2 classes = -6045.94 5 classes = -5865.64
3 classes = -5920.47 Saturated: -5858.45
4 classes = -5916.05

Latent Class Regression of ADVS Far Vision on Vision Variables[1]

Variable[3]	Comparison	OR	95% C.I.
Visual acuity (.3)	Severe vs None	**2.36**	(1.32,4.24)
	Steps vs None	1.35	(0.59,3.09)
	Signs vs None	**3.39**	(2.13,5.39)
	Nt-sign[2] vs None	1.43	(0.98,2.09)
Contrast sens. (6)	Severe vs None	**0.51**	(0.29,0.91)
	Steps vs None	**0.56**	(0.32,0.97)
	Signs vs None	0.77	(0.51,1.17)
	Nt-sign vs None	0.72	(0.51,1.02)
Glare sens. (6)	Severe vs None	1.89	(0.91,3.92)
	Steps vs None	**1.89**	(1.02,3.48)
	Signs vs None	**2.18**	(1.35,3.53)
	Nt-sign vs None	1.31	(0.93,1.84)
Sex (F v M)	Severe vs None	**4.22**	(1.91,9.33)
	Steps vs None	**3.03**	(1.71,5.37)
	Signs vs None	**3.84**	(2.09,7.05)
	Nt-sign vs None	**1.82**	(1.28,2.61)

MODEL CHECKING IS POSSIBLE!

Observed (solid) and Predicted (dash) Item Prevalence vs Acuity Plots



Summary

What Has Been Learned?

! **Analysis of summary scores**

- C Multiple aspects of vision independently, substantially associated with reported far vision functioning.
- C Age not associated with self-report, after accounting for vision (data not shown)

Summary

What Has Been Learned?

! **Analysis of Far Vision Items**

C Visual acuity association differentiated among tasks

Possible: missed dimension of functioning?

differential measurement?

C Between-item associations very strong

Summary
What Has Been Learned?

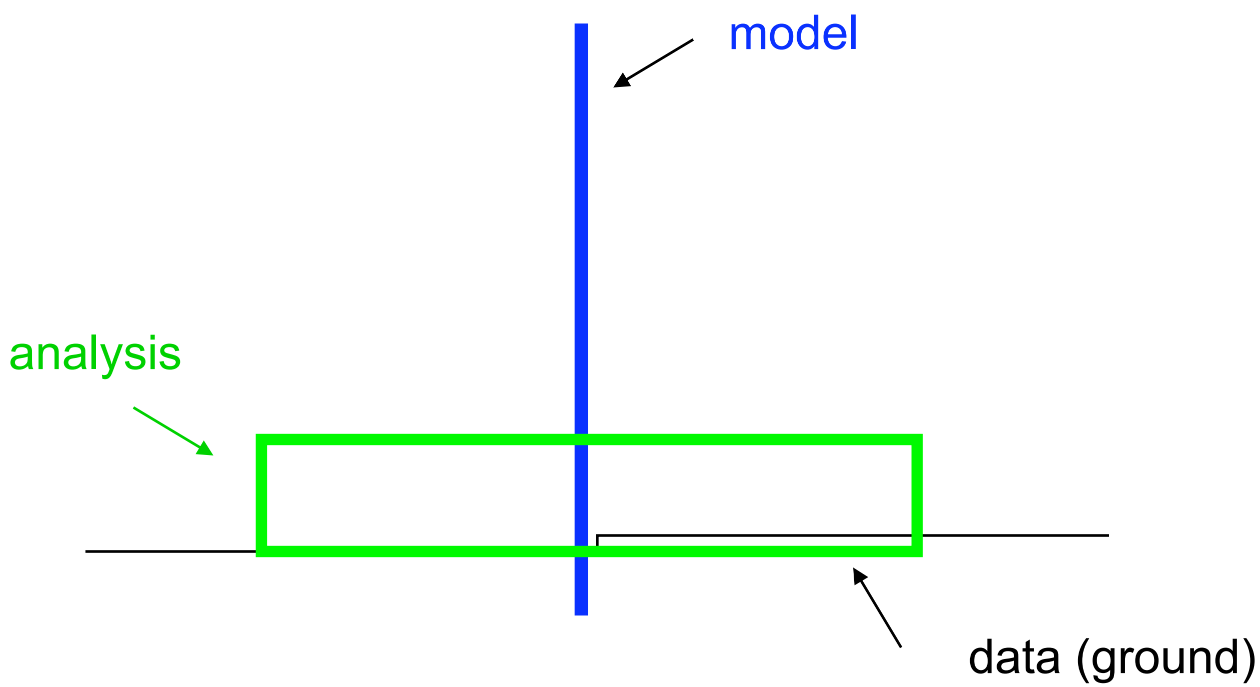
- **Summarize and analyze**
  - Distinct non-hierarchical difficulty patterns
    Distance acuity versus other far vision aspects
  - Specificity in visual associations
    Visual acuity: distance acuity tasks
    Contrast sensitivity: distance contrast tasks
    Glare sensitivity: global
    Stereoacuity, Visual field: primarily severe difficulty
  - Very general gender specificity in reporting
    Not driven by isolated items
  - TV: a rogue item?
    Possible masking (gender), inflation (acuity)?

One last issue
Identifiability

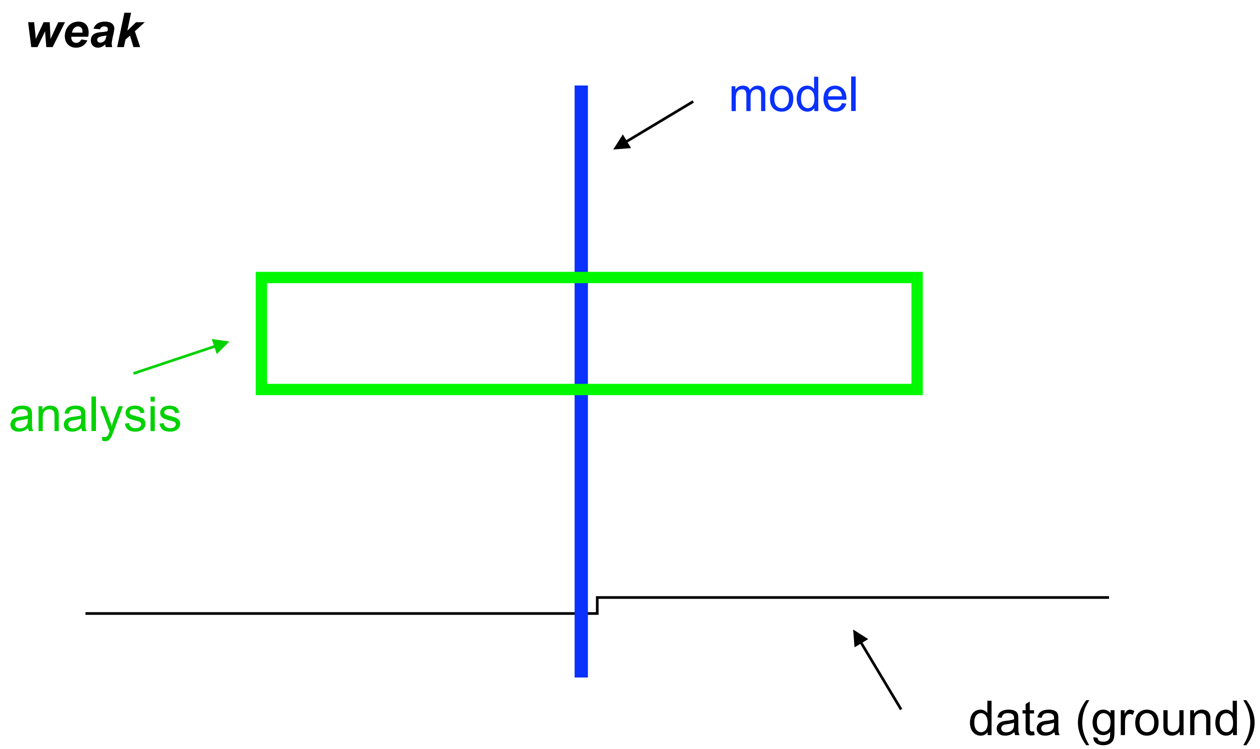
- Models can be too big / complex
- A model is non-identifiable if distinct parameterizations lead to identical data distributions
  - i.e. analysis not grounded in data
- Weak identifiability is common too:
  - Analysis only indirectly grounded in data (via the model)

Identifiability

strong

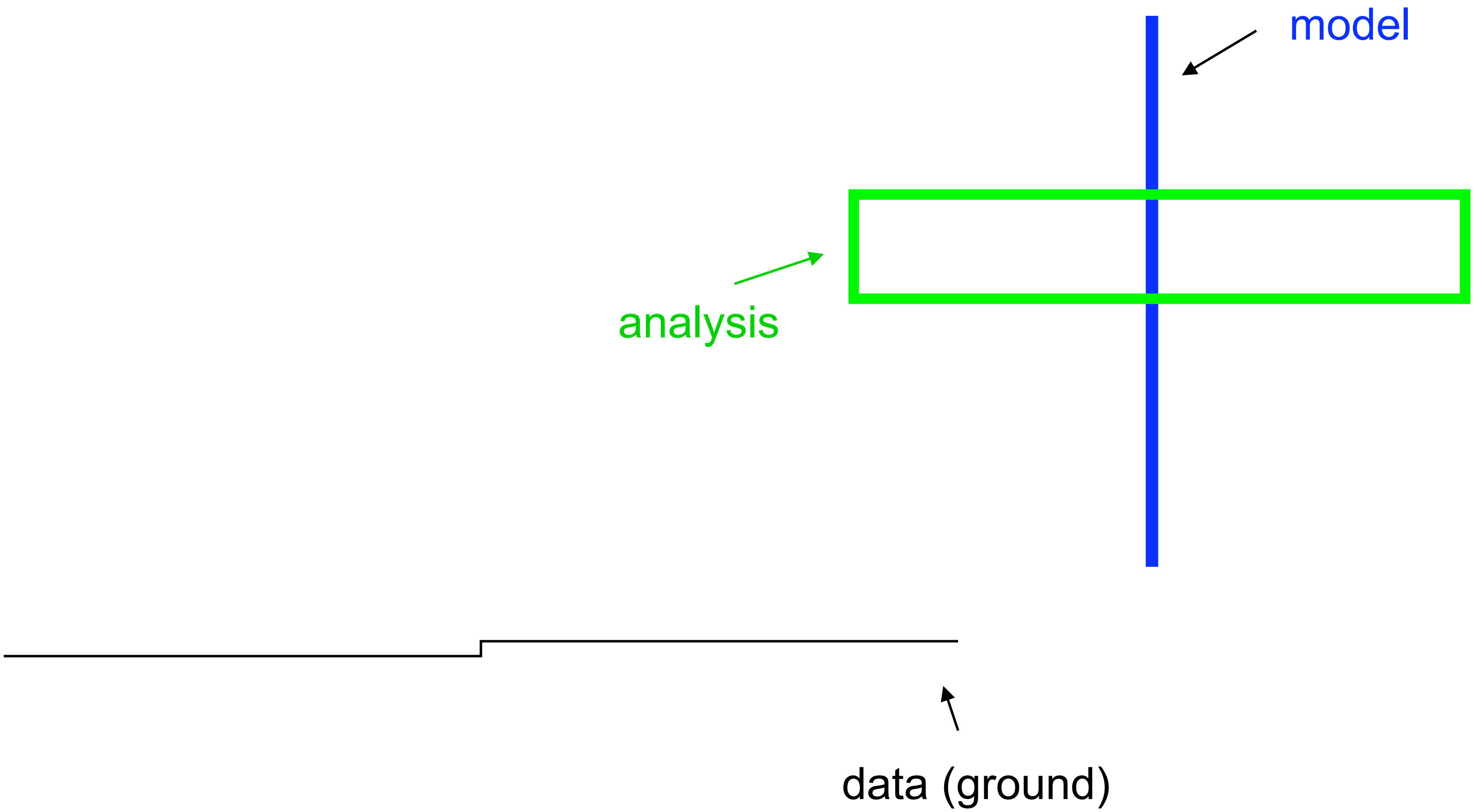


Identifiability



Identifiability

non



Objectives
For you to leave here knowing…

- What is a latent variable?
- What are some common latent variable models?
- What is the role of assumptions in latent variable models?
- Why should I consider using—or decide against using—latent variable models?

DISCUSSION
The Debate over Latent Variable Models

- **In favor**: they
  - acknowledge **measurement problems:** errors, differential reporting
  - **summarize** multiple measures **parsimoniously**
  - operationalize **theory**
  - describe population **heterogeneity**

- **Against**: their
  - **modeling assumptions** may determine scientific conclusions

  - **interpretation** may be ambiguous
    - Nature of latent variables (*existence*)?
    - Unique (***identifiability***)?
    - Comparable fit of very different models (*estimability*)?
    - Seeing is believing (*can the model be checked*)?