FROM LATENT VARIABLES TO CAUSAL INFERENCE:
UNDERSTANDING STRUCTURAL EQUATION MODELS

KENNETH A. BOLLEN

UNIVERSITY OF NORTH CAROLINA

AT CHAPEL HILL

USA

UNDERSTANDING STRUCTURAL EQUATION MODELS (SEMs)

Section 1: Motivation & Intuition

Section 2: SEM Core Concepts

Section 3: Model Estimation, Fit, & Coefficients

Section 4: Real-World SEM Examples

Section 5: Advanced SEM Capabilities

Section 6: Summary & Getting Started

WHAT ARE SEMs?

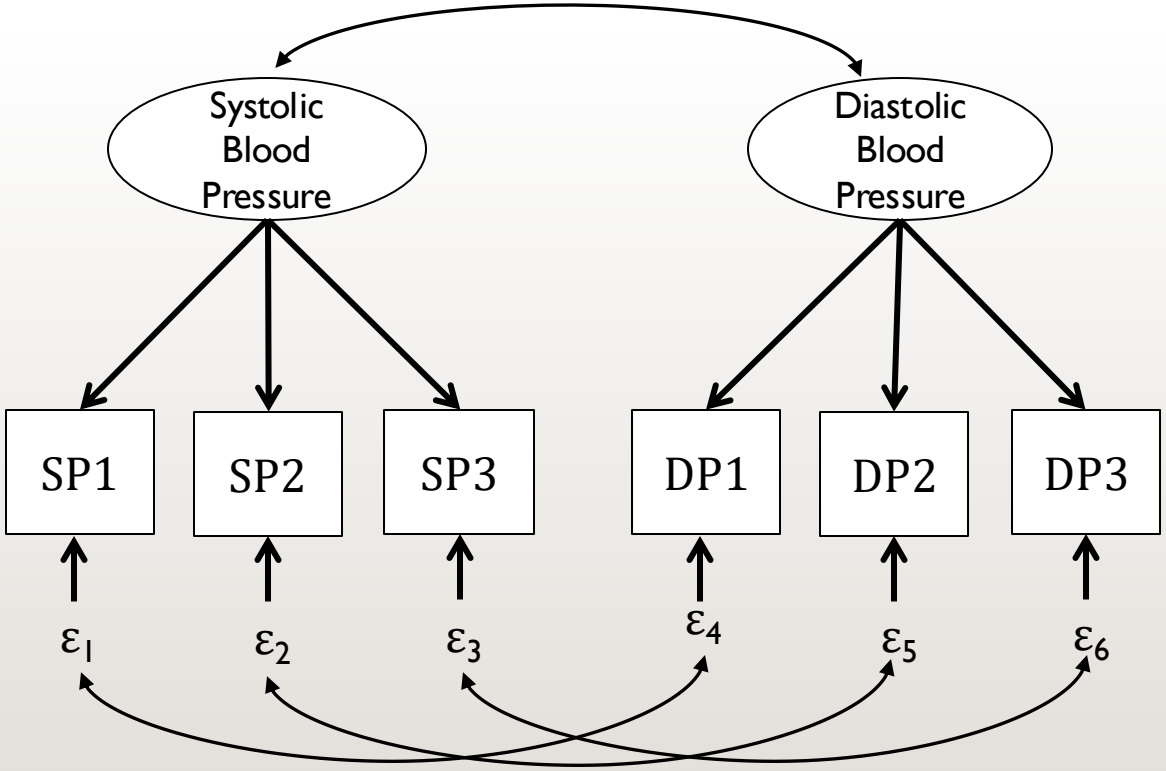
- *Structural equation models are formal depictions of the relationships between a set of variables. The relationships originate from subject-matter hypotheses that specify causal connections, noncausal associations, or lack of associations between variables. The variables can be latent or observed. Causal effects can be direct or indirect. The representation of the model is in a set of equations and assumptions or in a path diagram with the equivalent information.*

Bollen (2026). *Elements of Structural Equation Models*. Cambridge University Press

SECTION 1: MOTIVATION & INTUITION

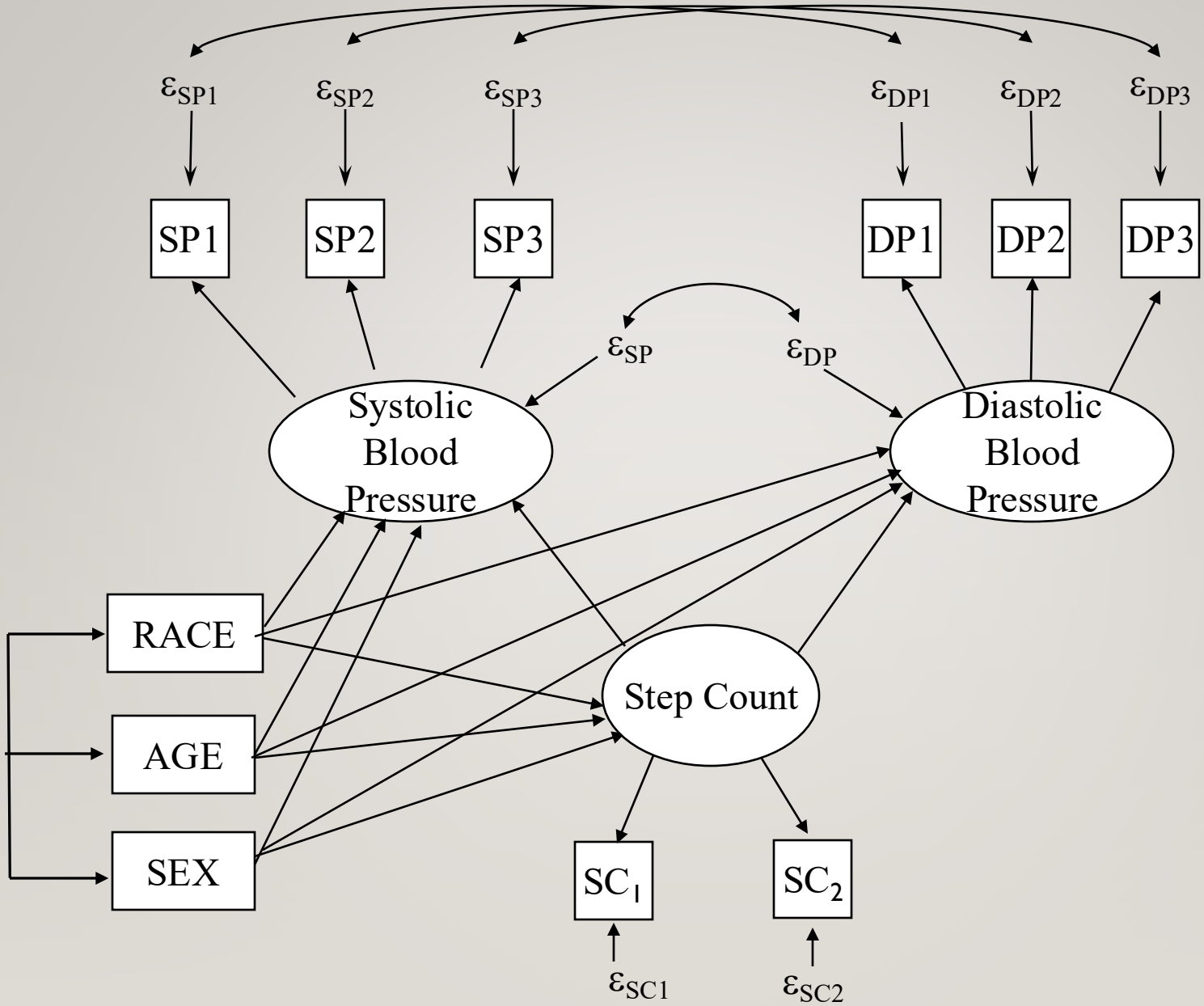
- Problem: How much measurement error in smart watch readings of blood pressure?
  - Measures: Systolic (SP) & diastolic (DP) blood pressure readings repeatedly measured 3 times: SP1, SP2, SP3, DP1, DP2, DP3
  - Sample: 250 individuals with smart watch measures
    - How reliable & valid are measures?
    - How correlated are systolic & diastolic blood pressure?
    - Are errors of same time measures of SP and DP correlated?

Path Diagram of Systolic and Diastolic Blood Pressure with 3 Smart Watch Measures of Each (SP1, SP2, SP3; DP1, DP2, DP3)



SECTION 1: MOTIVATION & INTUITION

- Problem: Impact of step-counts on blood pressure controlling for age, sex, and race?
  - Two measures of step-count: smart watch (SC1) and smart phone (SC2)
    - Measurement errors in step count
    - Estimate effects controlling for measurement errors
  - Measures of age, sex at birth, race
    - Affect step-counts and blood pressure


εSP1
εSP2
εSP3
εDP1
εDP2
εDP3
SP1
SP2
SP3
DP1
DP2
DP3
εSP
εDP
Systolic Blood Pressure
Diastolic Blood Pressure
RACE
AGE
SEX
Step Count
SC1
SC2
εSC1
εSC2

COMMON RESEARCH ISSUES

- Latent Variables (e.g., true blood pressure, step count)
- Measurement Errors
- Complex Causal Paths with Mediation
  - Direct, indirect, and total effects
- Missing Data
- Categorical and Count Measures
- Empirical Tests of Model Assumptions

THE SEM ADVANTAGE

- Unified modeling framework
  - Regression, factor analysis, multilevel models, longitudinal models, error in variables regression, simultaneous equations, and more
- Path analysis graphic diagrams to represent models
  - Pictorial representation of system of equations
  - Century old approach (Sewall Wright founder)
  - Precedes DAGs by decades
- Suitable for experimental and observational data
- Wide variety of estimators, fit statistics, and diagnostics unavailable with other approaches

UNDERSTANDING STRUCTURAL EQUATION MODELS (SEMs)

Section 1: Motivation & Intuition

Section 2: SEM Core Concepts

Section 3: Model Estimation, Fit, & Coefficients

Section 4: Real-World SEM Examples

Section 5: Advanced SEM Capabilities

Section 6: Summary & Getting Started

OBSERVED VS. LATENT VARIABLES

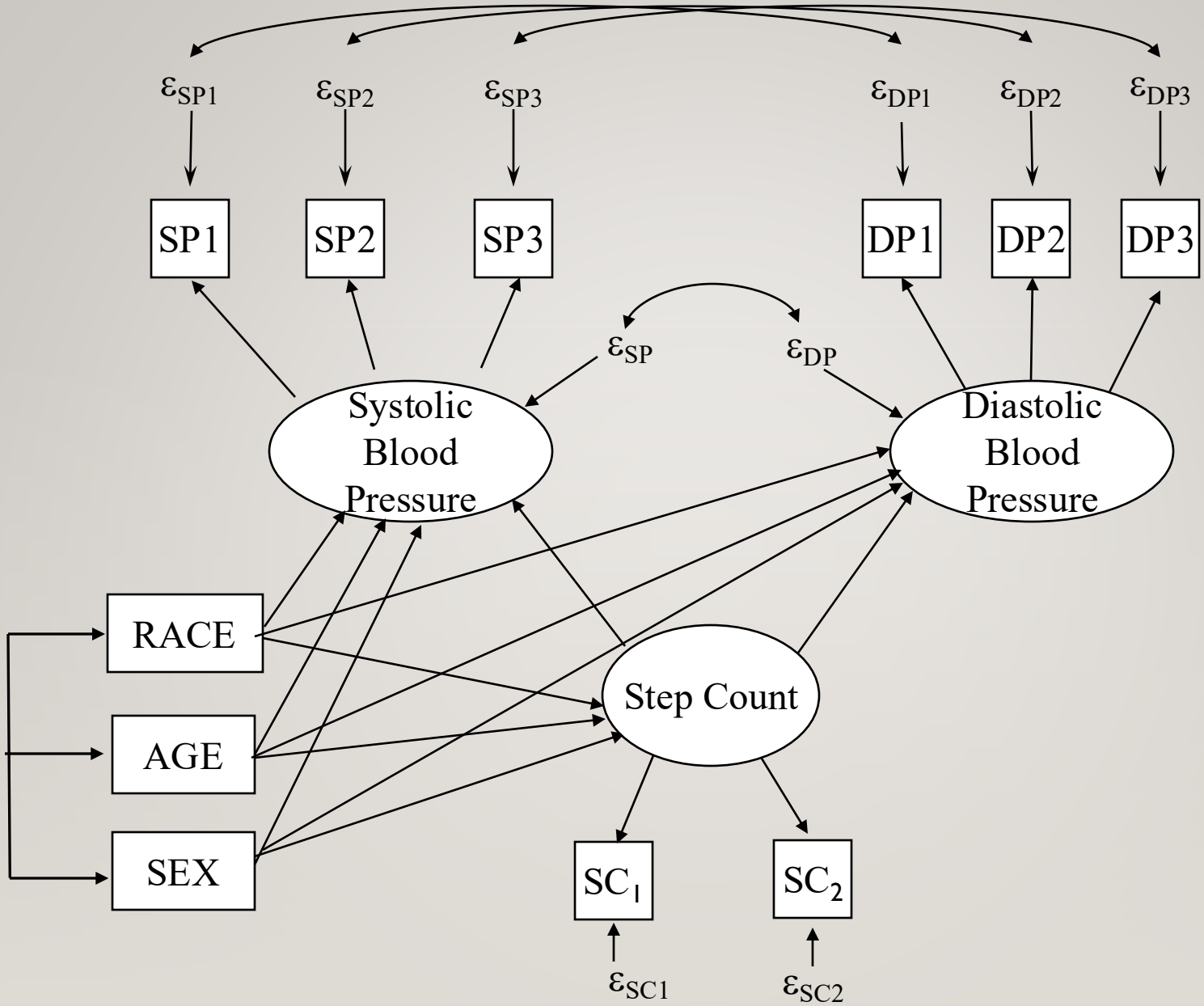
- *Observed Variables*
  - Measured variables that are part of our data set
  - E.g., Respondent's reported education or income; air particulates in city or density of city; LDL cholesterol of individuals; weight of objects
  - Key feature: values are in dataset
    - Variables' values are accessible

OBSERVED VS. LATENT VARIABLES

- *Latent Variables*
  - Variables of interest, but not part of our data set
  - E.g., education level with no measurement error; weight of object without errors; true systolic blood pressure
  - Latent variables can be relatively concrete (e.g., weight, blood pressure, age) or more abstract (e.g., happiness, depression, industrialization of country, conservativeness of political party)
  - Key feature: important substantively, but no sample realizations of variable (Bollen, 2002)
    - Variables' values not accessible

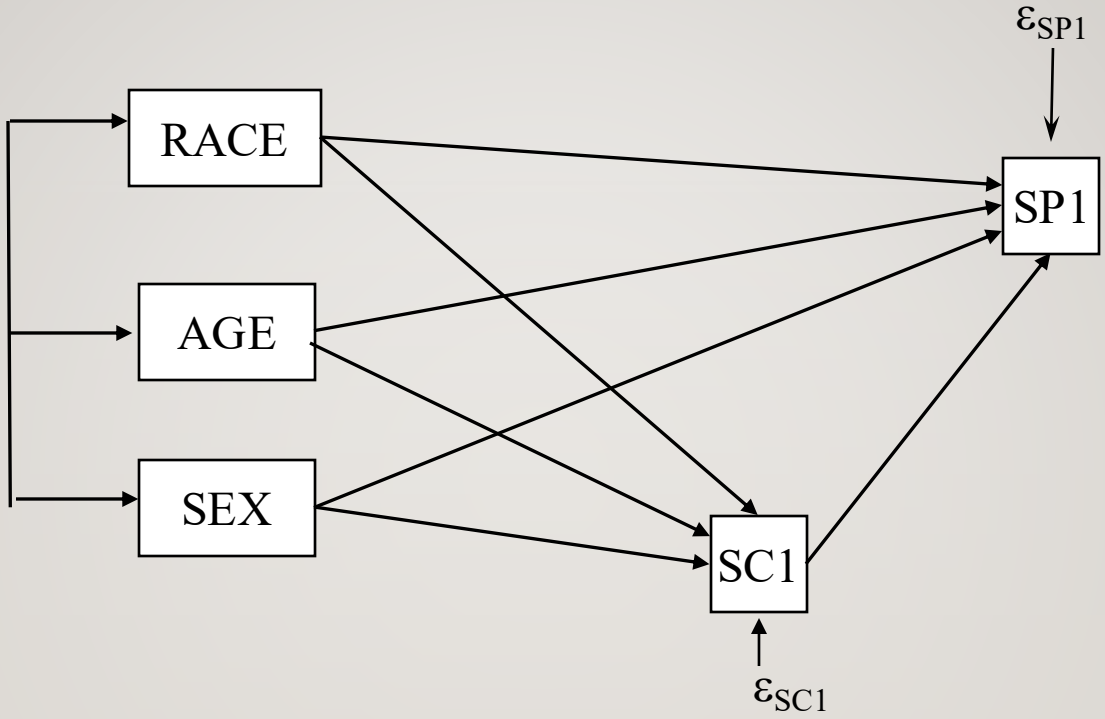
PATH DIAGRAMS 101

- Latent variables in ellipses or circles
- Observed variables in boxes
- Error (or disturbance) variables unenclosed
- Direct effect is single headed arrow from cause to effect
- Unanalyzed association is curved two-headed arrow
  - Correlation due to direct relations between connected variables, confounders, or any other unspecified reason

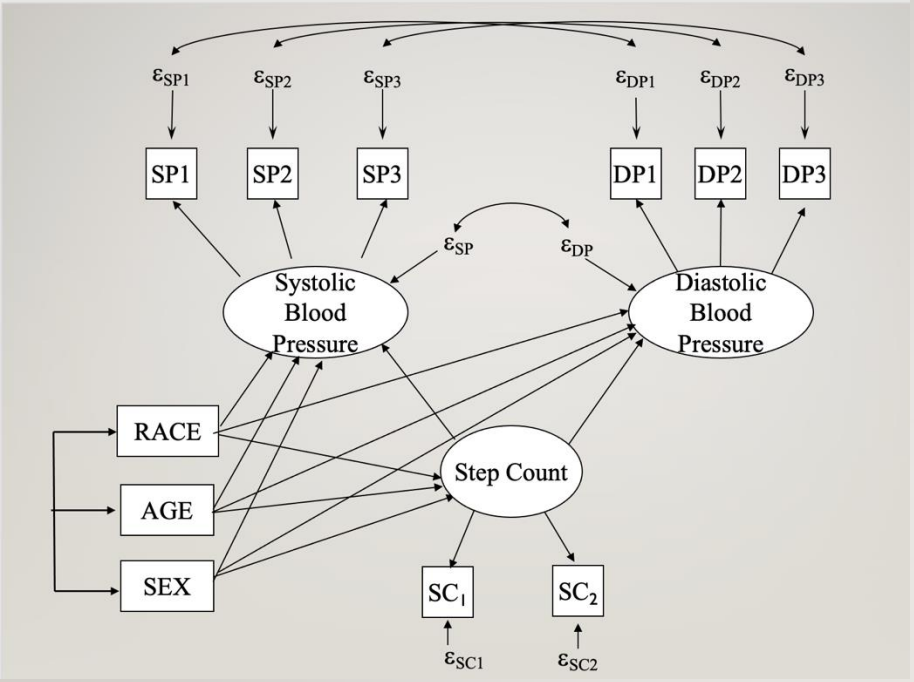

εSP1
εSP2
εSP3
εDP1
εDP2
εDP3
SP1
SP2
SP3
DP1
DP2
DP3
εSP
εDP
Systolic Blood Pressure
Diastolic Blood Pressure
RACE
AGE
SEX
Step Count
SC1
SC2
εSC1
εSC2

MEASUREMENT ERROR MATTERS

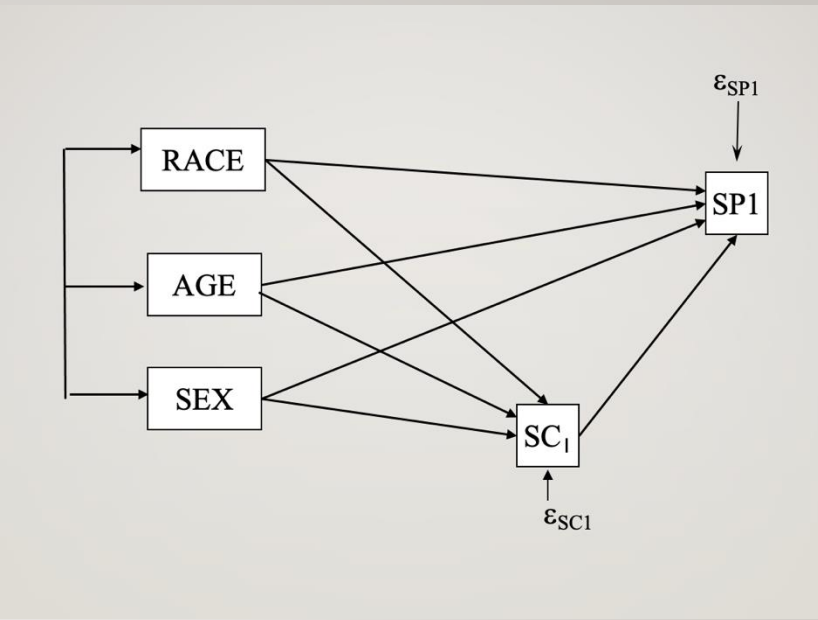
- Scientific hypotheses formulated between latent variables
  - Empirical tests with observed variables can bias tests
  - Direction of bias difficult to predict
- Accurate scientific descriptions affected by measurement errors
- Scientific research that ignores measurement errors creates uncertainty as to true effects



SC1 = 1st step count measure
SP1 = 1st systolic blood pressure measure



versus



Results with latent variable SEM:

1. Control for measurement error
2. Reliability estimates of measures
3. Mediation analysis while controlling measurement error
4. Demographic variable effects
5. Unbiased estimates of effects

Results with typical modeling of observed variables

1. No control for measurement errors
2. No reliability estimates
3. Distorted mediation analysis
4. Coefficients of all variables biased in unknown direction and magnitude

SEM:
A GENERAL MODELING FRAMEWORK

SEMs Allow:

- Latent and Observed Variables
- Random and Nonrandom Errors
- Errors-in-Variables Regressions
- Multiple Indicators
- Linear & Nonlinear Restrictions on Parameters
- Tests of Model Fit
- Nonnormal Variables
- Categorical Outcomes (dichotomous, ordinal, censored)

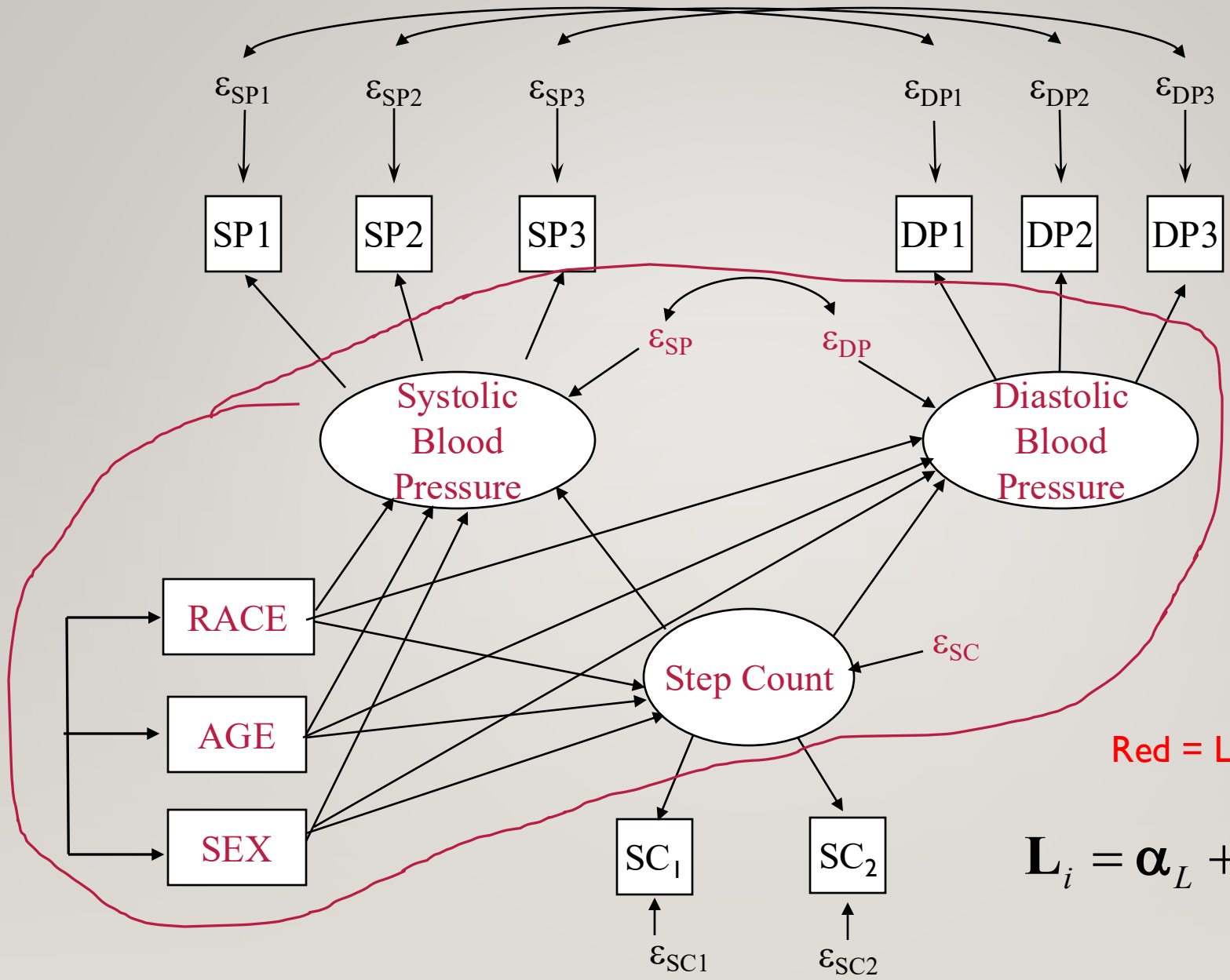
SEM:
A GENERAL MODELING FRAMEWORK

- Latent Variable Model

$$\mathbf{L}_i = \boldsymbol{\alpha}_L + \mathbf{B}\mathbf{L}_i + \boldsymbol{\varepsilon}_{Li}$$

- Measurement Model (continuous indicators)

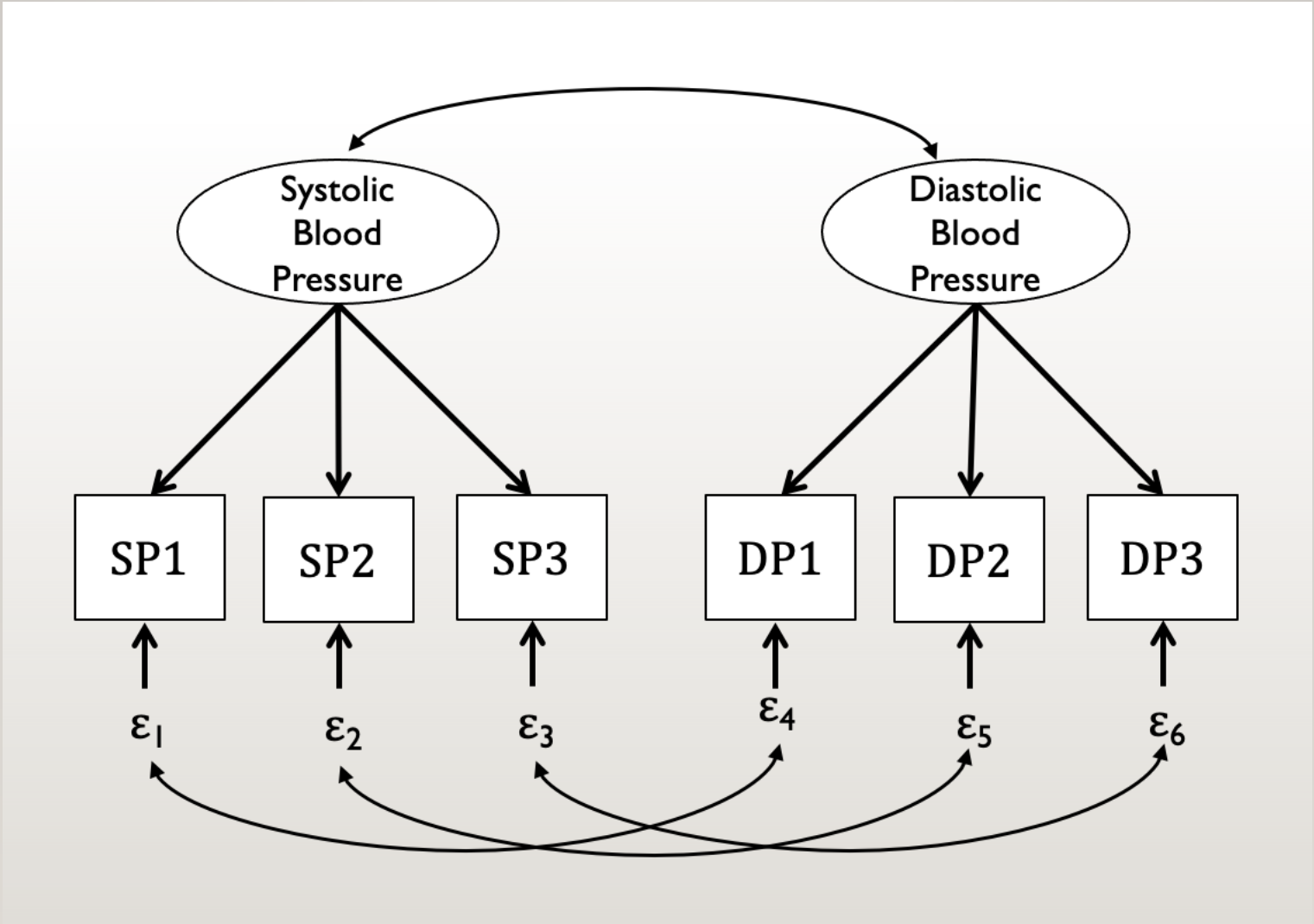
$$\mathbf{Z}_i = \boldsymbol{\alpha}_z + \boldsymbol{\Lambda}\mathbf{L}_i + \boldsymbol{\varepsilon}_{zi}$$



Red = Latent Variable Model

$$\mathbf{L}_i = \boldsymbol{\alpha}_L + \mathbf{B}\mathbf{L}_i + \boldsymbol{\varepsilon}_{Li}$$

Measurement Model: $\mathbf{Z}_i = \boldsymbol{\alpha}_z + \boldsymbol{\Lambda}\mathbf{L}_i + \boldsymbol{\varepsilon}_{zi}$



SEM: A GENERAL MODELING FRAMEWORK

$$\mathbf{L}_i = \boldsymbol{\alpha}_L + \mathbf{B}\mathbf{L}_i + \boldsymbol{\varepsilon}_{Li} \qquad \mathbf{Z}_i = \boldsymbol{\alpha}_z + \boldsymbol{\Lambda}\mathbf{L}_i + \boldsymbol{\varepsilon}_{zi}$$

- Other statistical models are special cases
  - E.g., Z equation alone handles all factor analysis models
  - OR Make Z = L (no measurement error), leads to simultaneous equations, multiple regression, ANOVA, ANCOVA, etc.
  - Other special cases: fixed & random effects, growth curve models, higher order factor analysis, regressions with errors in variables, and so on.
  - Extensions available for categorical and count outcomes

UNDERSTANDING STRUCTURAL EQUATION MODELS (SEMs)

Section 1: Motivation & Intuition

Section 2: SEM Core Concepts

Section 3: Model Estimation, Fit, & Coefficients

Section 4: Real-World SEM Examples

Section 5: Advanced SEM Capabilities

Section 6: Summary & Getting Started

ESTIMATION

- Systemwide (Full Information) Estimators
  - Simultaneous estimation of all parameters
  - Maximum likelihood (ML), Weighted Least Squares (WLS)
- Equation-by-equation (Limited Information) Estimators
  - Model Implied Instrumental Variables, 2SLS (MIIV-2SLS)
- Nonnormal distributions
  - Distribution-robust versions of ML, WLS
  - MIIV-2SLS "distribution-free"

MODEL FIT INDEXES

Fit Index	Guidelines on Fit
Model overidentification test (T_{ML}, *df*, *p-value*)	High *p-values* (>0.05 better than less)
Tucker-Lewis Index (TLI)	> 0.90 or 0.95
Relative Noncentrality Index (RNI/CFI)	> 0.90 or 0.95
Bayesian Information Criterion (BIC)	Lowest value across models
Root Mean Square of Approximation (RMSEA)	< 0.10 or 0.08

Caution: Good Fit ≠ Proof of True Model

INTERPRETING COEFFICIENTS AND EQUATIONS

Statistic	Interpretation
Equation overidentification test (T_S, *df*, *p-value*)	High *p-values* (>0.05 better than less)
Regression coefficients	Expected effects of variable on outcome, net of other covariates
Factor loadings	Expected effects of latent variable on measure, net of other variables
R-squares	Explained variances by determinants of dependent variable
Improper estimates (var< 0 or |correlation| > 1)	Misspecifications or sampling fluctuations

UNDERSTANDING STRUCTURAL EQUATION MODELS (SEMs)

Section 1: Motivation & Intuition

Section 2: SEM Core Concepts

Section 3: Model Estimation, Fit, & Coefficients

Section 4: Real-World SEM Examples

Section 5: Advanced SEM Capabilities

Section 6: Summary & Getting Started

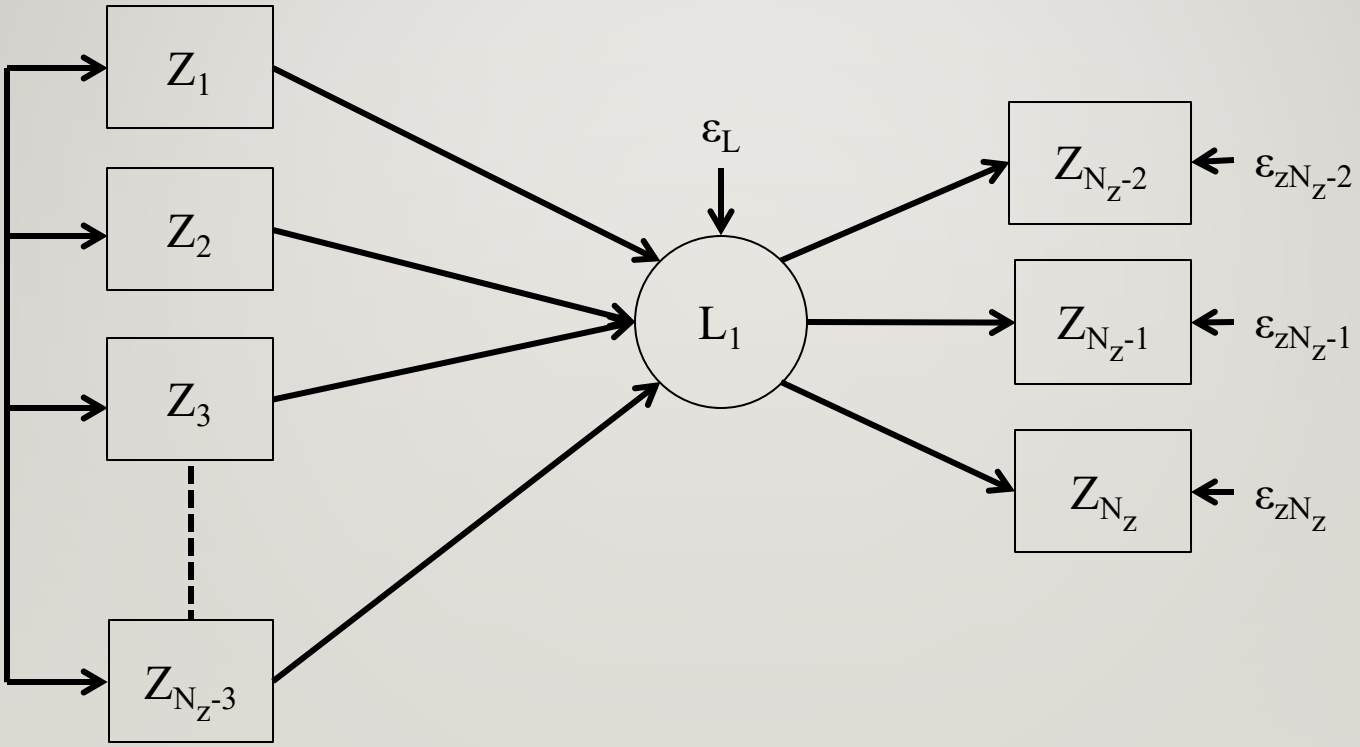
EXAMPLE 1: MEASURES AND DETERMINANTS OF HOME VALUES

- Robins, P. K., & West, R. W. (1977). Measurement error in the estimation of home value. *Journal of the American Statistical Association, 72*(359), 290–294.
- Problems:
  - 1) Measurement error in appraised, owner, and assessed values of home
  - 2) What variables impact the value of home (controlling for error)?

EXAMPLE 1: MEASURES AND DETERMINANTS OF HOME VALUES

- Robins and West (1977)
  - L_1 = value of home
  - Z_1 = lot size
  - Z_2 = square footage
  - Z_3 = number of rooms
  - Z_3 to Z_{N_z-3} = other causal indicators
  - Z_{N_z-2} = appraised value
  - Z_{N_z-1} = owner estimate
  - Z_{N_z} = assessed value
  - ε_L, ε_{zN_z-1}, ε_{zN_z-2}, ε_{zN_z-3} = disturbances (errors)
  - N_z = # of observed variables

EXAMPLE 1: MEASURES AND DETERMINANTS OF HOME VALUES



EXAMPLE 1: MEASURES AND DETERMINANTS OF HOME VALUES

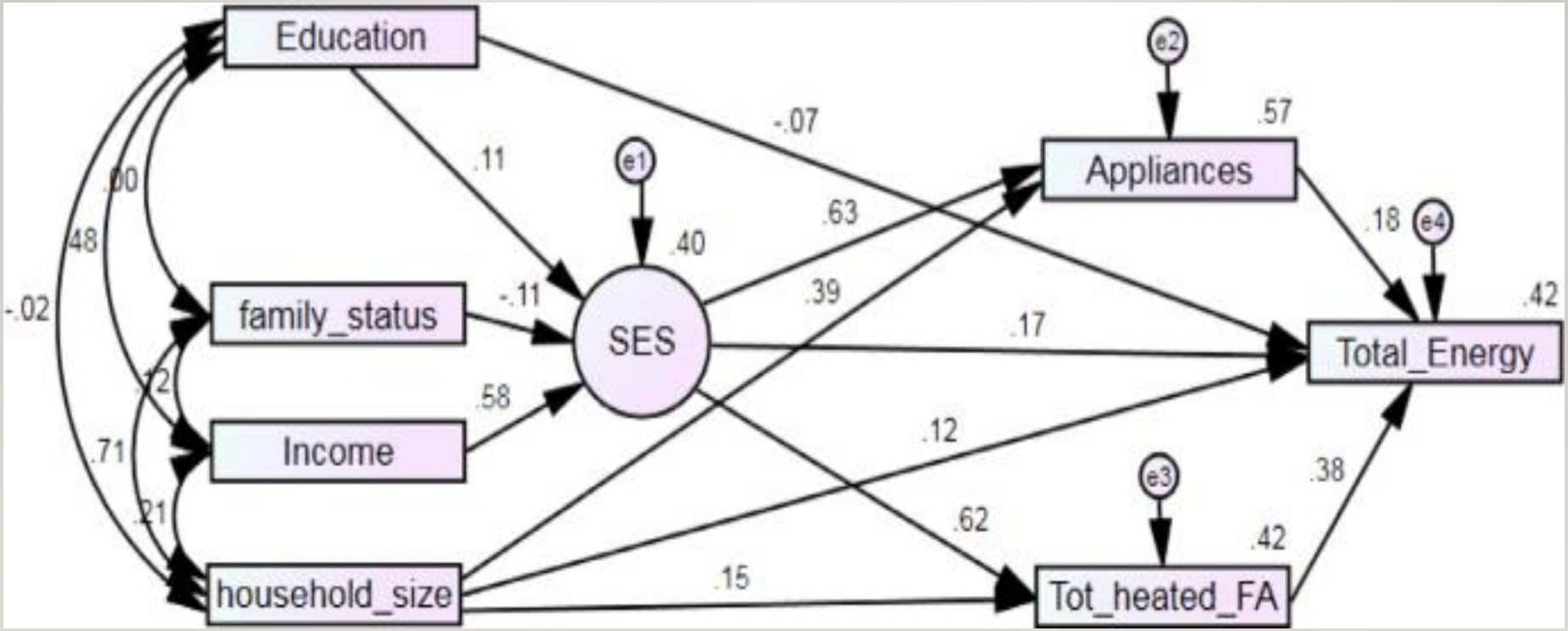
Maximum Likelihood Estimates (N = 138)				
a. Home value measurement equations				
Variable	α	Λ	V(ε)	R^2
Appraised value	0	1.000	2.415	0.553
		(-)	(0.158)	
Owner-estimate	1.257	0.973	2.771	0.470
		(0.119)	(0.177)	
Assessed value	-5.909	1.339	1.612	0.832
		(0.120)	(0.169)	

Maximum Likelihood Estimates (N = 138)		
b. Home value equation		
Z Variable	B	SE
Construction grade	1.077	0.201
1 if attached garage	0.134	0.555
1 if detached garage	0.693	0.248
1 if basement garage	1.537	0.433
Finished area (100s of sq ft)	0.223	0.045
1 if substandard storage	-1.391	0.679
1 if quality below neigh stds	-1.144	0.646
Number of stories	0.493	0.366
Number of built-ins	0.461	0.172
Effective age	-0.073	0.010
Number of rooms	0.287	0.129
Lot size (100s of sq ft)	0.014	0.006
α (intercept)	5.704	-

EXAMPLE 2: SOCIOECONOMIC STATUS (SES) AND RESIDENTIAL ENERGY CONSUMPTION

- Karatasou, S., & Santamouris, M. (2019). Socio-economic status and residential energy consumption: A latent variable approach. *Energy and Buildings, 198*, 100-105. https://doi.org/10.1016/j.enbuild.2019.06.013
- Problems:
  - 1) Determining the impact of SES on household energy consumption
  - 2) Best way to build model for measures of SES
  - 3) Look at mediating effects

EXAMPLE 2: SOCIOECONOMIC STATUS AND RESIDENTIAL ENERGY CONSUMPTION



EXAMPLE 2: SOCIOECONOMIC STATUS AND RESIDENTIAL ENERGY CONSUMPTION (N=5686)

Fit Index	Values
Model overidentification test (T_{ML}, *df*, *p-value*)	T_{ML} = *5.3, df = 3, p-value =0.15*
Tucker-Lewis Index (TLI)	0.998
Relative Noncentrality Index (RNI/CFI)	1.000
Bayesian Information Criterion (BIC)	-20.6
Root Mean Square of Approximation (RMSEA)	0.012

SEM SOFTWARE

- Numerous packages and procedures
  - Mplus (standalone)
  - R packages: *lavaan, blavaan, MIIVsem*, etc.
  - Proc Calis in SAS
  - AMOS in SPSS
  - sem procedure in Stata
- New SEM platform in JMP Pro
  - JMP Statistical Discovery LLC. (2025). JMP® 18.2. Cary, NC: JMP Statistical Discovery LLC.
  - Laura Castro-Schilo, PhD and Chris Gotwalt, PhD at JMP

Structural Equation Models

Model Specification

Specification

Model Name:

Model 2

Add Notes

From List

Filter

- Constant
- Appliances
- Education
- Household Size
- Income
- Family Status
- Total Energy
- Total Heated FA
- SES

To List

Filter

- Constant
- Appliances
- Education
- Household Size
- Income
- Family Status
- Total Energy
- Total Heated FA
- SES

Model Shortcuts

Latent1

Action

Fix To | Free | Undo | Run

Set Equal | Remove | Redo | Reset

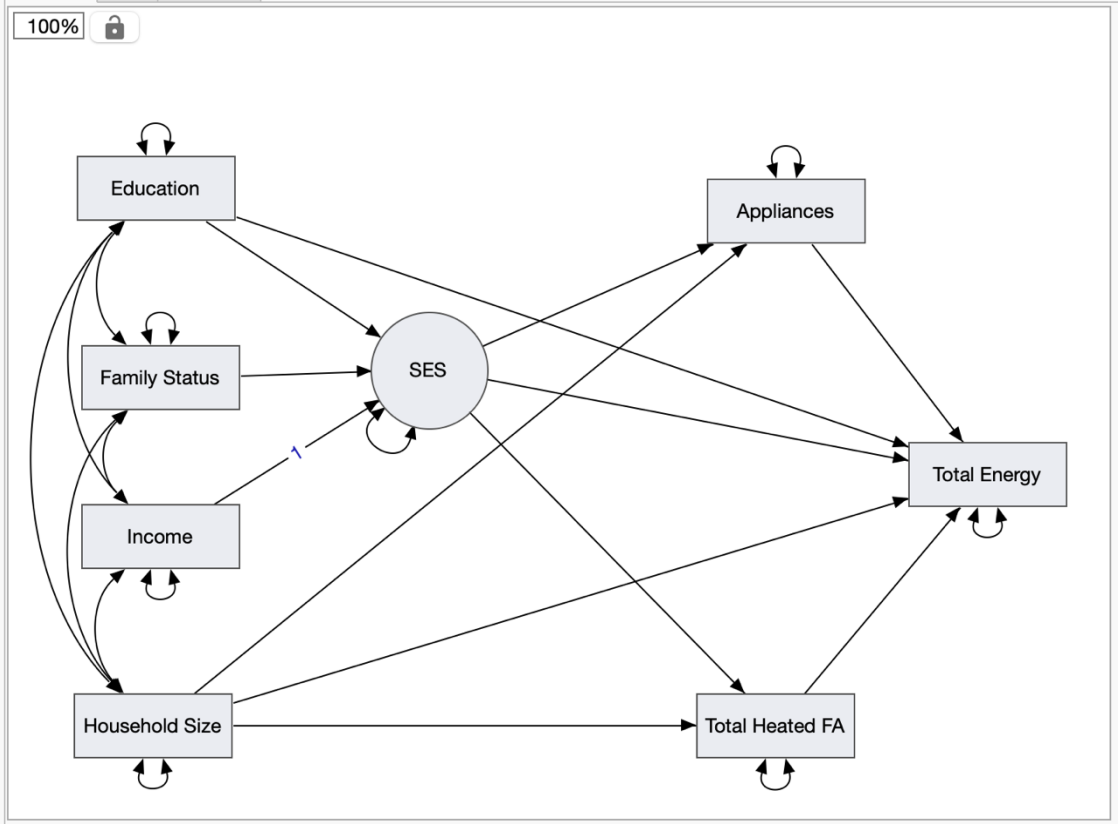
Diagram

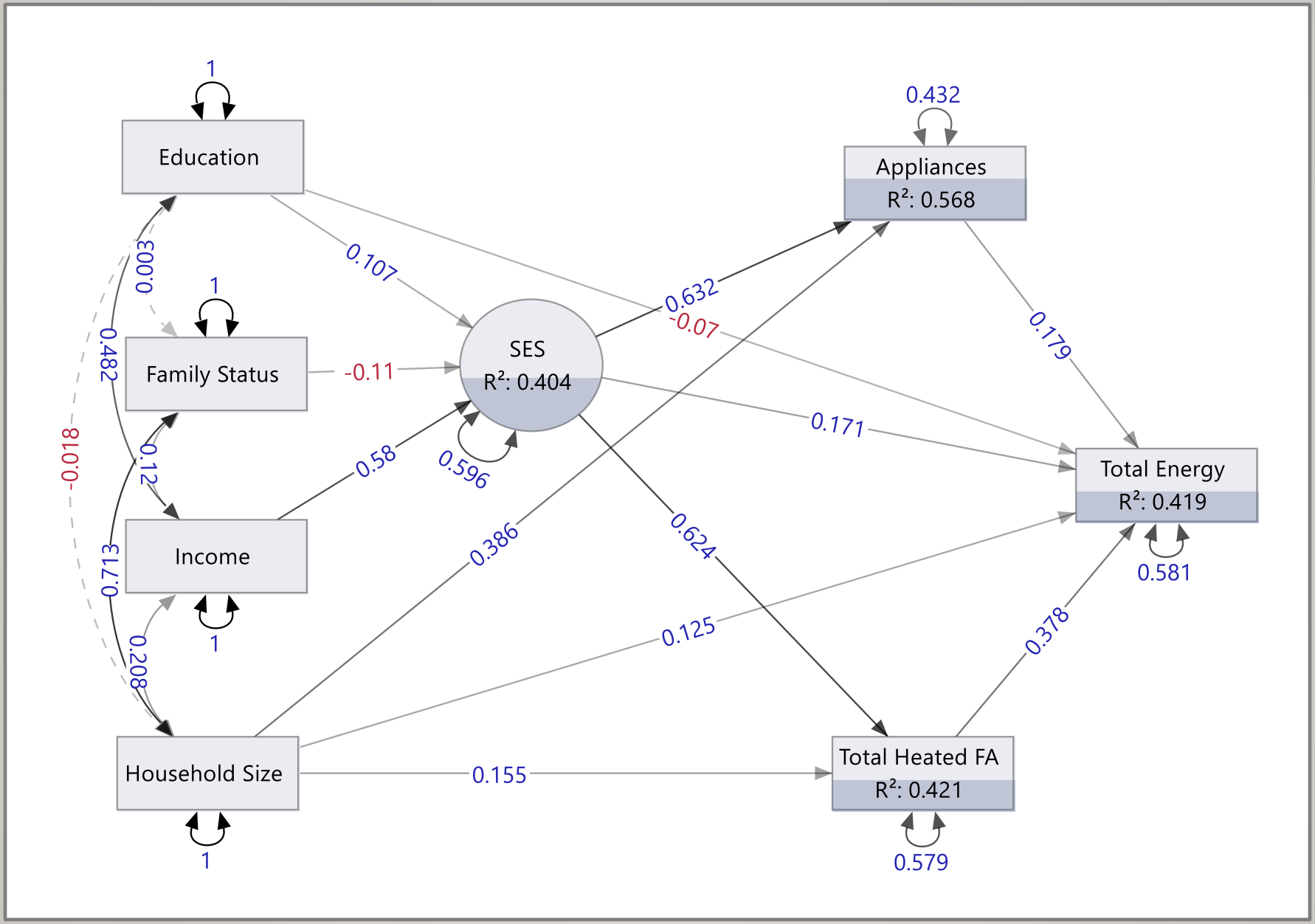
Details

Manifest	Latent	Free	DF	Iterations
7	1	32	3	1000

Diagram | Lists | Status

100%




1
Education
0.432
Appliances
R²: 0.568
0.003
0.107
0.632
-0.07
0.179
1
Family Status
-0.11
SES
R²: 0.404
0.171
Total Energy
R²: 0.419
0.482
-0.018
0.12
0.58
0.596
0.386
0.624
0.581
0.713
Income
1
0.208
0.125
0.378
Household Size
0.155
Total Heated FA
R²: 0.421
1
0.579

Summary of Fit

✓ Maximum Likelihood. Converged in Gradient.

Sample Size	5686
Rows with Missing	0
-2 Log Likelihood	140577.81
Iterations	8
Number of Parameters	32
AICc	140642.18
BICu	-20.62606
ChiSquare	5.3112284
DF	3
Prob>ChiSq	0.1503755
CFI	0.999822
RMSEA	0.0116401
Lower 90%	0
Upper 90%	0.0275405

Parameter Estimates

Means/Intercepts	Estimate	Std Error	Wald Z	Prob>|Z|
Constant → Family Status	0.3239536	0.0062062	52.198323	<.0001*
Constant → Total Heated FA	0.6271955	0.0451115	13.903234	<.0001*
Constant → Total Energy	1.8057238	0.2394557	7.5409529	<.0001*
Constant → Household Size	2.577383	0.0189919	135.70927	<.0001*
Constant → Education	3.1310236	0.0152099	205.85441	<.0001*
Constant → Income	3.669891	0.0295543	124.17449	<.0001*
Constant → Appliances	4.1248174	0.1376377	29.968669	<.0001*
Loadings	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
SES → Appliances	0.6170862	0.0201716	30.591897	<.0001*
SES → Total Heated FA	0.1911773	0.0069007	27.704209	<.0001*
Regressions	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Education → SES	0.3578676	0.0628015	5.6983916	<.0001*
Income → SES	1	0	.	.
Family Status → SES	-0.902809	0.1747157	-5.167305	<.0001*
Appliances → Total Energy	0.224212	0.0334605	6.7008034	<.0001*
Total Heated FA → Total Energy	1.5061149	0.0833788	18.063514	<.0001*
Education → Total Energy	-0.285541	0.0530409	-5.38341	<.0001*
SES → Total Energy	0.2086497	0.0533591	3.9102934	<.0001*
Household Size → Appliances	1.0122975	0.0383558	26.392307	<.0001*
Household Size → Total Energy	0.4093861	0.0523995	7.8127789	<.0001*
Household Size → Total Heated FA	0.1275401	0.0126714	10.065172	<.0001*
Variances	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Appliances ↔ Appliances	6.0939873	0.2082146	29.267821	<.0001*
Education ↔ Education	1.3154041	0.0246701	53.31979	<.0001*
Household Size ↔ Household Size	2.050906	0.0384643	53.31979	<.0001*
Income ↔ Income	4.9664765	0.0931451	53.31979	<.0001*
Family Status ↔ Family Status	0.2190077	0.0041074	53.31979	<.0001*
Total Energy ↔ Total Energy	12.809586	0.26425	48.475264	<.0001*
Total Heated FA ↔ Total Heated FA	0.80428	0.0225072	35.734365	<.0001*
SES ↔ SES	8.8145528	0.6558172	13.440563	<.0001*
Covariances	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Education ↔ Household Size	-0.029924	0.0217857	-1.373583	0.1696
Education ↔ Income	1.232137	0.0376291	32.744264	<.0001*
Education ↔ Family Status	0.0018738	0.007118	0.2632519	0.7924
Household Size ↔ Income	0.6624601	0.0432268	15.325203	<.0001*
Household Size ↔ Family Status	0.4777454	0.0109149	43.769927	<.0001*
Income ↔ Family Status	0.1248788	0.0139297	8.9649338	<.0001*

Specific Indirect Effects

Indirect Effect	Estimate	Std Error	Wald Z	Prob>|Z|	Standardized Estimate	.2 .4 .6 .8
Education → SES → Appliances → Total Energy	0.0495139	0.0105155	4.7086422	<.0001*	0.0120951	
Education → SES → Total Heated FA → Total Energy	0.1030426	0.017121	6.0184865	<.0001*	0.0251709	
Education → SES → Total Energy	0.0746689	0.0231517	3.2252068	0.0013*	0.0182399	
Income → SES → Appliances → Total Energy	0.1383581	0.0212817	6.5012821	<.0001*	0.0656721	
Income → SES → Total Heated FA → Total Energy	0.2879349	0.019236	14.968517	<.0001*	0.1366692	
Income → SES → Total Energy	0.2086497	0.0533591	3.9102934	<.0001*	0.0990362	

A subset of specific indirect effects:

UNDERSTANDING STRUCTURAL EQUATION MODELS (SEMs)

Section 1: Motivation & Intuition

Section 2: SEM Core Concepts

Section 3: Model Estimation, Fit, & Coefficients

Section 4: Real-World SEM Examples

Section 5: Advanced SEM Capabilities

Section 6: Summary & Getting Started

SECTION 5: ADVANCED SEM CAPABILITIES (PARTIAL LIST)

- Missing data
- Categorical or count dependent variables or measures
- More general longitudinal models
- Bayesian SEM
- Time series SEM
- Test equivalency of parameters across groups or over time

UNDERSTANDING STRUCTURAL EQUATION MODELS (SEMs)

Section 1: Motivation & Intuition

Section 2: SEM Core Concepts

Section 3: Model Estimation, Fit, & Coefficients

Section 4: Real-World SEM Examples

Section 5: Advanced SEM Capabilities

Section 6: Summary & Getting Started

SECTION 6: SUMMARY & GETTING STARTED

- SEM is both general statistical model and method of testing causal inferences
  - Widely used statistical models are special cases of general model
  - Overidentification tests and fit indexes provide assessments of causal assumptions and causal inferences
  - Incorporates latent variables, observed variables, measurement error, multiple indicators in multiequation system
- Common Practice of *impromptu* regressions & ignoring measurement error = bias results
- SEM common in social, behavioral, information, and ecology, but little known and underutilized in many sciences

SECTION 6: SUMMARY & GETTING STARTED

- Getting Started
  - Online video introductions (YouTube, etc.)
  - Workshops (often fees, sometimes free)
  - Semester courses (many universities offer courses)
  - Books and articles
    - Numerous sources available
  - Partner with colleagues or students
    - Learning as group can be helpful

THANK YOU !!