

Clase 16: Error de medición

¿Ahora qué?



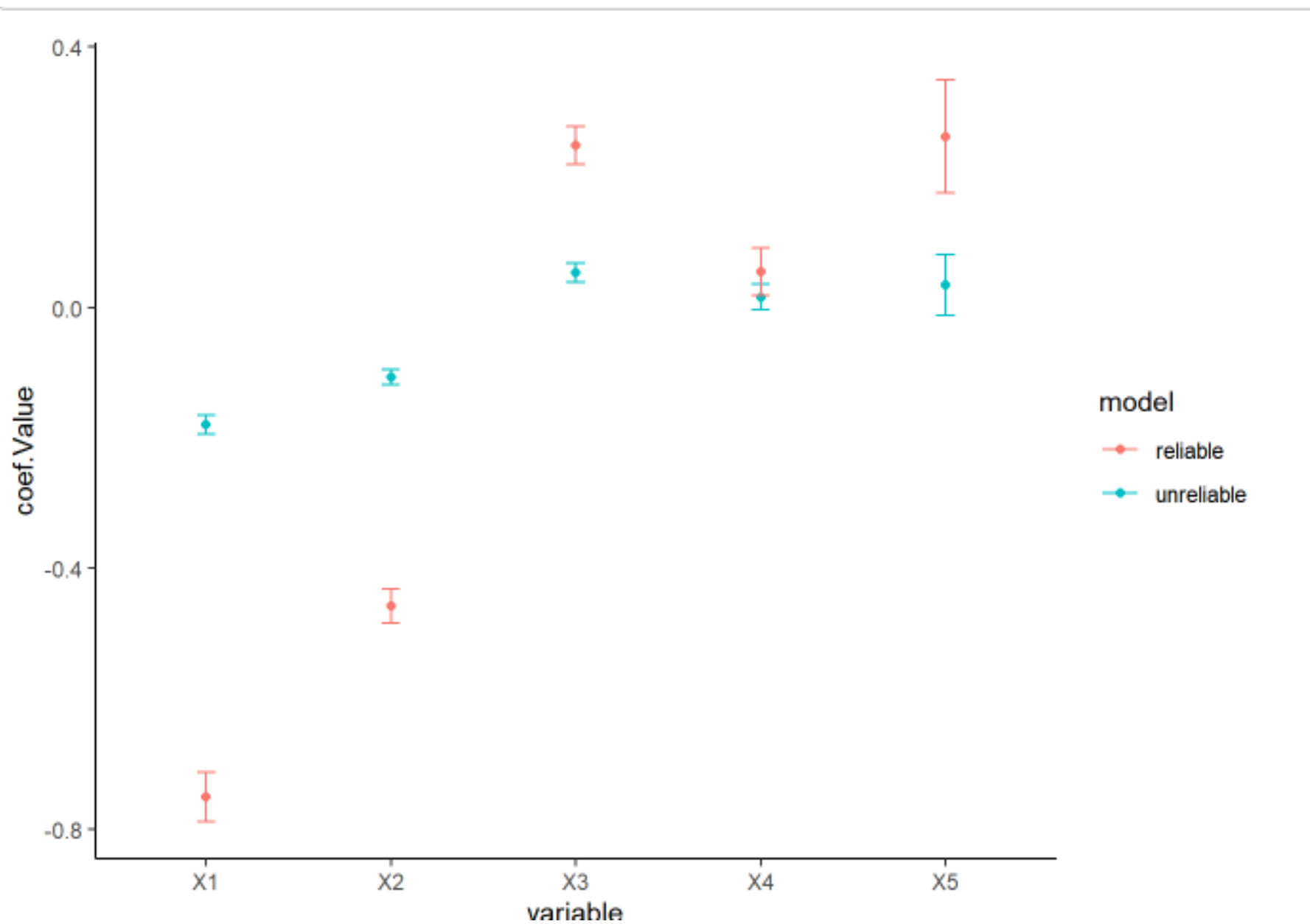
Parte 1



Error de medición

- El error de medición en la variable dependiente, generalmente, **atenúa** los coeficientes de una regresión β
- El error de medición en la variable independiente, generalmente, resulta en **sesgos e inconsistencias** en la β
- El error de medición resulta en errores de clasificación y, generalmente, resulta **sesgos e inconsistencias** en la β
- La atenuación:
 - Lo que es estimable es el error de medición
 - El tamaño de la atenuación es inobservable / latente
 - No equi-proporcional





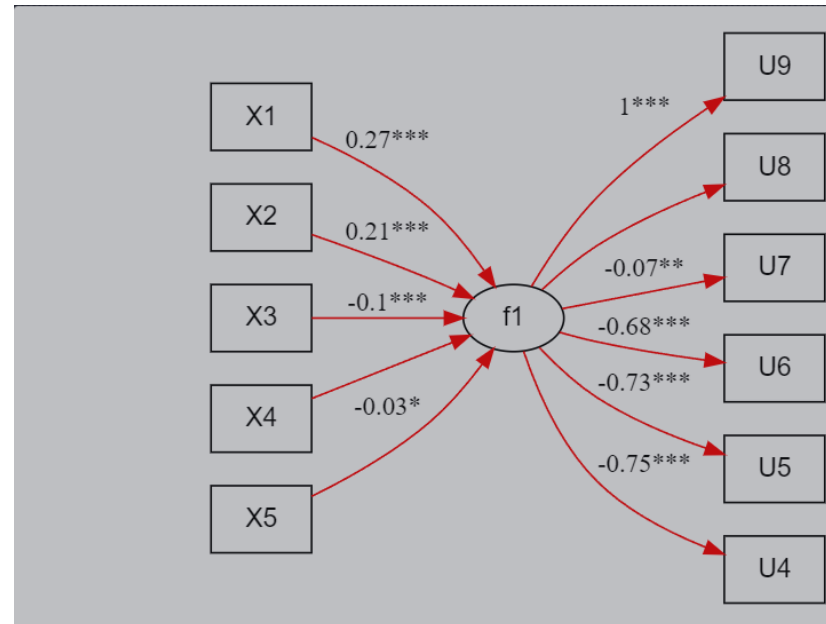
**Error de
medición
aleatorio en la
variable
dependiente**

¿Qué alternativas existen?

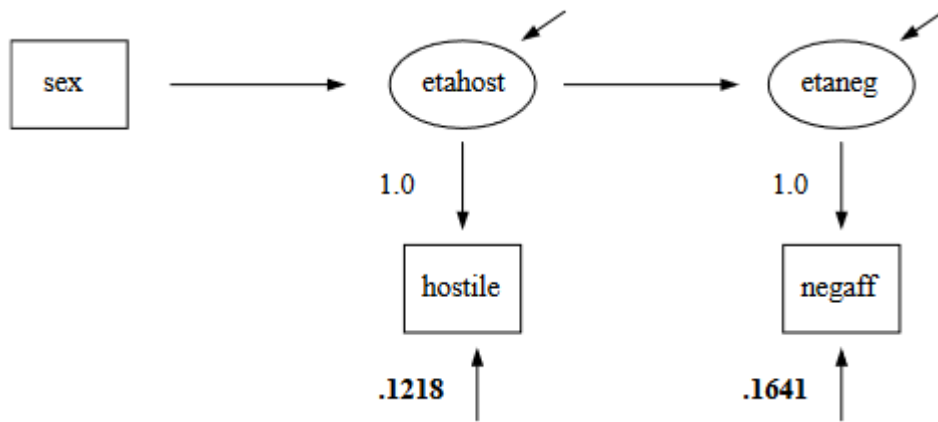
- Estimo confiabilidad, la reporto y no hago ninguna corrección
- Estimo confiabilidad, la reporto y busco hacer alguna corrección:
 - Dependiente: MIMIC model
 - Independiente: Varianza observada y confiabilidad
 - Clases latentes: clasificación
 - Corrección de Hausman: clasificación
 - Bayesian modelling (Hausman + otras)

SEM: Dependiente

Usar los scores del factor en lugar de los scores observados/manuales en un modelo MIMIC: Multiple Indicators
Multiple Causes



SEM: Dependiente



```
! Set variance to (1-reliability) (variance of each variable);  
!   Hostile measurement variance = (1 - .7) (.406) = .1218;  
!   Negaff measurement variance = (1 - .7) (.547) = .1641;  
  
hostile@.1218;  
negaff@.1641;
```

SEM: Dependiente

Uncorrected Results: Assumes No Measurement Error

(note: output format from an earlier version of Mplus, so it is formatted slightly differently)

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
ETAHOST BY HOSTILE	1.000	0.000	0.000	0.636	1.000
ETANEG BY NEGAFF	1.000	0.000	0.000	0.738	1.000
ETANEG ON ETAHOST	0.344	0.067	5.109	0.296	0.296
ETAHOST ON SEX	0.006	0.078	0.077	0.009	0.005
Residual Variances					
HOSTILE	0.000	0.000	0.000	0.000	0.000
NEGAFF	0.000	0.000	0.000	0.000	0.000
ETAHOST	0.405	0.035	11.640	1.000	1.000
ETANEG	0.497	0.043	11.640	0.912	0.912

R-SQUARE

Observed
Variable R-Square

HOSTILE 1.000
NEGAFF 1.000

Latent
Variable R-Square

ETAHOST 0.000
ETANEG 0.088

Results Correcting for Measurement Error Attenuation

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
ETAHOST BY HOSTILE	1.000	0.000	0.000	0.532	0.836
ETANEG BY NEGAFF	1.000	0.000	0.000	0.617	0.836
ETANEG ON ETAHOST	0.492	0.098	5.024	0.424	0.424
ETAHOST ON SEX	-0.009	0.077	-0.115	-0.017	-0.008
Residual Variances					
HOSTILE	0.122	0.000	0.000	0.122	0.301
NEGAFF	0.164	0.000	0.000	0.164	0.301
ETAHOST	0.283	0.035	8.135	1.000	1.000
ETANEG	0.312	0.043	7.182	0.820	0.820

R-SQUARE

Observed
Variable R-Square

HOSTILE 0.699
NEGAFF 0.699

Latent
Variable R-Square

ETAHOST 0.000
ETANEG 0.180



SEM: Independiente

Si pensamos que la confiabilidad mide la varianza de interés y conocemos la variabilidad total de la variable en cuestión:

$$X_{\text{varm}} = (1 - \omega) (X_{\text{var}})$$

Es decir, controlamos la varianza de X según lo que verdaderamente nos interesa de X



Parte 2



Hausman

Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left

Jerry Hausman

JOURNAL OF ECONOMIC PERSPECTIVES
VOL. 15, NO. 4, FALL 2001
(pp. 57-67)

Download Full Text PDF
(Complimentary)

Article Information

Comments (0)

Abstract

The effect of mismeasured variables in the most straightforward regression analysis with a single regressor variable leads to a least squares estimate that is downward biased in magnitude toward zero. I begin by reviewing classical issues involving mismeasured variables. I then consider three recent developments for mismeasurement econometric models. The first issue involves difficulties in using instrumental variables. A second involves the consistent estimators that have recently been developed for mismeasured nonlinear regression models. Finally, I return to mismeasured left hand side variables, where I will focus on issues in binary choice models and duration models.



Hausman variables categóricas dependientes

Measurement Error in the Left-Hand Side Variables: Probit and Logit

In the usual linear regression specification, a mismeasured left-hand side variable does not lead to a biased coefficient, as discussed above, but only to less statistical precision in estimation. However, in certain contexts, misclassification of the left-hand side variable can lead to estimators that are biased and inconsistent. In particular, this situation often arises when the dependent variable is limited in some way: for example, in a probit or logit estimation where the dependent variable takes on only two values, zero or one.¹⁰ For example, consider a case in which the left-hand variable is whether a person has changed jobs or not.

Corrección de Hausman y Abrevaya

- Abrevaya, Jason and Jerry Hausman. 1999. “Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells.” *Annales D’Economie et de Statistique*. 55-56, pp. 243-75.
- Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998): “Misclassification of a Dependent Variable in a Discrete-Response Setting,” *Journal of Econometrics* 87: 239-269.



Table 1
Monte Carlo simulation results ($n = 5000$)

	True	Probit Coefficient estimates	Ratio to constant	MLE ($\alpha = \alpha_0 = \alpha_1$) Coefficient estimates	Ratio to constant
α	0.02	—	—	0.0192 (0.0054)	—
β_0	— 1.0	— 0.787 (0.069)	—	— 0.990 (0.068)	—
β_1	0.2	0.158 (0.001)	0.20	0.199 (0.008)	0.20
β_2	1.5	1.27 (0.06)	1.61	1.49 (0.08)	1.51
β_3	— 0.6	— 0.158 (0.023)	0.66	— 0.598 (0.026)	0.60
α	0.05	—	—	0.0497 (0.0076)	—
β_0	— 1.0	— 0.567 (0.073)	—	— 1.007 (0.084)	—
β_1	0.2	0.114 (0.010)	0.20	0.201 (0.010)	0.20
β_2	1.5	1.06 (0.05)	1.87	1.50 (0.08)	1.50
β_3	— 0.6	— 0.431 (0.019)	0.76	— 0.599 (0.032)	0.60

Table 3
CPS coefficient estimates

	Probit	MLE $\alpha_0 = \alpha_1$	MLE $\alpha_0 \neq \alpha_1$	MRC/IR
α_0	—	0.058 (0.007)	0.061 (0.007)	0.035 (0.015)
α_1	—	0.058 (0.007)	0.309 (0.174)	0.395 (0.091)
Married	− 0.108 (0.049)	− 0.073 (0.077)	− 0.103 (0.100)	− 0.161 (0.191)
Last grade attended	0.026 (0.009)	0.063 (0.015)	0.080 (0.026)	0.052 (0.043)
Age	− 0.022 (0.003)	− 0.028 (0.005)	− 0.033 (0.007)	− 0.035 (0.021)
Union membership	− 0.434 (0.061)	− 0.707 (0.148)	− 0.811 (0.199)	− 0.794 (0.503)
Earnings per week	− 0.001 (0.0001)	− 0.003 (0.0004)	− 0.004 (0.0009)	− 0.003 (0.0015)
Western region	0.214 (0.054)	0.301 (0.086)	0.367 (0.127)	0.367 (—)
Constant	0.051 (0.162)	0.171 (0.259)	0.581 (0.495)	—
Log likelihood	− 1958.1	− 1941.4	− 1940.9	—
Number of obs.	5221	5221	5221	—

Note: Standard errors are in parentheses. The MRC coefficient estimates have been normalized to have the same value for western region as the MLE with $\alpha_0 \neq \alpha_1$. There is no associated standard error on ‘western region’ due to the normalization. The IR estimates are the point estimates from the first and last steps of the isotonic regression step-function estimate.

Error de clasificación

Hausman et al. Correction y ML

Abrevaya tiene el *.ado file en un repositorio en la Universidad de Texas

[Home](#) > [Behaviormetrika](#) > [Article](#)

Logistic regression with misclassification in binary outcome variables: a method and software

Note | Published: 23 August 2017 | 44, 447–476 (2017)



Behaviormetrika

[Aims and scope](#) →

[Submit manuscript](#) →

[Haiyan Liu](#) ✉ & [Zhiyong Zhang](#)

[Access this article](#)



Parte 3





Bayesian Hausman correction

Research Article

Misclassification Error, Binary Regression Bias, and Reliability in Multidimensional Poverty Measurement: An Estimation Approach Based on Bayesian Modelling

Hector Najera 

Pages 63-81 | Published online: 21 Jun 2023

 Cite this article  <https://doi.org/10.1080/15366367.2022.2026104>

 Check for updates

 Full Article

 Figures & data

 References

 Citations

 Metrics

 Reprints & Permissions

Read this article

ABSTRACT

Measurement error affects the quality of population orderings of an index and, hence, increases the misclassification of the poor and the non-poor groups and affects statistical inferences from binary regression models. Hence, the conclusions about the extent, profile, and distribution of poverty are likely to be misleading. However, the size and type (false positive/negatives) of classification error have remained untraceable in poverty research. This paper draws upon previous theoretical literature to develop a Bayesian-based estimator of population misclassification and binary-regression coefficient bias. The study uses the reliability values of existing poverty indices to set up a Monte Carlo study based on factor mixture models to illustrate the connections between measurement error, misclassification, and bias and evaluate the procedure and discusses its importance for real-world applications.

Related

People also read

The Importance of Multidimensional Poverty Measurement: Illustration of the Index for Low Income and Exclusion

Héctor E. Ibarra
The Journal of Development Studies
Published online: 21 Jun 2023



Cuadro 1. Estimación del error de clasificación del índice de privación social¹². México 2008-2018

Año	Omega	Error FN [ICr 95%]	Error FP [ICr 95%]
2008	0.75	6 [2-10]	1 [0-2]
2010	0.72	8 [5-11]	1 [0-2]
2012	0.71	11 [8-14]	1 [0-3]
2014	0.72	9 [6-12]	1 [0-2]
2016	0.68	13 [11-16]	1 [0-2]
2018	0.68	14 [11-17]	0 [0-1]

Fuente: cálculos propios con datos de las ENIGH 2008, 2010, 2012, 2014, 2016 y 2018

Cuadro 2. Cambios en la prevalencia de hogares sin carencia sin y con ajuste por error de clasificación. México 2008-2018

Año	% Sin corrección	% Con corrección
2008	27 [26-27]	25 [24-26]
2010	27 [26-27]	23 [21-24]
2012	27 [27-28]	22 [20-24]
2014	29 [29-30]	25 [23-27]
2016	32 [31-32]	24 [22-26]
2018	30 [30-31]	22 [20-25]

Fuente: cálculos propios con datos de las ENIGH 2008, 2010, 2012, 2014, 2016 y 2018



Medida Mexicana 2018

Comparison of binary unadjusted and adjusted models. Mexico 2018

Parameter	Unadjusted model			Adjusted model		
	50%	2.5%	97.5%	50%	2.5%	97.5%
Intercept	1.66	1.52	1.79	2.79	2.51	3.09
Household size	0.03	0.01	0.04	0.05	0.04	0.07
Rural (Ref=Urban)	0.73	0.68	0.78	1.23	1.11	1.34
Household head indigenous	0.45	0.32	0.58	0.59	0.45	0.74
Low secondary education (ref=Primary)	-1.13	-1.19	-1.07	-1.60	-1.74	-1.46
Secondary education (ref=Primary)	-1.51	-1.58	-1.44	-2.04	-2.20	-1.89
Tertiary (ref=Primary)	-1.99	-2.06	-1.92	-2.60	-2.77	-2.43
Household head female	0.09	0.04	0.14	0.08	0.03	0.13
Household head age	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03
% False-Positive	NA	NA	NA	0.00	0.00	1.00
% False-Negative	NA	NA	NA	14.00	11.00	17.00

*Estimation of the adjusted model 300 secs (8-Core 4.0 GHz processor)



Evolución del sesgo de estimación

Cuadro 5. Evolución del sesgo en la estimación del coeficiente medio de la distribución posterior (B_1/B_{1a}) de cada variable 2008-2018

	2008	2010	2012	2014	2016	2018
Constante	0.760	0.767	0.725	0.775	0.652	0.595
Tamaño hogar	1.000	0.875	0.500	0.667	0.667	0.600
Rural (Ref=Urbano)	0.907	0.856	0.800	0.851	0.758	0.593
<u>Indígena jef@ hogar (ref=No indígena)</u>	0.951	0.861	0.824	0.911	0.850	0.763
Secundaria (ref=Primaria o inferior)	0.909	0.874	0.803	0.849	0.755	0.706
Media superior (ref=Primaria o inferior)	0.916	0.877	0.820	0.863	0.783	0.740
Superior (ref=Primaria o inferior)	0.936	0.899	0.848	0.886	0.817	0.765
Mujer jefa hogar	1.000	0.857	0.889	0.750	0.818	1.125
<u>Edad jef@ hogar</u>	1.000	1.000	1.000	1.000	0.667	0.667

Fuente: cálculos propios con datos de las ENIGH 2008, 2010, 2012, 2014, 2016 y 2018

Parte 4

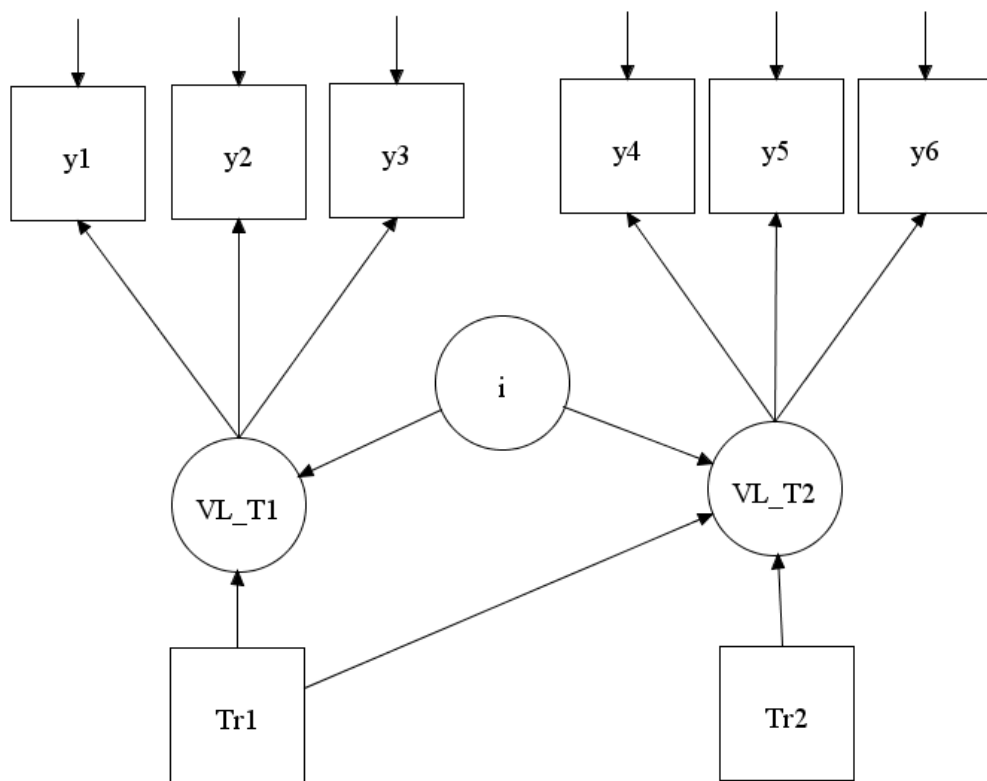


SEM Extensiones: Datos perdidos

- En el análisis de datos hay un principio general en el que la información usada en el diseño de la recolección de los datos debe incluirse en el análisis (i.e., diseño muestral).
- Es un hecho bien sabido que en el análisis de datos lo adecuado es siempre dar cuenta de cualquier característica grupal que prediga la inclusión en el conjunto de datos (i.e., sesgo de selección).
- Prácticamente en todos nuestros análisis tenemos algo tipo de sesgo derivado de datos faltantes o perdidos



SEM Extensiones: Datos perdidos I



Datos perdidos:

¿Pierdo poder? (Tamaño del efecto indetectable)

¿Se sesga mi estimación?

¿Dependiente, independiente o ambas?

Proceso de pérdida:

MCAR: MI, ML o listwise.

MAR: Función de las auxiliares. ML

NMAR: ML (Bayes) + Clases latentes

Simulación de Monte Carlo



SEM Extensiones: Datos perdidos

NMAR II

Psychological Methods
2011, Vol. 16, No. 1, 17–33

© 2011 American Psychological Association
1082-989X/11/\$12.00 DOI: 10.1037/a0022634

Growth Modeling With Nonignorable Dropout: Alternative Analyses of the STAR*D Antidepressant Trial

Bengt Muthén and Tihomir Asparouhov
Muthén & Muthén, Los Angeles, California

Aimee M. Hunter and Andrew F. Leuchter
University of California, Los Angeles

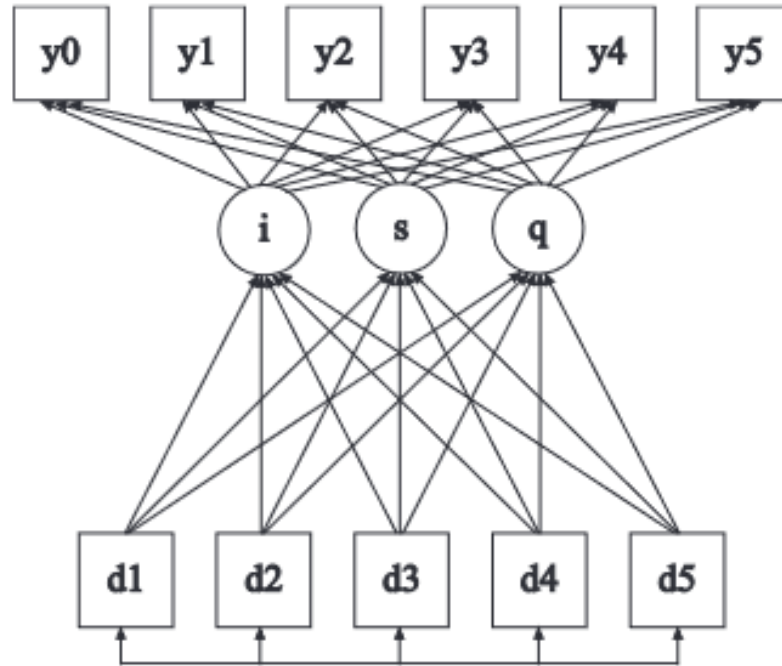


Figure 2. Pattern-mixture modeling (d s are dropout dummy variables).

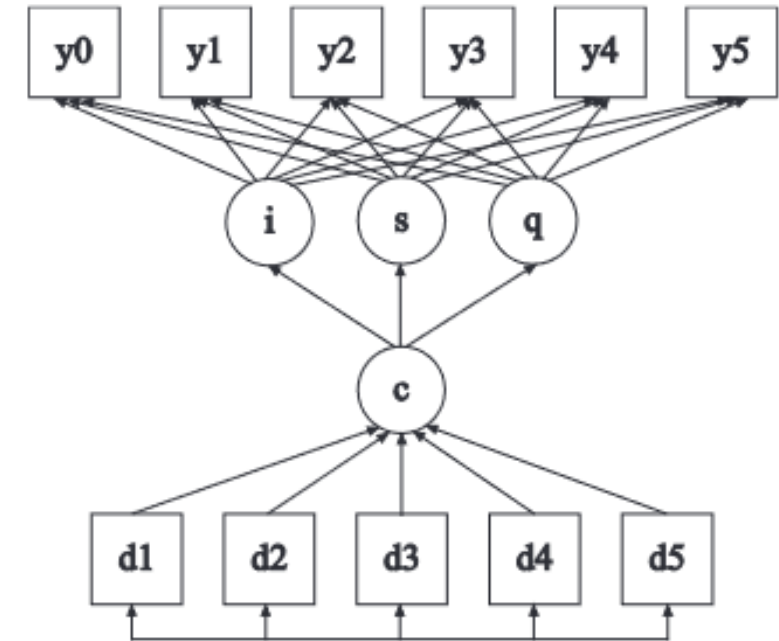


Figure 6. Roy latent class dropout modeling.

SEM ¿Qué ganamos?

GROWTH MODELING WITH NONIGNORABLE DROPOUT

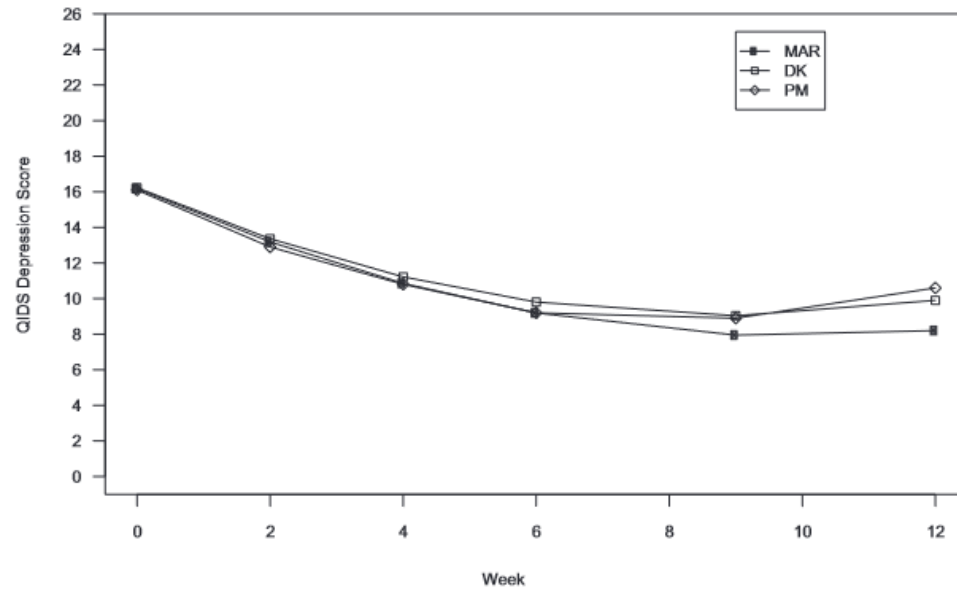


Figure 3. Depression mean curves estimated under missing-at-random (MAR), pattern-mixture (PM), and Diggle-Kenward (DK) selection modeling. QIDS = Quick Inventory of Depressive Symptoms–Clinician Rated.

GROWTH MODELING WITH NONIGNORABLE DROPOUT

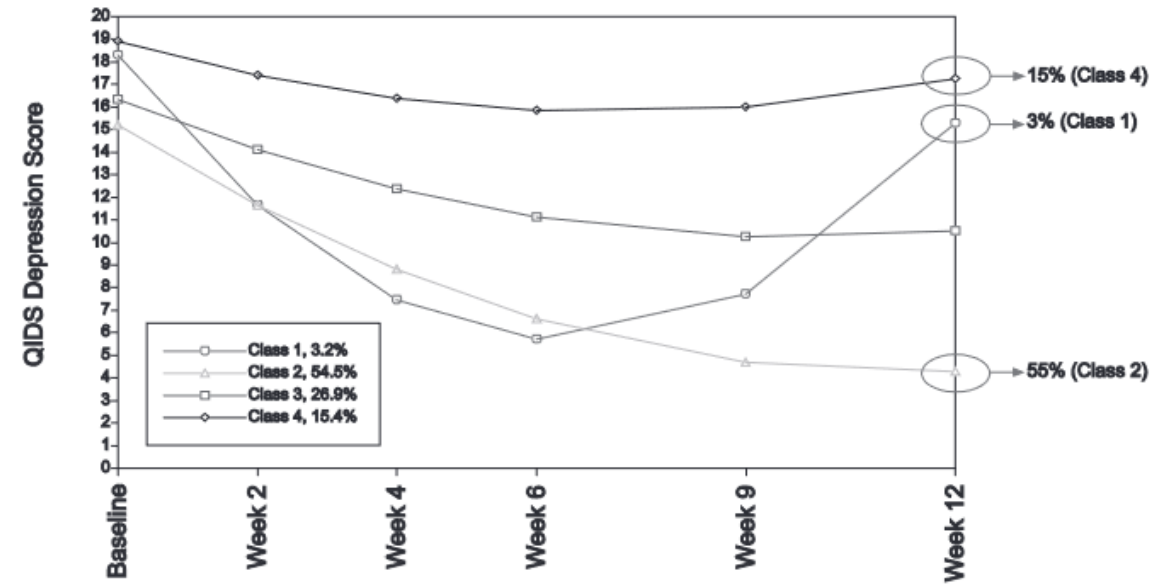


Figure 5. Four-class growth mixture model estimated under missing at random. QIDS = Quick Inventory of Depressive Symptoms–Clinician Rated.

ENOE y Pobreza laboral

- Error 1: Valores perdidos en los ingresos
 - Missing Completely At Random, Missing Not Random (sistemático)
- Error 2: Medición de pobreza laboral
 - Error aleatorio
 - Error sistemático: falsos positivos y negativos

ENOE y pobreza laboral

Si mi objetivo es estimar la pertenencia a grupos ¿Qué errores son decisivos?

Paso 1:

- Error aleatorio: Imputación
- Error sistemático: Modelación del proceso de pérdida
- Recalculo los estatus de pobreza

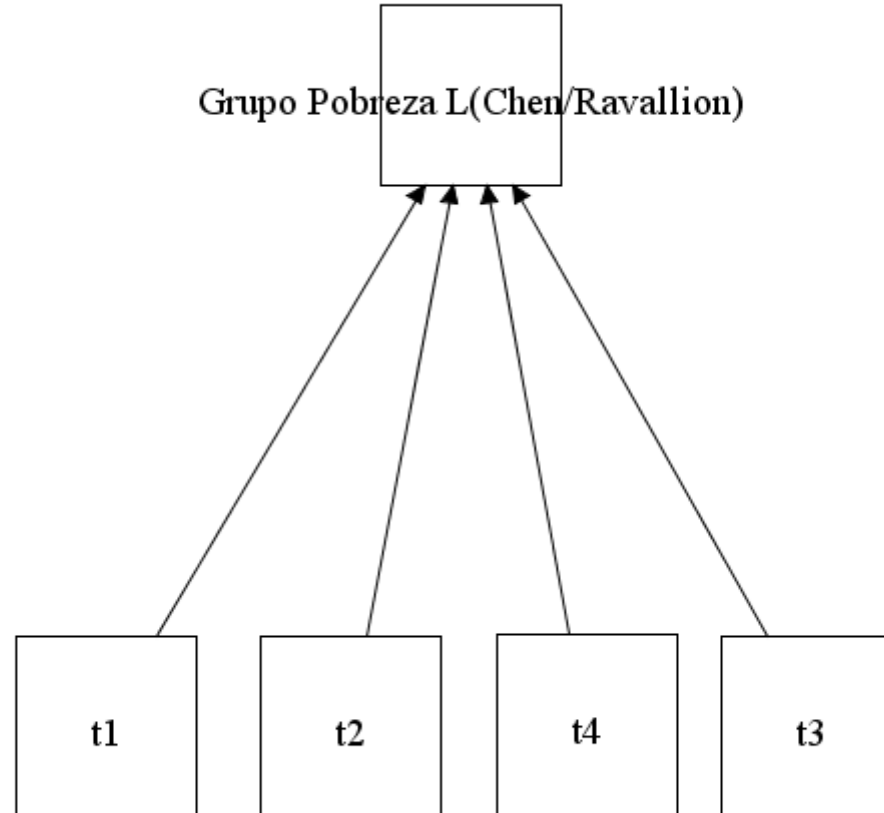
Paso 2:

- Error de clasificación: Estimo las tasas de FP y FN del estatus pobre

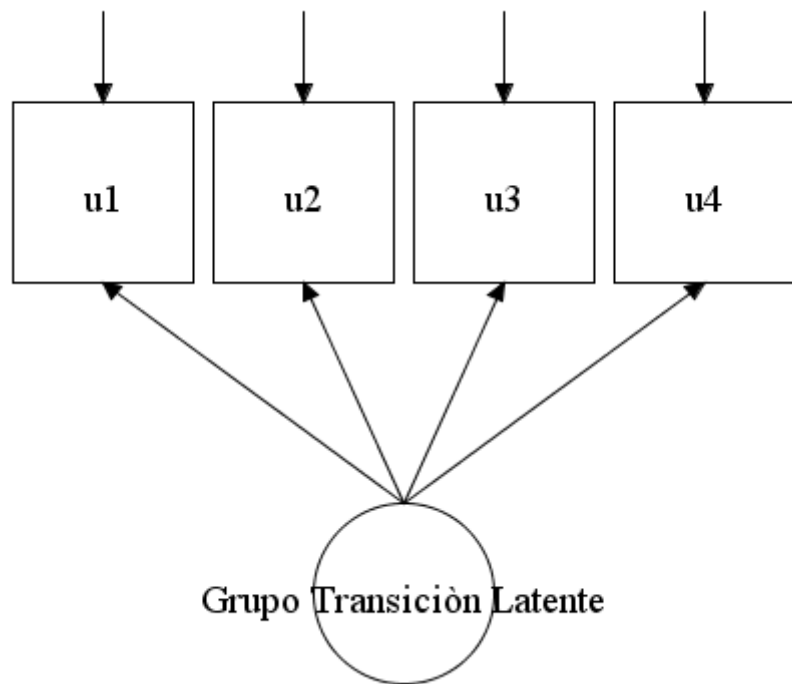
Paso 3:

- Estimo mi modelo de datos panel con efectos fijos o aleatorios

Modelo inicial



Modelo inicial

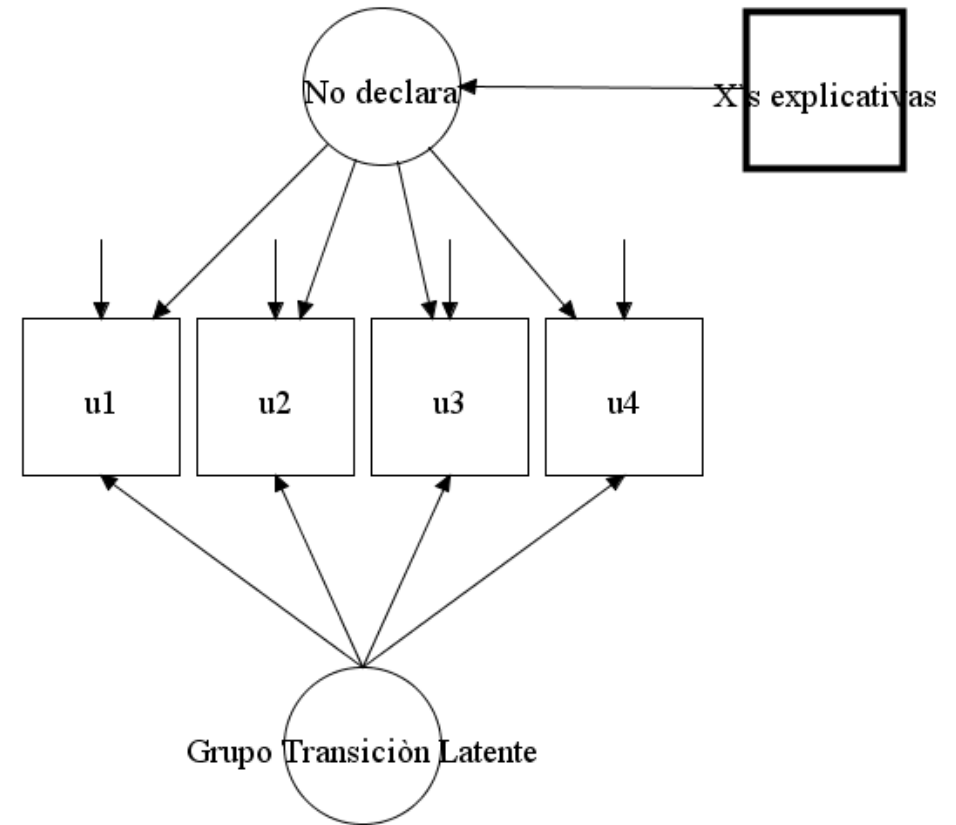
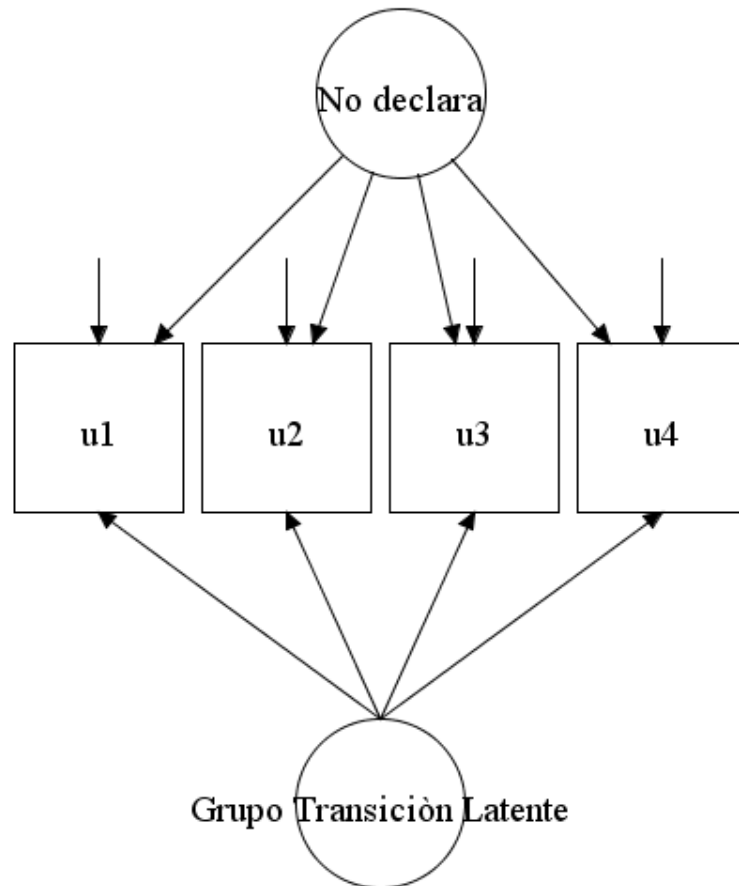


Se pobre crónico causa
observar: 1, 1, 1, 1

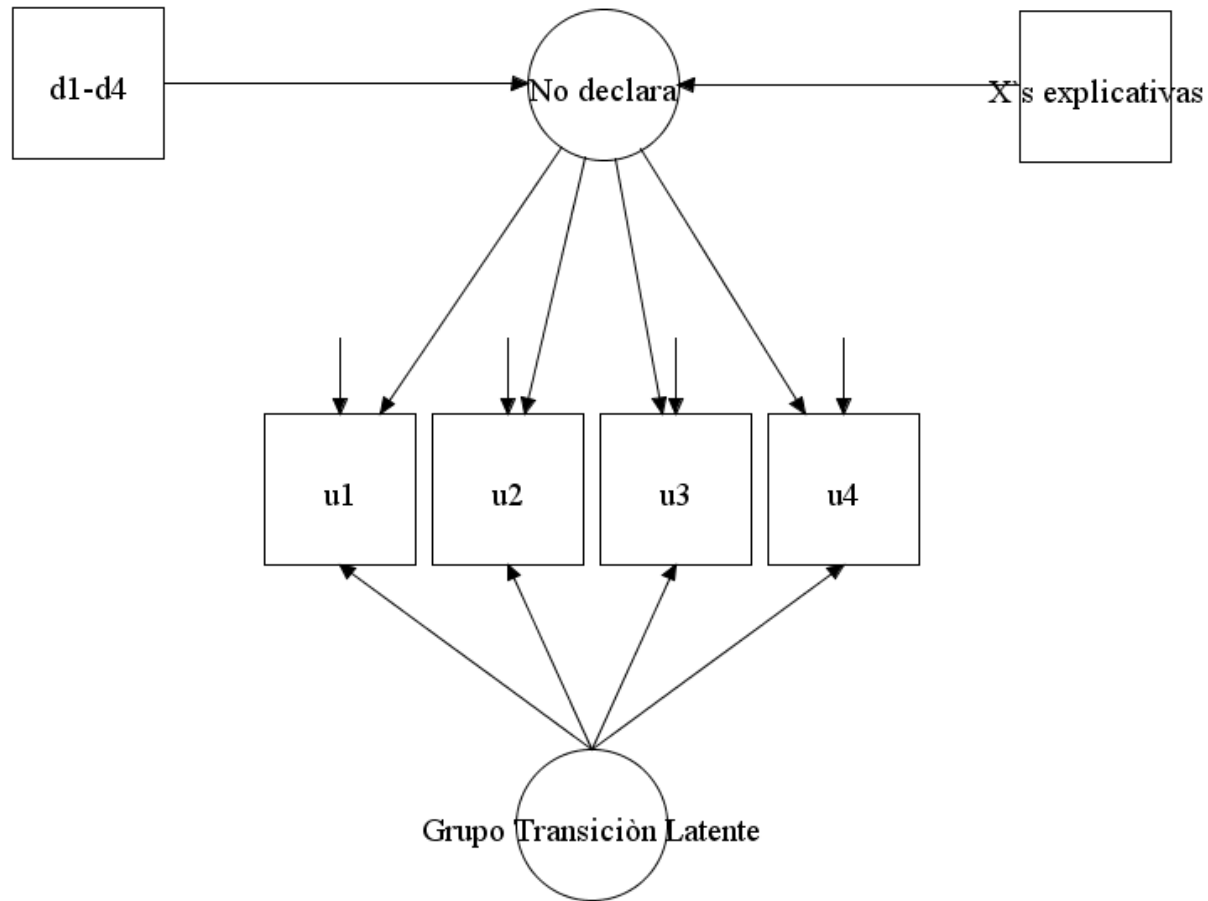
Se pobre consistente causa
observar: 0, 1, 0, 1

El error de
clasificación
ahora es
condicional
al modelo

Modelo ajustado por subdeclaración



Modelo ajustado por subdeclaración



$d1-d4$: la deserción del panel

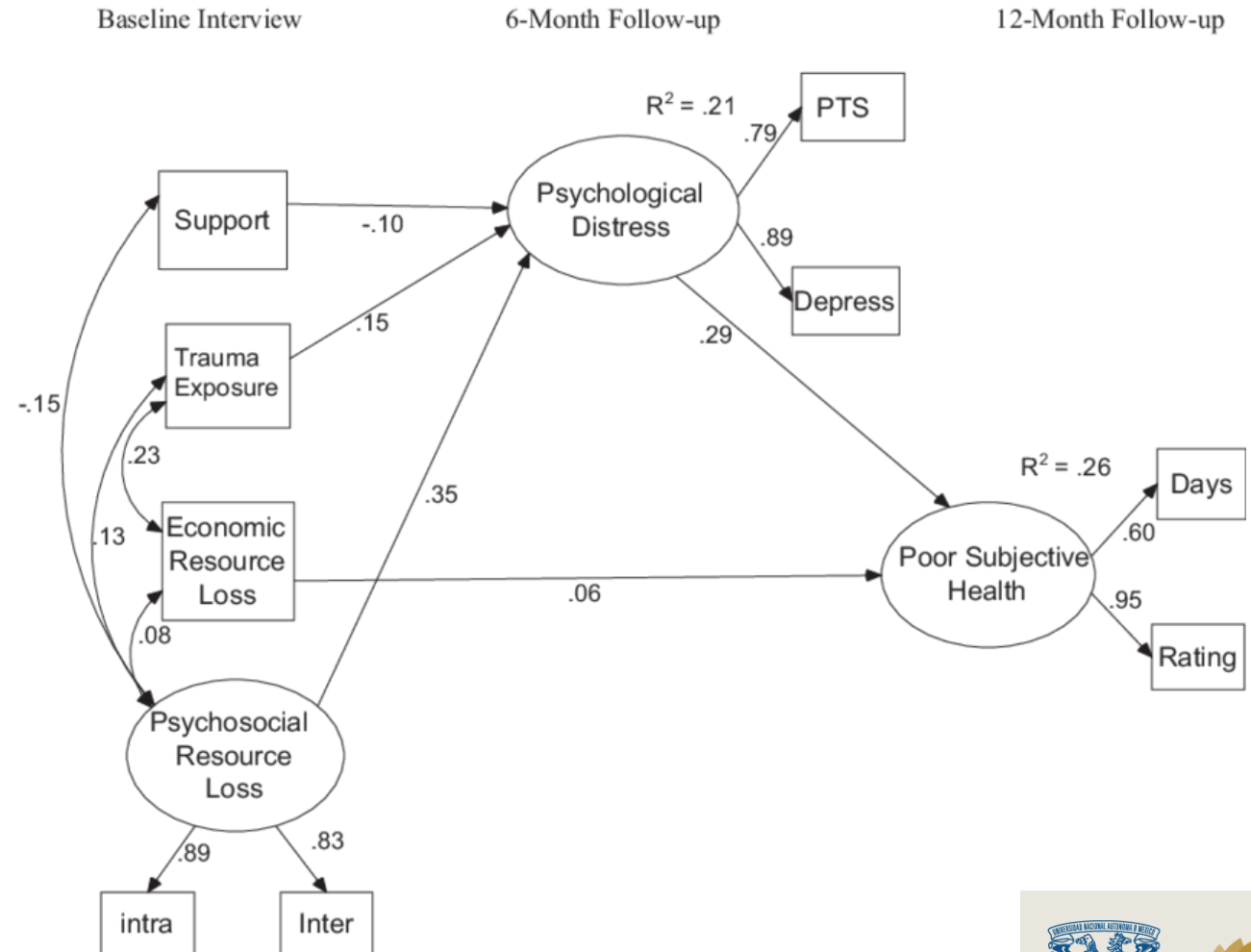
Parte 5

Fronteras en SEM I

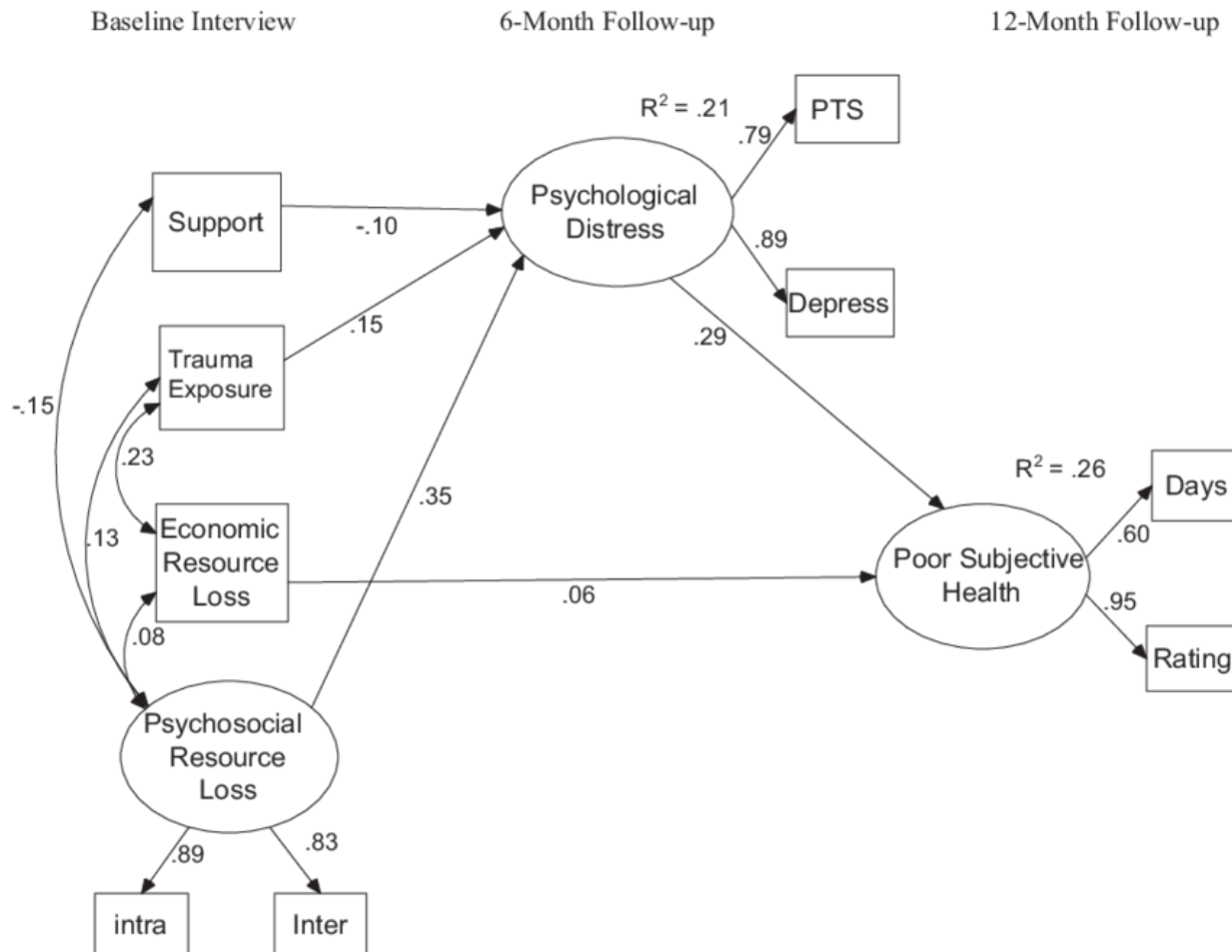


SEM y pruebas de hipótesis

- Los experimentos para poner bajo prueba una hipótesis se han vuelto más complejos
- La econometría clásica es generalmente insuficiente:
 - $Y = \alpha + \beta X_1 + e$
 - $P_i = \alpha + \beta X_1$
 - $P_i = \frac{1}{1 + e^{-\alpha + \beta X_1}}$
- Estamos interesados en relaciones y efectos múltiples entre distintas variables
- Prácticamente cualquier modelo econométrico de libro de texto podemos plantearlo bajo SEM
- ¿Cuándo sí y cuándo no?



SEM y pruebas de hipótesis



Estadística aplicada: El arte de hacer afirmaciones de conocimiento bajo incertidumbre

$$\frac{B}{SE} = \text{Descubrimiento}$$

¿Es la hipótesis nula la forma de hacer descubrimiento?

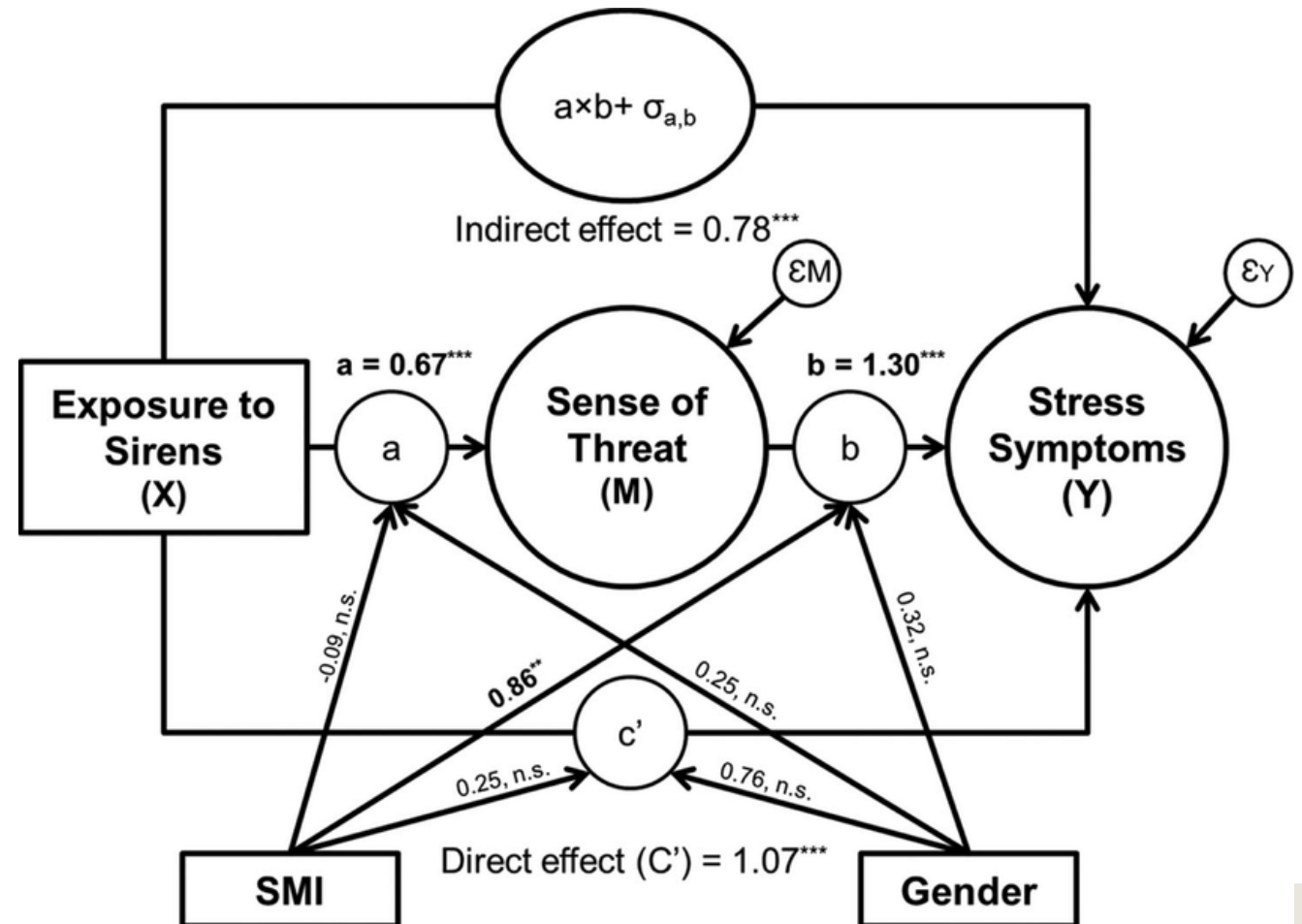
$$p < .05$$

Si es baja la probabilidad de que el valor de mi beta sea aleatorio, entonces afirmo que el efecto es de dicha magnitud?

SEM. Path Analysis

En ocasiones no queremos estimar el efecto directo sino el efecto “mediado” por otra variable

Efectos parciales

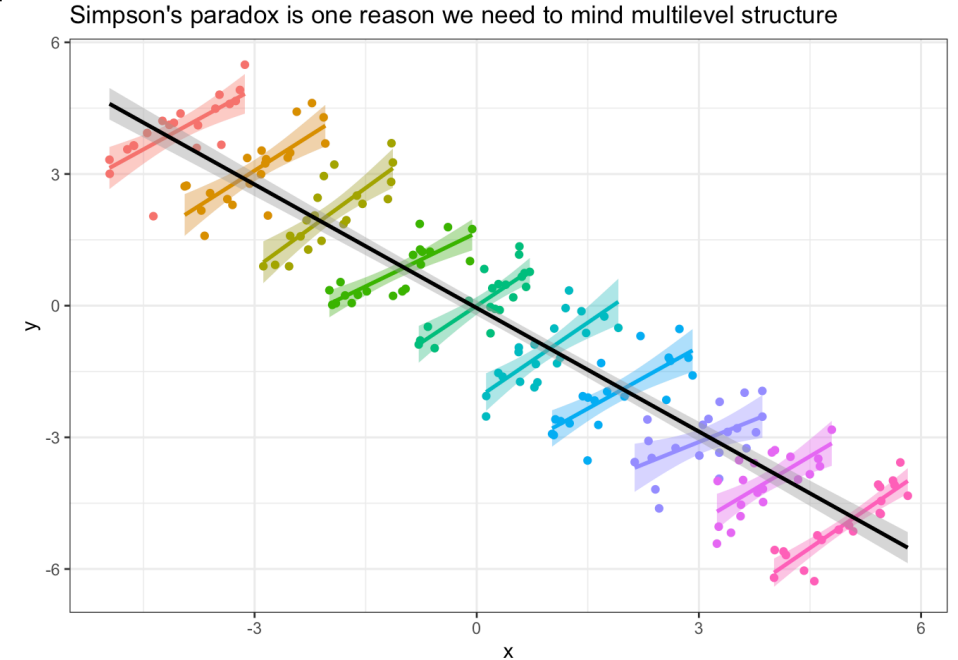
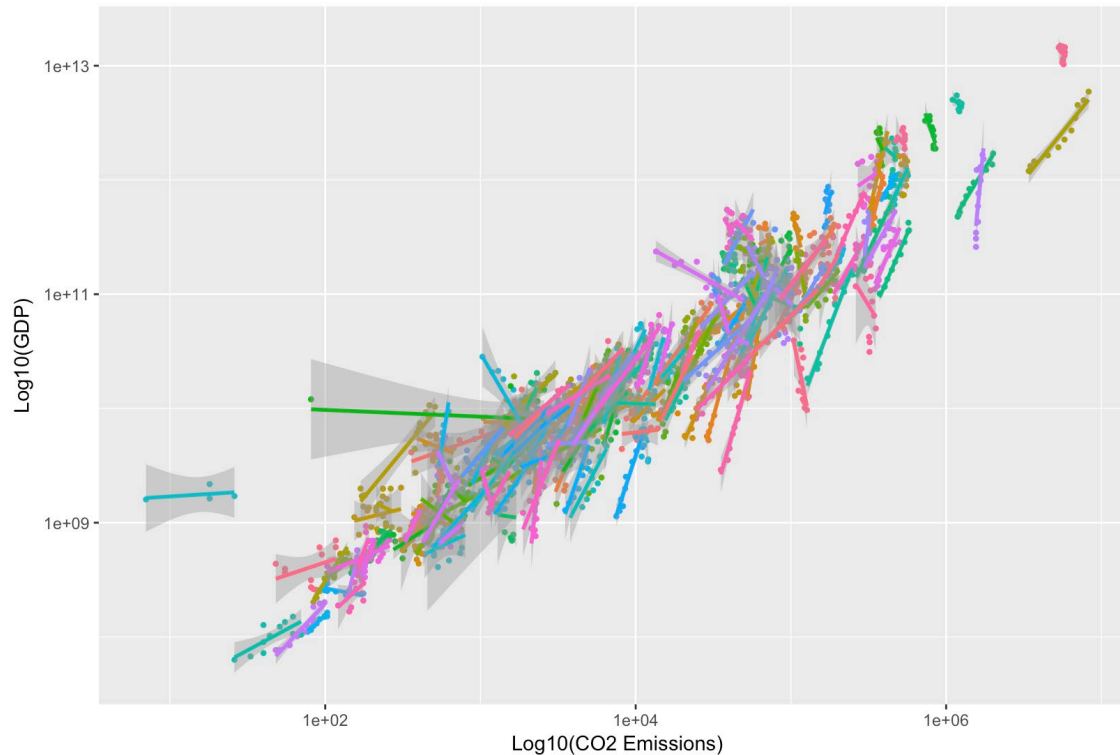


SEM Multinivel



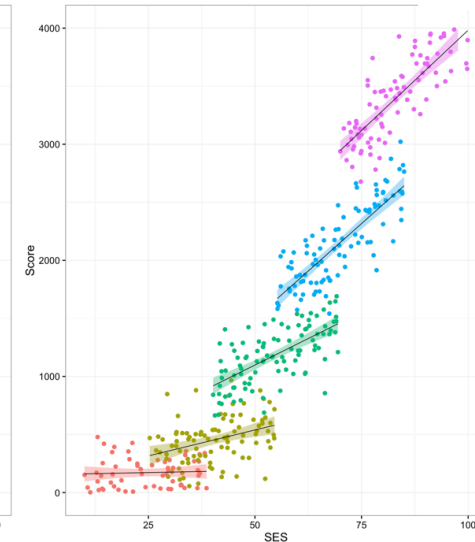
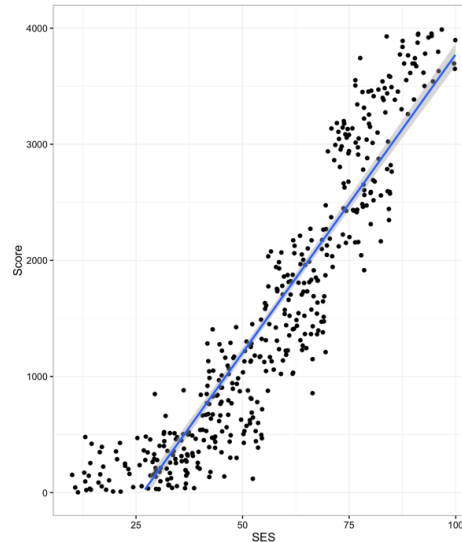
SEM, modelación jerárquica

- Hoy en día la modelación es jerárquica:
 - Estructura de datos:

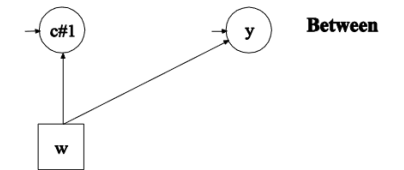
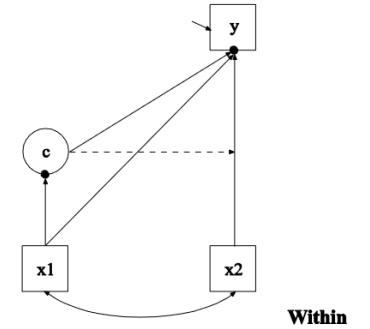


SEM, modelación jerárquica

- Ejemplos:
 - Dos mediciones en dos o más puntos en el tiempo
 - Datos panel
 - Mediciones para distintos grupos de población
 - Efectos de algún programa en distintos grupos de población
 - Efectos macro con variaciones regionales
 - Efectos macro con variaciones según su interacción: Inversión por sectores



Examples: Multilevel Mixture Modeling



SEM y vectores autoregresivos

> [Multivariate Behav Res.](#) 2000 Jan 1;35(1):51-88. doi: 10.1207/S15327906MBR3501_3.

A Structural Modeling Approach to a Multilevel Random Coefficients Model

M J Rovine, P C Molenaar

PMID: 26777231 DOI: [10.1207/S15327906MBR3501_3](#)

> [Multivariate Behav Res.](#) 2007 Jan-Mar;42(1):67-101. doi: 10.1080/00273170701340953.

Structural Equation Modeling of Multivariate Time Series

Stephen H C du Toit ¹, Michael W Browne ²

Affiliations + expand

PMID: 26821077 DOI: [10.1080/00273170701340953](#)

Residual Structural Equation Models

Tihomir Asparouhov & Bengt Muthén

To cite this article: Tihomir Asparouhov & Bengt Muthén (2022): Residual Structural Equation Models, Structural Equation Modeling: A Multidisciplinary Journal, DOI: [10.1080/10705511.2022.2074422](#)

To link to this article: <https://doi.org/10.1080/10705511.2022.2074422>

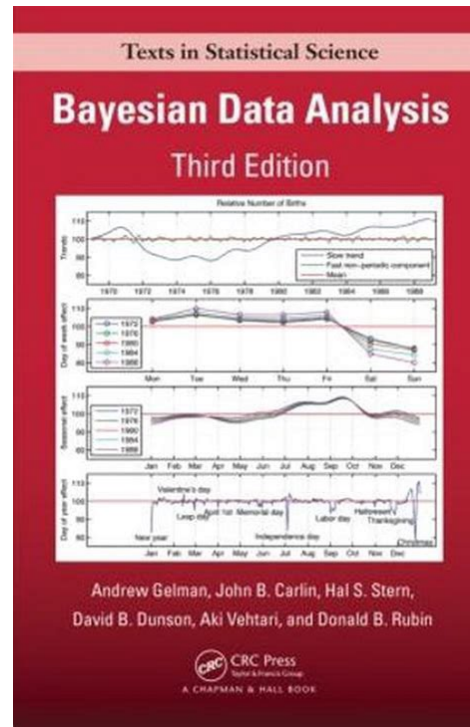
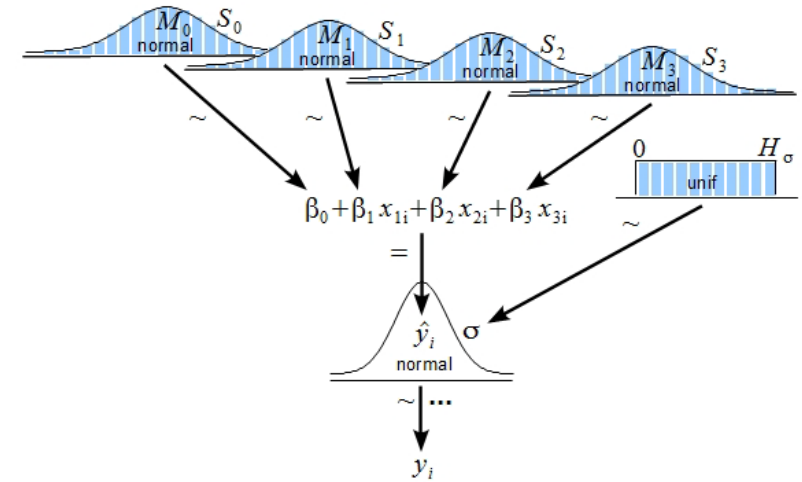


SEM, modelación jerárquica

Los modelos SEM son generativos

Los modelos Bayesianos son generativos

A más parámetros más dimensiones y más limitaciones de los algoritmos tradicionales



SEM: Monte carlo

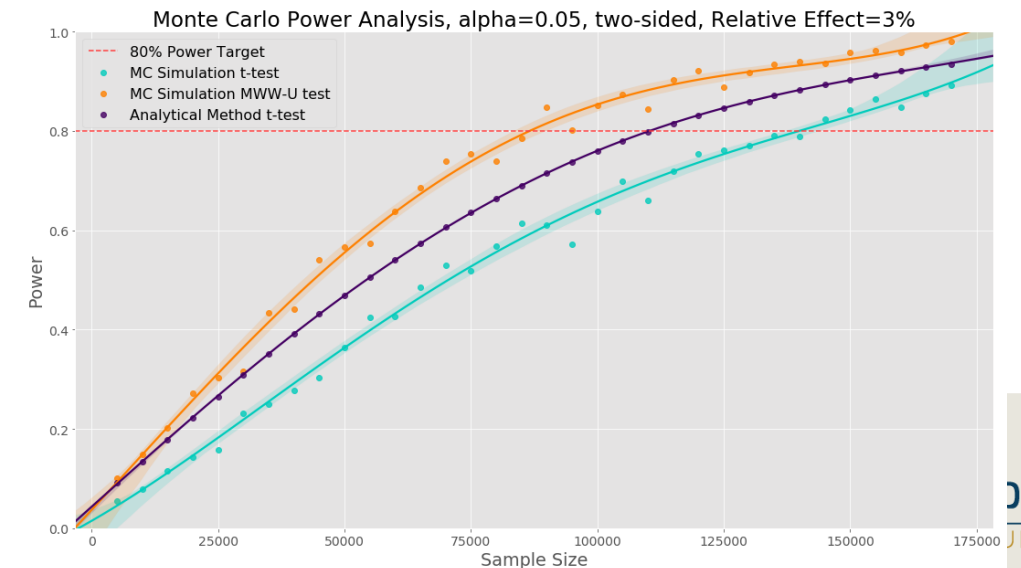
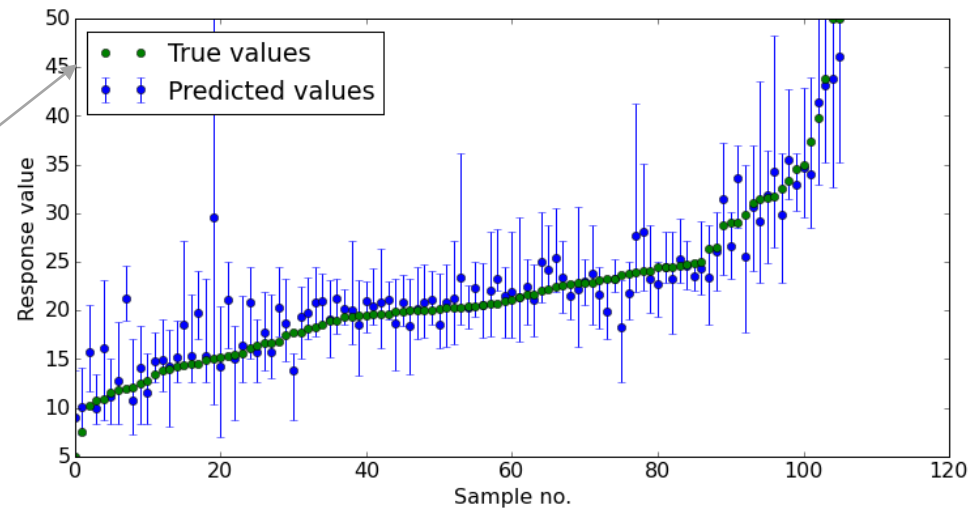


SEM y simulación (SEM y estudios de monte carlo)

Mecanismo generador de datos

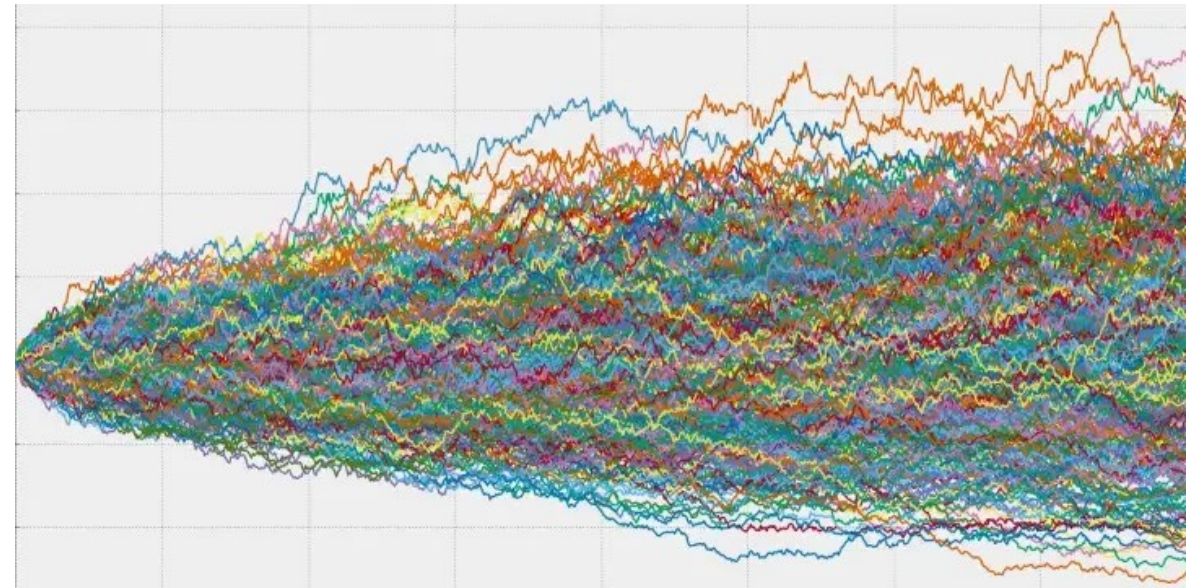
Los modelos SEM son generativos

Los modelos bayesianos son generativos




SEM y simulación

- Incertidumbre respecto al cumplimiento de los supuestos de un modelo o respecto a la capacidad del modelo para responder lo que quiero contestar
 - Datos perdidos
 - Transformación de variables
 - Sub-identificación
 - Modelo complejo para rastrear efectos puntuales
 - Tipo de distribución de la dependiente
 - Comportamiento de los errores
 - Diferencias entre algoritmos
 - Diferencias de especificación



SEM, simulación y software



HOME ORDER CONTACT US LOGIN MPLUS DISCUSSION


Chapter 12: Monte Carlo Simulation Studies

Download all Chapter 12 examples

Example	View output	Download input	Download data
12.1: Monte Carlo simulation study for a CFA with covariates (MIMIC) with continuous factor indicators and patterns of missing data	ex12.1	ex12.1.inp	none
12.2: Monte Carlo simulation study for a linear growth model for a continuous outcome with missing data where attrition is predicted by time-invariant covariates (MAR)	ex12.2	ex12.2.inp	none
12.3: Monte Carlo simulation study for a growth mixture model with two classes and a misspecified model	ex12.3	ex12.3.inp	none
12.4: Monte Carlo simulation study for a two-level growth model for a continuous outcome (three-level analysis)	ex12.4	ex12.4.inp	none
12.5: Monte Carlo simulation study for an exploratory factor analysis with continuous factor indicators	ex12.5	ex12.5.inp	none
12.6 Step 1: Monte Carlo simulation study where clustered data for a two-level growth model for a continuous outcome (three-level analysis) are generated, analyzed, and saved	ex12.6step1	ex12.6step1.inp	none
12.6 Step 2: External Monte Carlo analysis of clustered data generated for a two-level growth model for a continuous outcome using TYPE=COMPLEX for a single-level growth model	ex12.6step2	ex12.6step2.inp	ex12.6replist.dat
12.7 Step 1: Real data analysis of a CFA with covariates (MIMIC) for continuous factor indicators where the parameter estimates are saved for use in a Monte Carlo simulation study	ex12.7step1	ex12.7step1.inp	ex12.7real.dat
12.7 Step 2: Monte Carlo simulation study where parameter estimates saved from a real data analysis are used for population parameter values for data generation and coverage	ex12.7step2	ex12.7step2.inp	ex12.7estimates.dat
12.8: Monte Carlo simulation study for discrete-time survival analysis	ex12.8	ex12.8.inp	none

rstan 2.21.2 Vignettes Functions Other Packages ▾ Stan

Simulation Based Calibration (sbc)

 Check whether a model is well-calibrated with respect to the prior distribution and hence possibly amenable to obtaining a posterior distribution conditional on observed data.

```
sbc(stanmodel, data, M, ..., save_progress, load_incomplete=FALSE)
# S3 method for sbc
plot(x, thin = 3, ...)
# S3 method for sbc
print(x, ...)
```

