Day 1: of the UNL **SSP** Missing Data Workshop Series: *Introduction to Missing Data*

by Levente (Levi) Littvay
Central European University
Department of Political Sciece
levente@littvay.hu

What is item nonresponse?

- Unit Nonresponse vs. Item Nonresponse

ID	Q1	Q2	Q3
456	1	1	2
457	4	2	1
458	?	?	?
459	3	2	1

ID	Q1	Q2	Q3
456	1	1	2
457	4	?	1
458	?	2	1
459	3	2	?

Unit Nonresponse Examples

- Person who is not at home
- Person who does not pick up the phone
- Person who hangs up on you
- Rat that dies before the study
- The country you could not get data on
- etc.

Item Nonresponse

- “I Don’t Know”
- Refusals to respond
- Questions left blank
- Failed measurement
- etc.

Best way to deal with Missing Data:

Don’t have any!!!

Minimizing Unit Nonresponse

- Call back if not home.
- Refusal conversion
- Don’t mess up
- Clear and understandable questionnaire
- Polite request
- Incentives

Minimizing Item Nonresponse

- Well written questions
- Minimize misunderstandings
  - cross-cultural example
  - Standardized vs. non-standardized
- Minimize skip patterns
- Use decent measurement instruments
- Know what you are doing

What kind of Missing Data Should be Modeled?

- If an item is missing from your dataset but you suspect that it has a true value.
- I don’t know might simply mean I don’t know
  - Don’t model it as if there was a true value
  - Model it as censor-inflated (for example)
- Attrition: Do you want to model dead people?
  - Well being of people who died
  - Consumer behavior of people who died

Distribution of Missing Data (Missingness)

ID	Q1	Q2	Q3
456	1	1	2
457	4	?	1
458	?	2	1
459	3	2	?

ID	Q1	Q2	Q3
456	0	0	0
457	0	1	0
458	1	0	0
459	0	0	1

Mechanism: Why is it missing?

- **MCAR** - Missing Completely at Random
  - Very rarely can you assume this
  - Testable
- **MAR** - Missing at Random
  - Variables in your model predict missingness
  - Theoretically you can often assume MAR
  - You can include (auxiliary) variables that will satisfy this
  - Not testable
- **NMAR** - Not Missing at Random (**MNAR**)
  - Missingness depends only on its own value
  - If you have this, you usually have a problem

Classical Missing Data Treatments

- Whatever you do, you are doing something
  - Case Deletion
    - Listwise
    - Pairwise
    - Multi-Item: Standardize and average available info
  - Single Imputation
    - (Unconditional) Mean Imputation
    - Conditional Mean Imputation (expected value)
    - Unconditional Distribution Imputation (Hot Deck)
    - Conditional Distribution Imputation
  - Reweighting - Can Perform Well
    - Has received more attention lately

Listwise Deletion and Multi-Item

- Exclude the whole case
- Default in most software
- Works if mechanism is MCAR
  and if pattern and sample size allows.
  (Need to have enough complete cases.)

- Item standardization & averaging
  - Works if MCAR (hard to justify)
  - If items measure construct without error
  - Better off with an IRT model

Pairwise Deletion

- An option for correlation/covariance matrixes.
  - Calculates each correlation in the matrix based on available information.
  - Each cell of the correlation matrix is calculated from a different sample (with different sample size).
  - Threat of sample selection error in each cell.
  - Very unpredictable bias if you estimate off this covariance matrix. (Structural Equation Model)

Pairwise Deletion

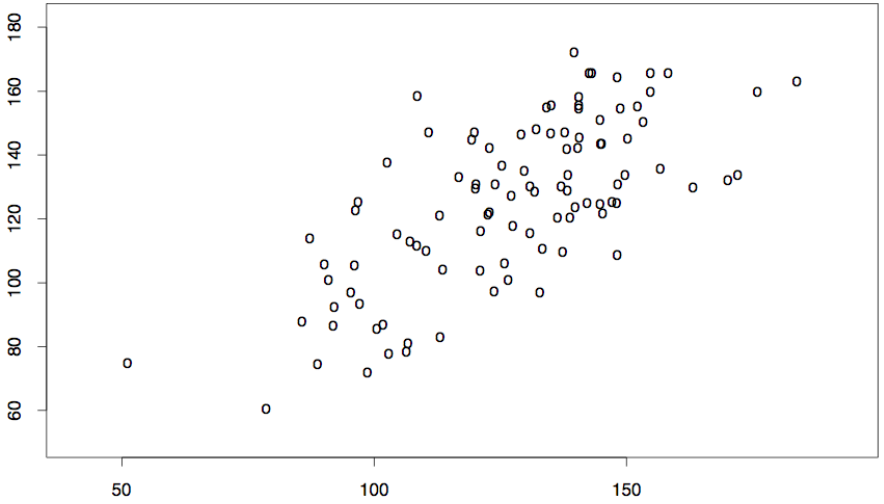
- An option for correlation/covariance matrixes.
  - Calculates each correlation in the matrix based on available information.
  - Each cell of the correlation matrix is calculated from a different sample (with different sample size).
  - Threat of sample selection error in each cell.
  - Very unpredictable bias if you estimate off this covariance matrix. (Structural Equation Model)

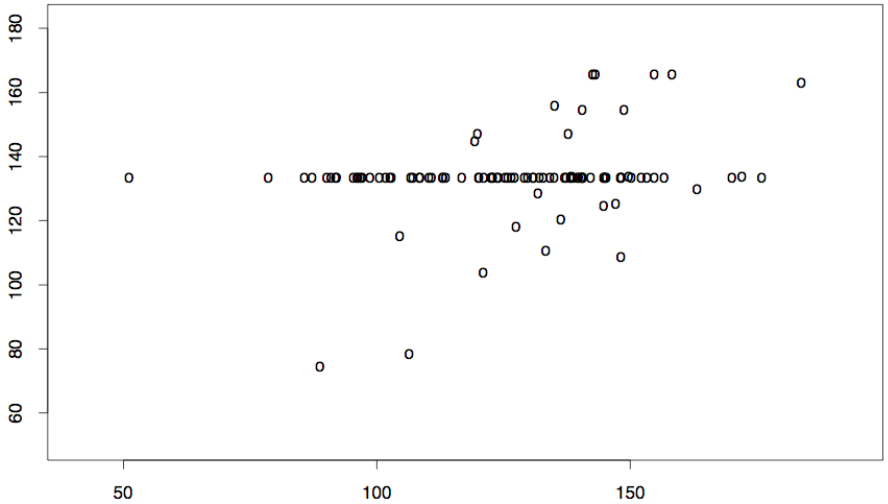
Mean Imputation (Very Bad)

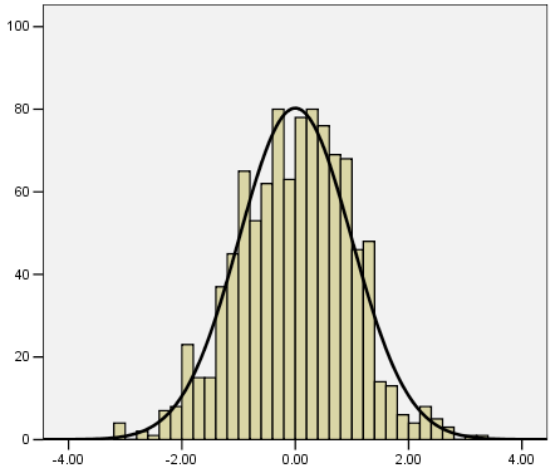
- Fallacy of Single Imputation Methods
- Properties of a Good Estimator

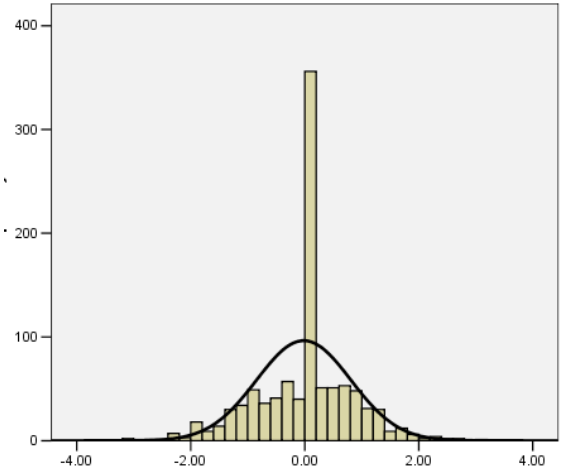
- Unconditional Mean Imputation: Biased
- Messes with parametric properties of data. (Leptokurtosis)
- Decreased standard errors.
- Overconfident Estimates
- Increased Probability of Type I error.

(Unconditional) Mean Imputation





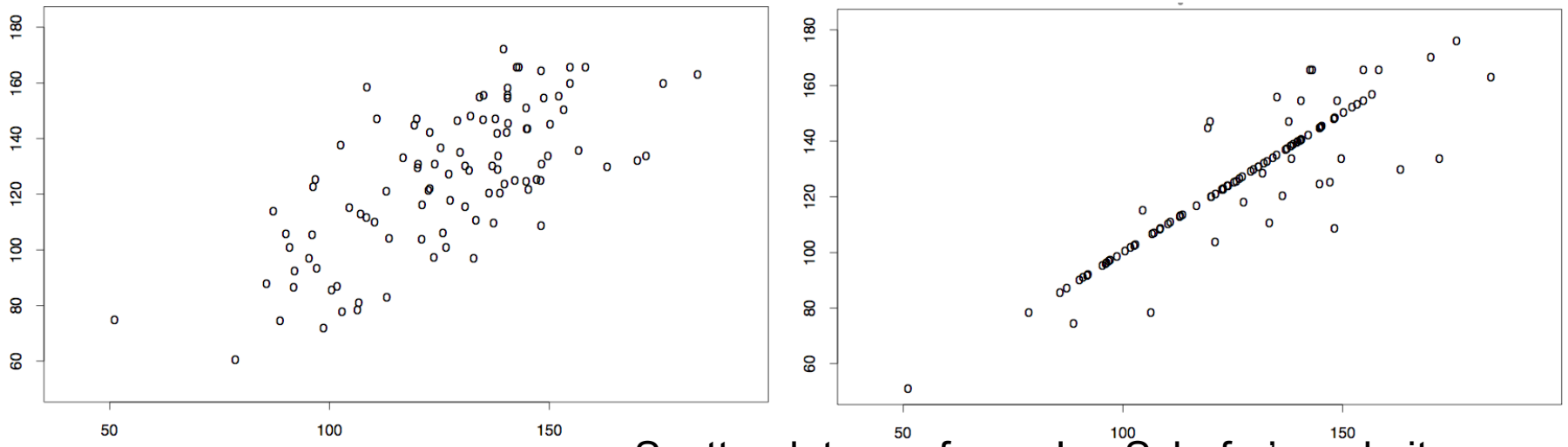




Scatterplots are from Joe Schafer's website

Imputation of Expected Value

- Usually uses the EM Algorithm
- SPSS Missing Value Analysis (SPSS MVA) $599.00
- Good for creating expected values
- Bad for multivariate analysis
  - Decreases standard errors
  - Creates overconfident outcomes
  - Increases probability of Type I error



Scatterplots are from Joe Schafer's website

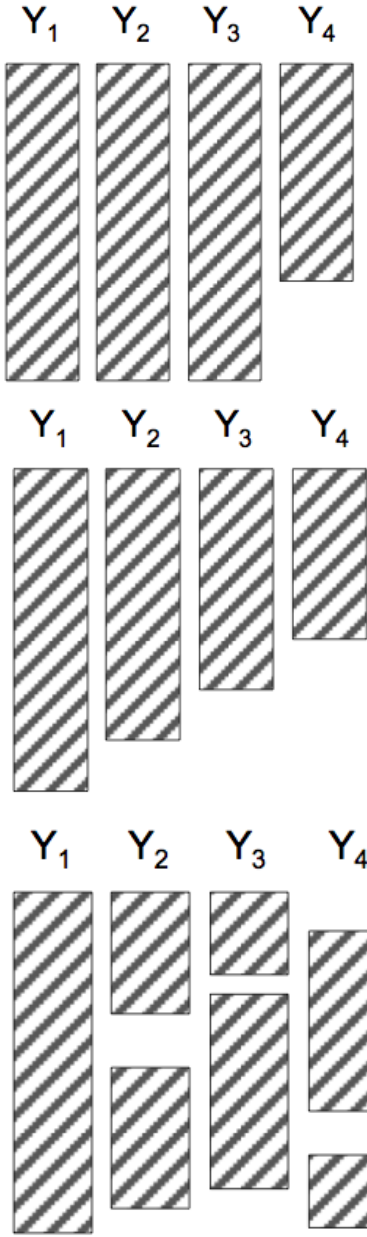
Distribution Imputation

- Unconditional Distribution Imputation (Hot Deck)
  - Non-Parametric
  - Preserves Distribution
  - Distorts Relationships

- Conditional Distribution Imputation (Cold Deck)
  - Nonparametric
  - Have to find the appropriate conditional model
    - Balance between modeling uncertainty and expectation
  - Works for univariate pattern
  - Near impossible to calculate for multivariate pattern

Missing Data Patterns

- Univariate
  - MCAR is easily testable
- Monotone
- Arbitrary
  - How can you satisfy MAR?



Why are patterns important?

- From a computational perspective
  - Who cares?

- Can offer obvious solutions
  - Example
  - Horizontal (Listwise) vs. Vertical Deletion
  - Think about your theory!

Increasing the Confusion: Models

- Analysis Model
- Missing Data Model
  - Model to predict missingness
    (To satisfy the MAR asumption)
  - Model to predict the values of missing data
    (To alleviate NMAR bias)
- Auxiliary Variables
- Specification of a Missing Data Model
  - Include all analysis variables
  - Include all other available variables
  - Cannot really overspecify (though it might not converge)
  - Structured vs Unstructured Model

Advanced Missing Data Methods

- Multiple Imputation

- Direct Estimation
  - Full Information Maximum Likelihood
  - Bayesian Estimation with Metropolis-Hastings or Markov Chain Monte Carlo

- NMAR Procedures (Usually uses one of these procedures or their extensions.)

Note: If you lose it from here don't feel too bad.
Tomorrow's session will probably answer a lot of your questions.

Multiple Imputation

- Models both expected value and uncertainty.
- Using the *Missing Data Model* you specify it simulates and imputes missing values “multiple” times creating M complete datasets. (M=5 is usually OK. It is a good idea to simulate more.)
- Analyze each dataset independently.
- Combines results to get unbiased estimates. Models both uncertainty and expectation

Multiple Imputation

- Pros
  - Can be used with any statistical procedure that produces parameter estimates and standard errors.
  - You can use any stats package you like.
  - More robust to parametric violations than Full Information Maximum Likelihood

Multiple Imputation

- Cons
  - Pain in the…
  - Most imputation software limits you to certain distributions (categorical, continuous, count.)
  - No unbiased way to combine goodness of fit stats
  - Empirical confidence intervals are not available
  - Model convergence is in the eye of the beholder for more complex models
  - Less efficient than FIML because of unstructured imputation model and limited number of imputations

Multiple Imputation Procedures

- Bayesain (Pros and Cons)
  - NORM for Windows
  - S-Plus 4 & R modules: NORM, PAN, CAT, MIX
  - S-Plus 6 and above: Missing Data Module
  - SAS: Proc MI
  - MlWin (multilevel)
  - AMOS (SEM and Structured Imputation)
- Bootstrap based (Pros and Cons)
  - Amelia II
- Chained Equations (Pros and Cons)
  - SAS: IVEWare
  - S-Plus module: MICE (R clone is also available)
  - Stata module: ICE
  - WinMICE (multilevel)

Combining Results, Rate of Missing Information and Efficiency

- Combining Results with Rubin's Rules:

  It Is Easy! You can do it manually with Excel.

  – Average parameter estimates $\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}_j.$

  – Simple formula to combine standard errors

Within $\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$ .etween $B = \frac{1}{m-1}\sum_{j=1}^{m} (\hat{Q}_j - \bar{Q})^2.$ $T = \bar{U} + \left(1 + \frac{1}{m}\right) B.$

- Rate of missing information df $df = (m-1)\left(1 + \frac{m\,\bar{U}}{(m+1)B}\right)^2.$

  – Simple Formula

$$\gamma = \frac{r + 2/(df+3)}{r+1} \qquad r = \frac{\left(1+m^{-1}\right)B}{\bar{U}}$$

- Efficiency

  – Simple Formula

  – Table $\left(1 + \frac{\gamma}{m}\right)^{-1}$

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Combining Results

- Software to do this:
  - **Norm for Windows** will do it (**best**)
  - Proc MIANALYZE
  - Numerous modules for R and Stata.

- Automated Support:
  - Mplus
  - HLM
  - MlWin (Fully Automated?)
  - Zeilig - R module (Fully Automated?)
  - etc.

Common Mistakes

- Confusing Missing Data and Analysis Model
- Confusion about what needs to be combined
- Rubin's Rules (Again It Is Easy!!!)
- Ignoring the Rate of Missing Information
- Conforming to Observed Data
  - In Range
  - Rounding

Direct Estimation (FIML and Bayes)

- Missing data is not imputed
- Information is borrowed from cases where the information is available
- It assumes that missing cases follow the same **multivariate** distribution as available cases. (Makes assumption MAR)
- If same **univariate** distribution was assumed you would have to assume MCAR

Full Information Maximum Likelihood

- ML is calculated for each pattern
- Contribution to the likelihood function is summed across patterns
- For MAR have to use observed (vs estimated) information matrix otherwise have to assume MCAR
- Bottom line: it works if it converges

Bayesian

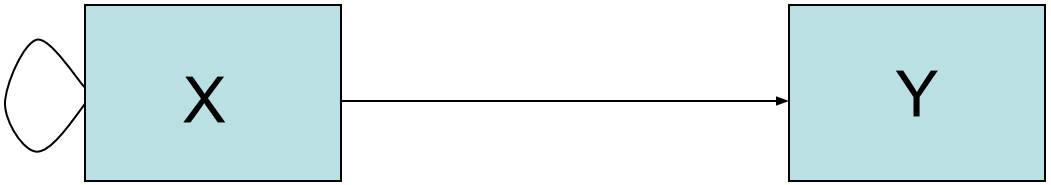
- A Bayesian procedure for estimation
- Requires prior distribution
- Performs Bayesian updating using prior distribution and distribution of available data
- Takes the solution as prior and repeats step
- After lots of iterations hopefully converges to a distribution. Solution is sampled from this.
- Solution is unbiased if mechanism was MAR

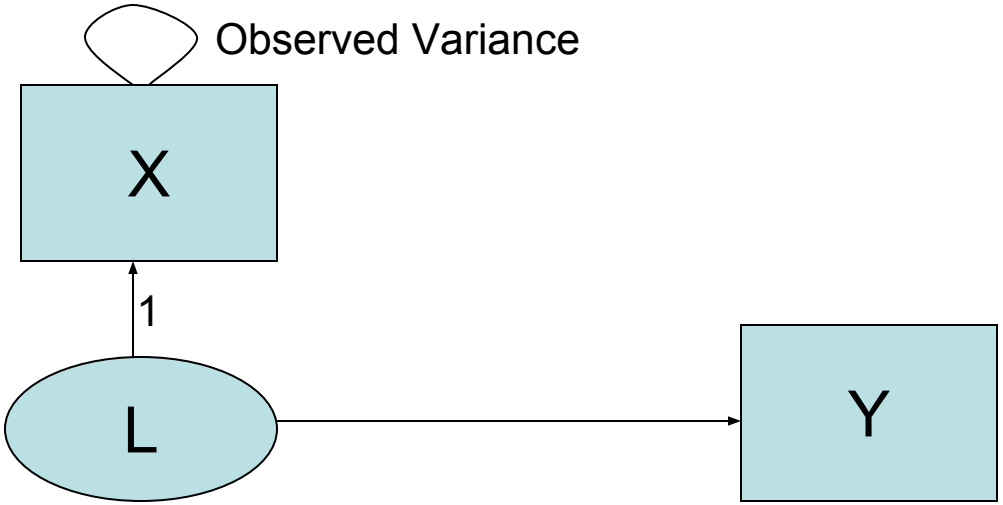
Direct Estimation Issues

- Missing Data on Dependent (or Endogenous) Variable: Not a Problem

- Missing on the Independent (or Exogenous) Variable:
  - Have to make them endogenous
  - Have to make a distributional assumption

SEM Analogy

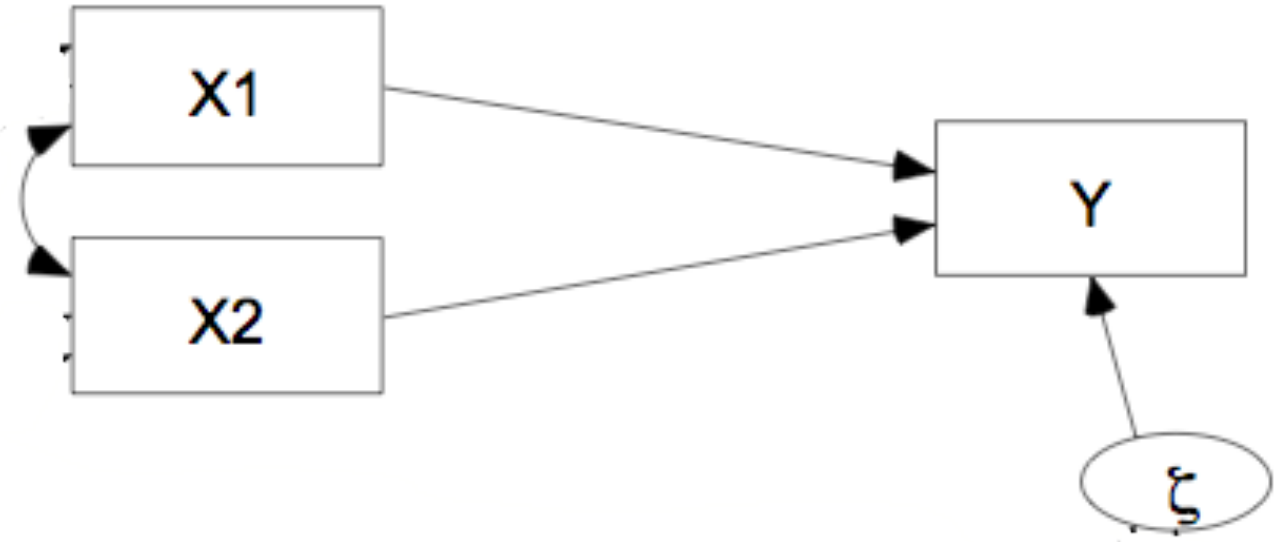




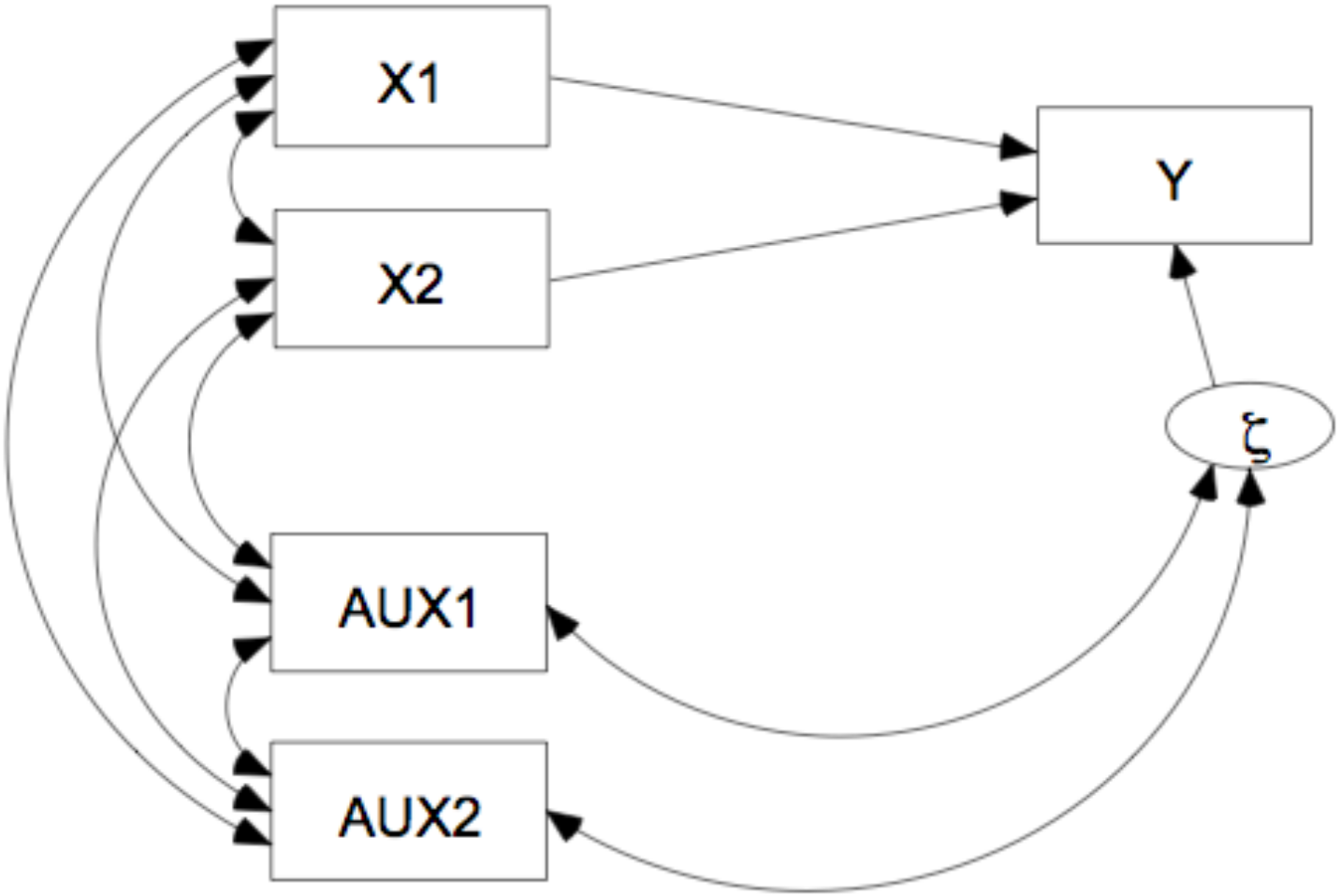
Including Auxiliary Variables

- Correlate selection/auxiliary variables with
  - All observed exogenous variables
  - All residuals of endogenous variables
- Selection/Auxiliary variables will not contribute to misfit.
- The Auxiliary/ Selection variable part of the model will be **completely saturated**.

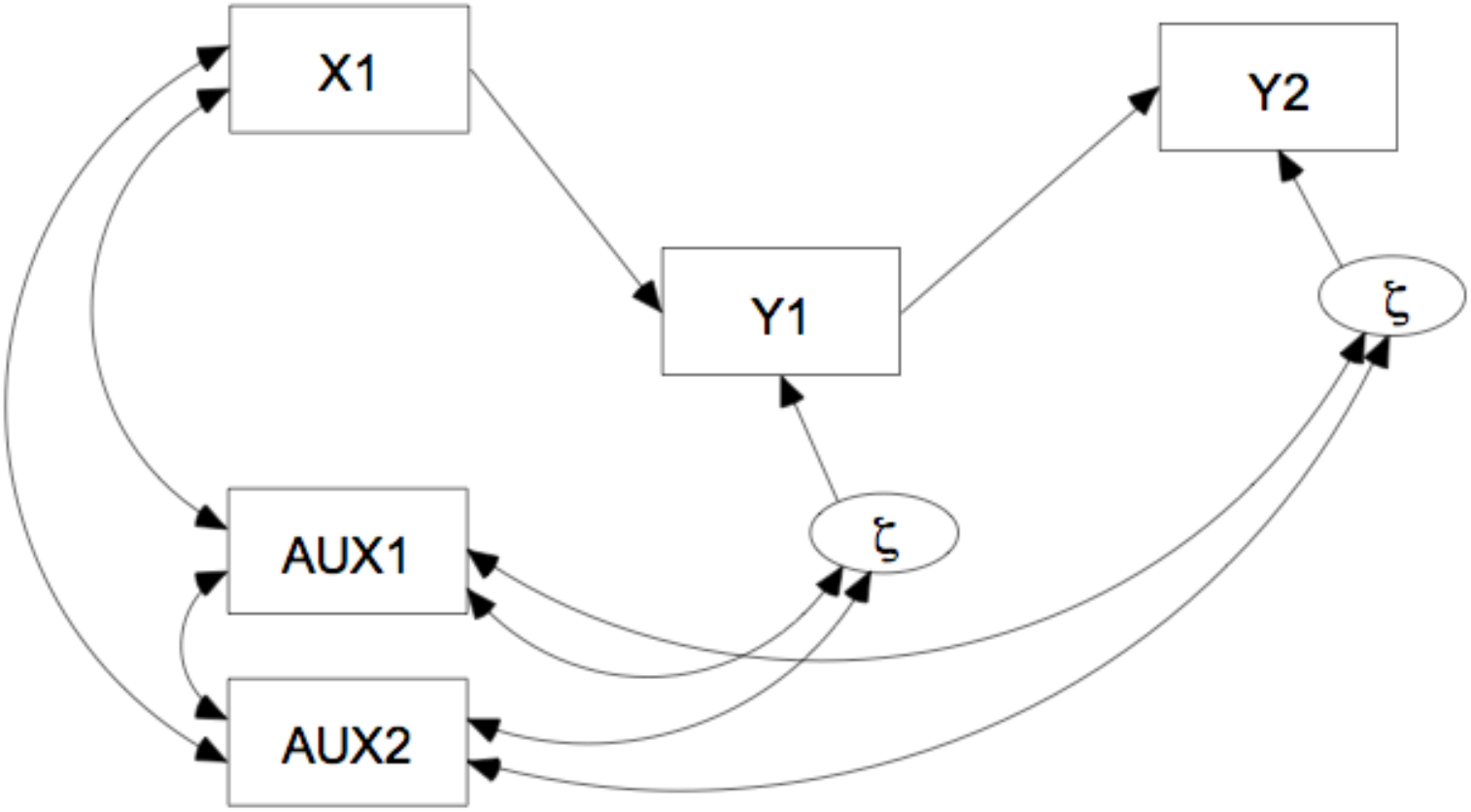
A Regression Model



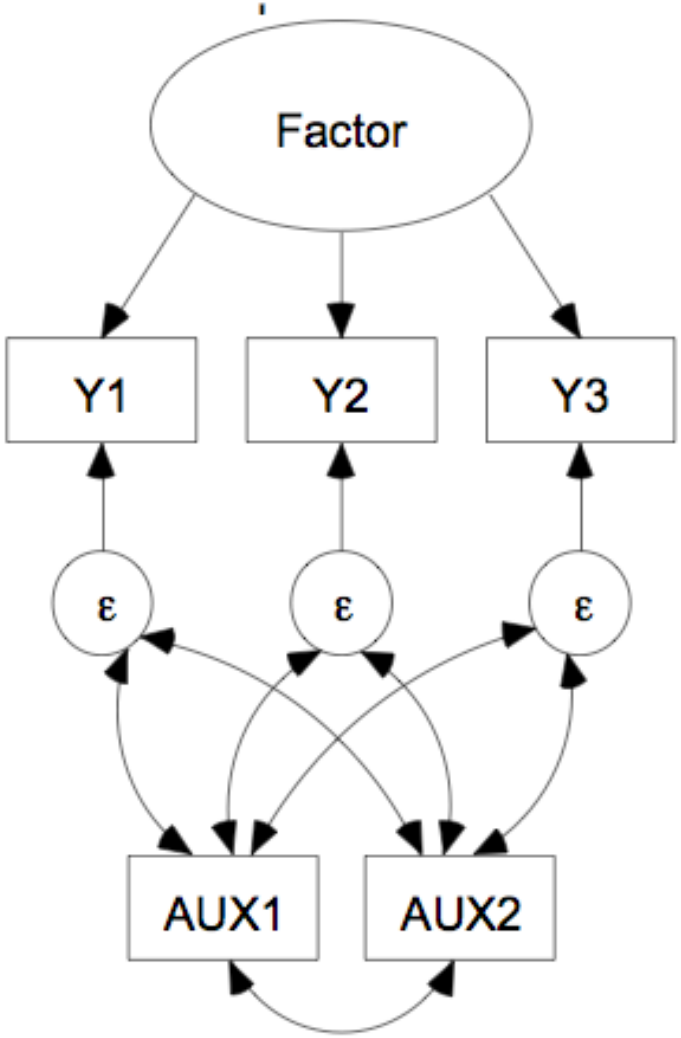
Auxiliary Variables for Regression



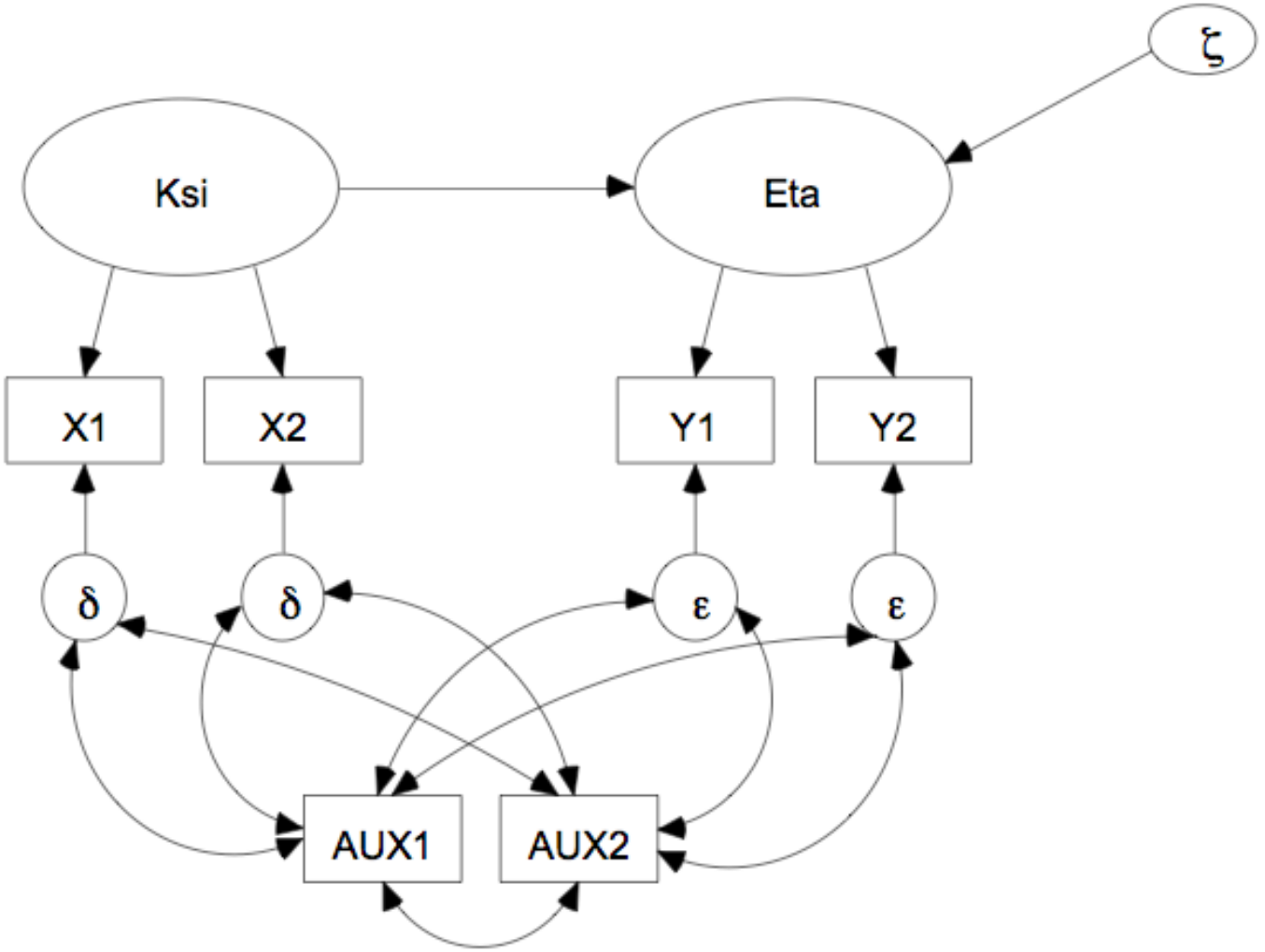
Auxiliary Variables for Path Models



Aux for Confirmatory Factor Models



Aux for Full Structural Models



NMAR Missing Data

- Three Approaches
- Auxiliary Variables
  - Can alleviate NMAR bias
    (If correlates highly with missing values)
- Pattern
  - Models missingness **and** model simultaneously
- Mixture
  - Makes certain assumptions about missing data

Auxiliary Approach

- We used auxiliary variables to satisfy MAR
- We needed the auxiliary variables to predict the missingness

but…

- If the auxiliary variables can also predict the values of missing data NMAR bias is decreased.
- NMAR bias will not be eliminated unless auxiliary variables predict perfectly

Pattern Approach

- Heckman Regression
- NMAR missing on the dependent variable
- Two models:
  - One that includes the predictors of the dependent variable
  - One that predicts missingness
- Implemented in Stata and R

Heckman Regression Estimation

- Originally estimated with two steps
- Only identified if the two models are different
- Now estimated jointly with likelihood methods
  - Difficult to estimate
  - Some areas of the likelihood function are flat
  - Quasi Likelihood Methods have been developed to overcome these difficulties
- Depending on estimation MAR missing on independent variables could be OK

What is of interest

- Of interest: correlation between the residuals of the two jointly estimated models.
- Controlling for all predictors if the leftover variance of the dependent variable is correlated with the leftover variance of missingness that is by definition NMAR
- This approach controls for this correlation

Mixture approach

- When we expect NMAR missing data we are dealing with a mixture of two distributions
  - One we observe
  - The other we do not
- Based on the observed we can make assumptions about the missing one
- Results depend heavily on the assumptions

Mixture approach example

- People who do not report their weight are 10 lbs. heavier than those who report.
- Mean is different
- Distribution is the same
- Problems
  - You need to make untestable assumptions.
  - You need to be a good enough expert to make such judgment calls
- One Solution: extensive sensitivity analysis
  - What if it is 15 lbs.? What if it is 20?

Mixture Distribution Implementation

- Multiple Imputation
  - You have an observed set of data
  - You have an imputed set of data
  - Make necessary transformations to imputed set of data
  - The rest is the same as MAR multiple Imputation
- Can be done with likelihood methods but there are no pre-packaged solutions available

Tie It Back to Design

- When you design your study try to avoid missing data
- When you cannot avoid it anticipate it
- Collect data that predict missingness
- Collect data that correlate with variables you expect NMAR missing data on.

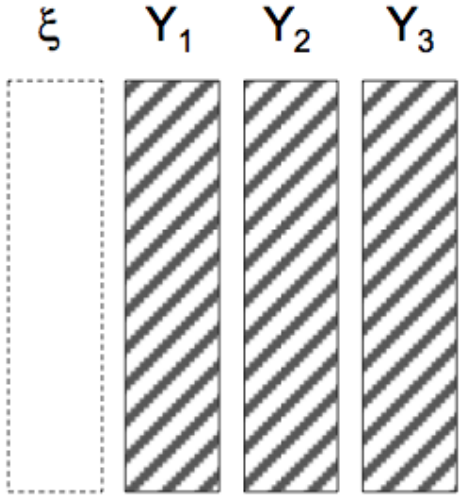
Examples

- Income
  - What kind of car you drive
  - This year how much money you plan to spend on vacations
- Attrition
  - On a scale of 1-7 how likely are you to drop out of this study if…
    - The drugs work
    - The drugs do not work
    - If you are too busy around the time of your next examination

Statistical Issues as Missing Data Problems

Latent Variable Models

- Most Simple Missing Data Problem
- Pattern



- Observed values predict missing values
- Solutions are also somewhat similar

Missing by Design

- Why put missing data in your study?
  - Fatigue
  - Cost (ie. respondents are paid by the question)
  - Context Effect
- How?
  - Randomize what will be missing for everyone
  - This creates MCAR Missing Data
  - Analyze results with discussed methods

Economic Incentives for Cognitive Tasks

playful

comforting

irritated

bored

Motivation and Design

- Theory of the Mind and Empathy (Sautter)
- Heritability of Behavior (the sample is twins)
- Funding Agency: Behavior Economics
- The latter: How well do they do when we pay them vs when we do not pay them?
  - Payment must be “big enough” even per face ($1?)
  - We have 38 faces (19 paid, 19 unpaid)
  - Missing by Design
  - Randomly Sample the Faces
  - Run it as IRT model with Missing Data

Warning

- The biggest problem with these missing data models is convergence
- It is a REALLY good idea to pre-test if your design is going to work

- How the hell can you do that?

Monte Carlo Simulations

- One of the methods for Power Analysis
  - Draw up the results you expect
  - Simulate an imaginary population
  - Sample from it
  - Create the missing data
  - Analyze it
  - Repeat
- You can look at power, % of models converged, etc.

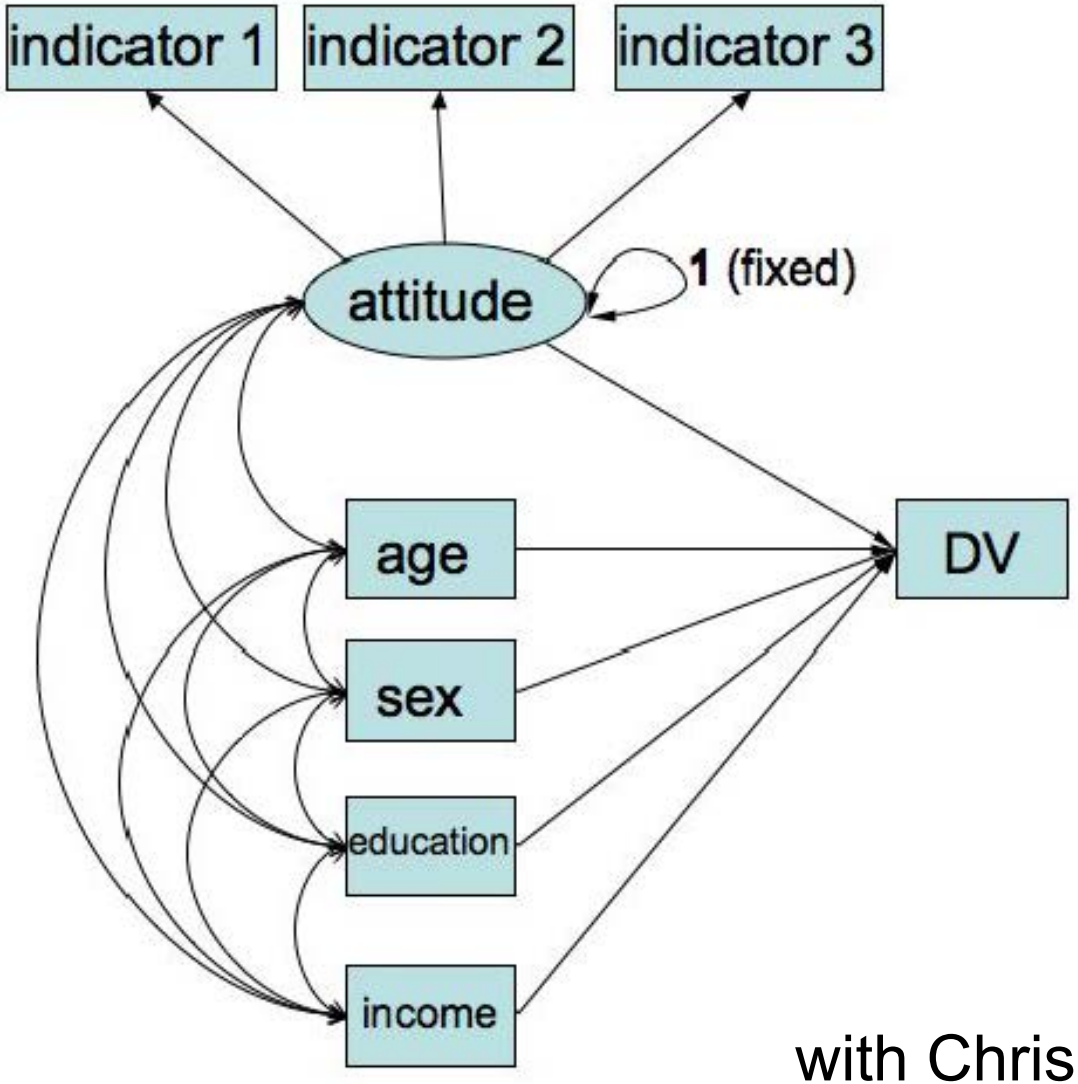
Context effect

- Survey Psychologists claim that if you ask multiple questions about the same construct context effects can emerge
- Measurement people claim that a single question cannot accurately measure a single construct. Therefore you need multiple questions and measurement models.

Fine. I Can Satisfy Both Groups

- It is a Missing Data problem
  - Prepare multiple questions
  - Randomly select one
  - Ask only that one from respondent
  - The rest is MCAR missing data
- Problems
  - Covariance Coverage
  - Convergence
  - Identification
- Possible Solutions
  - Multi-Group Structural Equation Model
  - Bayesian Estimator with WinBUGS (we did this)

The Model



with Chris Dawes

Merging Datasets

- Want to look at the the relationship between A and B
- You have information on A in one dataset
- You have information on B in another
- There is no available dataset with both A and B
- You discard the project

Merging Datasets as Missing Data

- The two datasets are sampled from the same population (around the same time)
- The people who were asked question A were randomly selected to get question A
- They have MCAR missing data on question B. And vice versa.
- FIML is problematic: no covariance coverage
- You can Impute or use Bayesian Estimation
- Expect very high rates of missing information

Another Warning

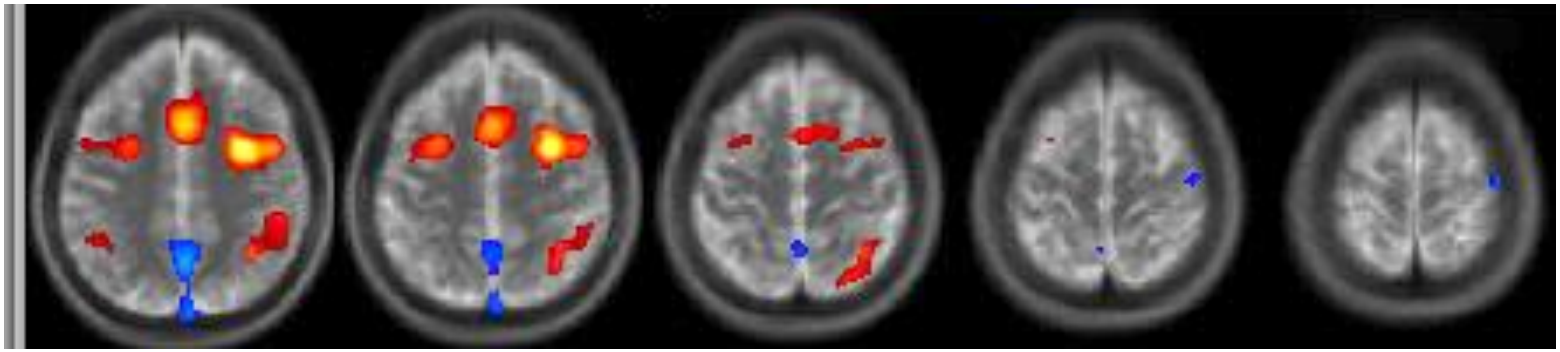
- These approaches are considered crazy
- The final word is always the reviewer's
- You might find and use the best missing data treatment available but if it scares your reviewer… Listwise it is... (Especially if results are the same. You can always use the footnote to show off.)
- Chris and I were talked out of pursuing publication on the context effect piece

What if the problem is a given

- Creating missing data by design is radical
- Not having the right data… Well… Get the money and collect it
- What if the problem is a given that everyone knows will not go away?

Functional Magnetic Resonance Imaging of the Brain

- They have a problem
- At any given time the brain can only be read in a 2 dimensional slice
- Could take 1-2 seconds to scan all slices



- All sorts of estimations are developed to interpolate all the slices into one time point
- Isn’t this an MCAR missing data problem?

Causal Analysis as a Missing Data Problem

- Experimental vs. Quasi-Experimental Design
- Experimental Design in Practice
  - Draw Sample
  - Assign to two groups randomly
  - Treat one group
  - Compare treatment group with control group
  - Isolates treatment as the sole cause

Quasi-Experimental Design

- Quasi-experimental Design in Practice
  - Collect sample
  - Test for the impact “treatment” (main predictor)
  - Control for every other possible correlate
- Example
  - Treatment = Smoking
  - Outcome = Lung Cancer
- Reformulation into a Missing Data Problem
  - You observe respondent 1 who smokes and got lung cancer. You did not observe what would have happened if s/he did not smoke

Causal Analysis - Growing Field

- New and cutting edge missing data research
  - I do not yet understand it
  - But I know where to go if I'll ever need it
- Missing Data Methods have allowed for
  - Possible isolation of causal direction
  - Better methods matching of samples
  - Legally solid model building
- Use weighting methods
- More on this: propensity scoring

Questions