A Very Basic Intro to SLCMA

Joshua A. Goode, PhD
INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

April 17, 2024

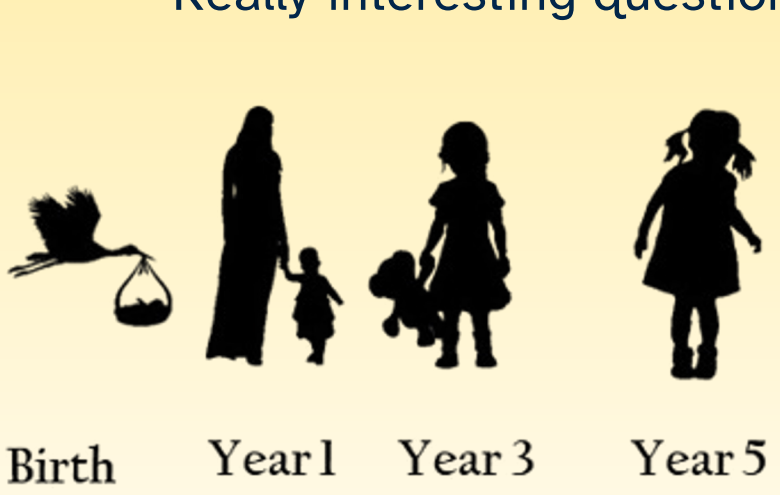
99 Problems

What We Have

- Several large panel studies
- Many years of data
- A plethora of data types
- Really interesting questions

What We Need

- Way to use our vast data to test hypotheses across the life course
  - Systematic to avoid false-positive results
  - Efficient to accommodate analysis of high-dimensional data
  - Easy to use



What Is SLCMA?

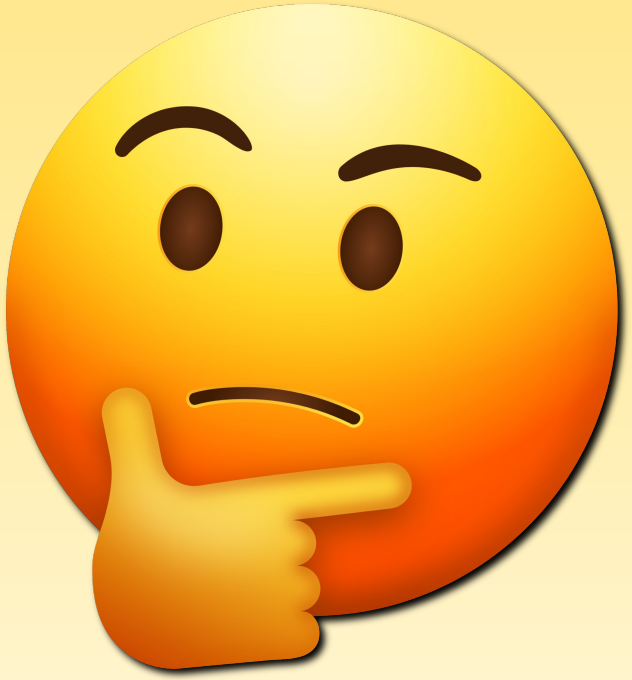
Structured

Life

Course

Modelling

Approach



Which Life Course Hypothesis Best Fits Our Data?

SLCMA R Package

- R Package (`slcma`)
  - Created by Andrew Smith
  - Under review by CRAN
- I'll be talking in the broadest terms
  - No code
  - Reach out for more info

SLCMA Steps

1. **Fit a regression model for each single life course hypothesis of interest, as well as groups of hypotheses**
   - **Manually:** Fit a model for each hypothesis, as well as compound hypotheses
   - **`slcma`:** Uses LARS to fit each model
2. **Measure the goodness-of-fit of each model and select the best one**
   - **Manually:** Create an elbow plot
   - **`slcma`:** Generates an elbow plot with a single command
3. **Calculate appropriate *p*-values for the selected model**
   - **Manually:** Apply a Bonferroni correction
   - **`slcma`:** Uses fixed LASSO inference or max-|t| test to correct *p*-values

High Throughput

Methyl Array Data

- A LOT OF MODELS!!
  - $850\text{K CpGs} \times 5\text{ Hyp} = 4.25\text{M Regs}$
- SLCMA can select the best-fitting hypothesis for each probe
  - Compound hypotheses are difficult
- SLCMA can typically find more associations than cross-sectional or ever-exposed EWAS models
  - Multiple-testing correction raises significance threshold significantly

SLCMA Hypotheses

Big 3

- Sensitive Periods
- Accumulation
- Recency

Others

- Mobility
- Change
- Always Exposed
- Ever Exposed

Hypotheses (Big 3)

Sensitive Periods (SP at t_j)

- The developmental timing of an exposure has the strongest effect on the outcome at a specific time point due to heightened levels of plasticity or reprogramming
- Just the exposure variable at each age
  - $SP_j = x_j$
- Can be continuous or binary

Hypotheses (Big 3)

Accumulation (Acc)

- Every additional time point of exposure affects the outcome in a dose-response manner, independent of the exposure timing
- Add up the exposure variable across ages
  - $Acc = \sum_{j=1}^{m} x_j$
- Can be continuous or binary

Hypotheses (Big 3)

Recency (Rec)

- More proximal exposures (closer in time to the of the outcome) are more strongly linked to the outcome than are more distal exposures
- Add up the products of each exposure variable multiplied by its age of observation
  - $Rec = \sum_{j=1}^{m}(x_j t_j)$
- Can be continuous or binary

Fitting Models

~~$Y \sim SP_1$~~

~~$Y \sim SP_2$~~

~~$Y \sim SP_3$~~

$Y \sim Acc$

~~$Y \sim Rec$~~

Test each hypothesis independently.
Accumulation is the best fit.

Fitting Models

$Y \sim Acc + SP_1$

~~$Y \sim Acc + SP_2$~~

~~$Y \sim Acc + SP_3$~~

~~$Y \sim Acc + Rec$~~

Models

$Y \sim Acc$

Test Accumulation w/ each remaining hypothesis.
Accumulation + Sensitive Period 1 is the best fit.

Fitting Models

~~$Y \sim Acc + SP_1 + SP_2$~~

$Y \sim Acc + SP_1 + SP_3$

~~$Y \sim Acc + SP_1 + Rec$~~

Models

$Y \sim Acc$

$Y \sim Acc + SP_1$

Test Accumulation + Sensitive Period 1 w/ each remaining hypothesis.
Accumulation + Sensitive Period 1 + Sensitive Period 3 is the best fit.

Fitting Models

~~$Y \sim Acc + SP_1 + SP_3 + SP_2$~~

~~$Y \sim Acc + SP_1 + SP_3 + Rec$~~

Models

$Y \sim Acc$

$Y \sim Acc + SP_1$

$Y \sim Acc + SP_1 + SP_3$

Model fit is identical in both remaining models.
Accumulation & Recency are linear combinations of SP_1, SP_2, and SP_3.

Fitting Models

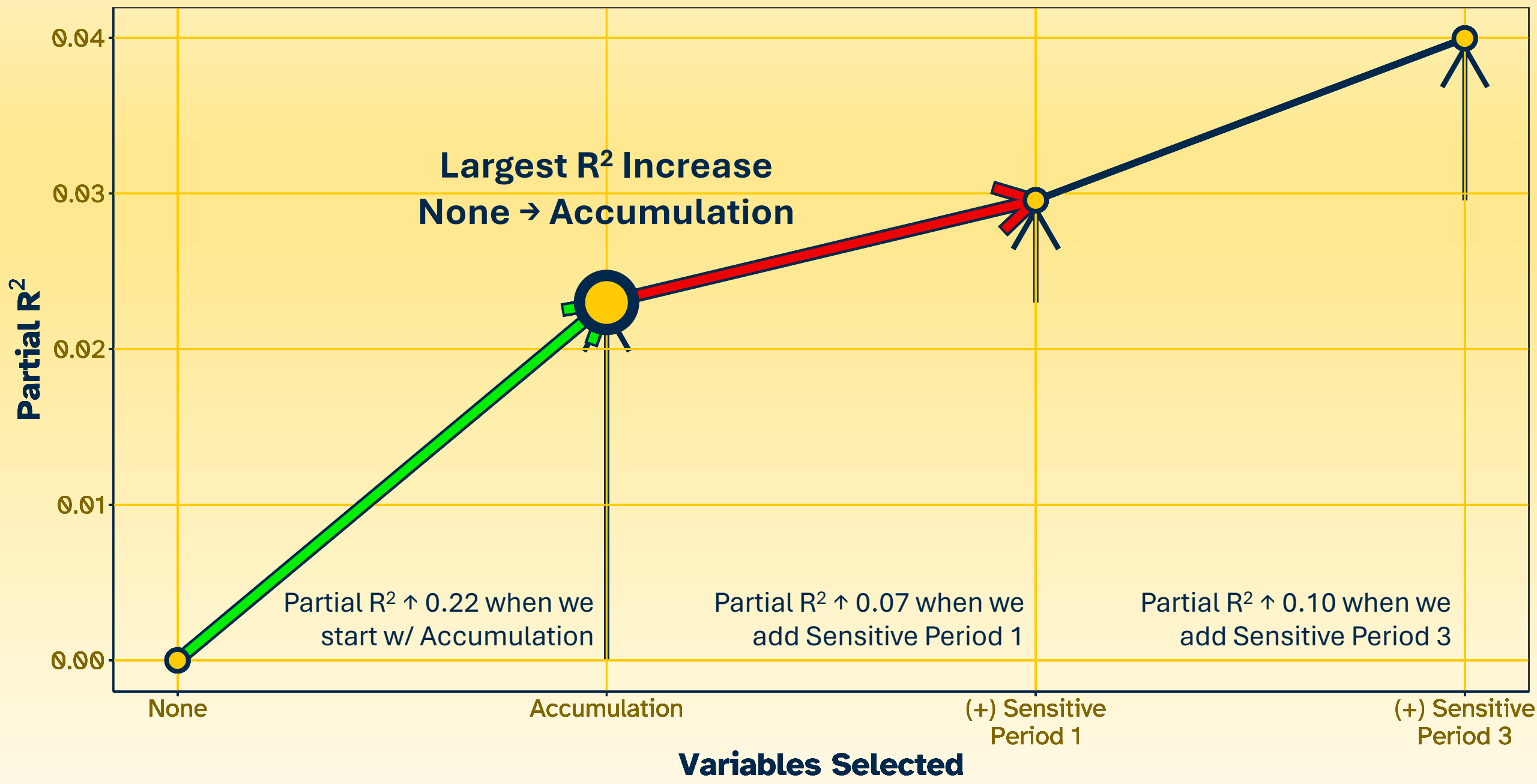
Models

$Y \sim Acc$

$Y \sim Acc + SP_1$

$Y \sim Acc + SP_1 + SP_3$

How Do We Choose?


Largest R² Increase
None → Accumulation
Partial R² ↑ 0.22 when we start w/ Accumulation
Partial R² ↑ 0.07 when we add Sensitive Period 1
Partial R² ↑ 0.10 when we add Sensitive Period 3
Partial R²
0.04
0.03
0.02
0.01
0.00
None
Accumulation
(+) Sensitive Period 1
(+) Sensitive Period 3
Variables Selected

Multiple Testing

One Hypothesis In Model, But We Tested Five

- We need to correct for multiple tests

- Naive Calculation
  - Inflated FWER; Biased *p*-values
  - Fast computation time
- Bonferroni Correction
  - FWER Controlled; Unbiased *p*-values
  - Overly conservative
- Fixed LASSO Inference
  - FWER Controlled; Unbiased *p*-values
  - Slow computation time
- Max-|t|-Test
  - FWER Controlled; Unbiased *p*-values
  - Slow computation time

Other Hypotheses

Mobility (t_j to t_k)

- The direction of change in exposure status between 2 time periods, rather than the absolute state at each individual time point, affects the outcome
- Upward mobility equal to 1 if exposed after unexposed, 0 if not
  - $Mob_{jk}^{+} = (1 - x_j)x_k$
- Downward mobility equal to 1 if unexposed after exposed, 0 if not
  - $Mob_{jk}^{-} = x_j(1 - x_k)$

Other Hypotheses

Change (t_j to t_k)

- The change in exposure status between 2 time periods, rather than the absolute state at each individual time point, affects the outcome
- Subtract the exposure at the later age from the earlier one
  - $Chn_{jk} = x_k - x_j$

Other Hypotheses

Always Exposed

- Constant exposure over time affects the outcome
- Equal to 1 if always exposed, 0 if not
  - $Alw = \prod_{j=1}^{m}(x_j)$

Other Hypotheses

Ever Exposed

- Exposure at any time point affects the outcome
- Equal to 1 if ever exposed, 0 if not
  - $Evr = 1 - \prod_{j=1}^{m}(1 - x_j)$

Theory Is Key

- SLCMA is inseparable from theory
- Better tools for machine learning outside of the theoretical framework
- Choose theoretically motivated hypotheses
  - Each hypothesis tested inflates p-values
  - Be parsimonious
- You can construct your own hypotheses

Thank You

Feel Free To Reach Out With Any Questions

Joshua A. Goode

jagoode@umich.edu