The parametric g-formula in SAS

JESSICA G. YOUNG

CIMPOD 2017

CASE STUDY 1

Structure of the workshop

Part I: Motivation

- Why we might use the parametric g-formula and how it works in general

Part II: GFORMULA SAS macro

- Structure of the macro
- Sample code

GFORMULA macro

Contributors: Miguel Hernán, Sarah Taubman, Roger Logan, Jessica Young, Sara Lodi, Sally Picciotto, Goodarz Danaei

Version on web: 2.0

Version for today: 3.0

Contact info

Updates to macro and documentation:

https://www.hsph.harvard.edu/causal/software/

My email:

jessica.gerald.young@gmail.com

PART I: MOTIVATION

Case study

Lajous et al. Changes in fish consumption in Midlife and the Risk of Coronary Heart Disease in Men and Women. *American Journal of Epidemiology* (2013).

Interested in estimating the **causal effect of different time-varying (sustained) fish consumption interventions on 18 year risk of coronary heart disease (CHD) in a study population**.

Causal effect

Population causal effects can be formally defined in terms of contrasts in counterfactual outcome distributions associated with different intervention rules:

- What would happen to the population 18 year risk of CHD if, *possibly contrary to fact*, we intervened on fish consumption at each time over the 18 year follow-up period in one way versus another?

Lajous et al:

Causal 18 year risk ratio/difference comparing :

1. Always eat at least 3 servings of fish per week versus
2. “Natural Course”: no intervention on fish

Ideal RCT

If we could, we would estimate such an effect in an ideal randomized controlled trial:

- Baseline randomization: subjects randomized to one of two treatment arms (at least 3 servings or do nothing)
- Full compliance with protocol until CHD event or 18 years later (whichever comes first)
- Eliminate “censoring events” (e.g. study drop out, incomplete follow-up)

Ideal RCT

- No confounding (by design)
- No selection bias (by design)
- Unbiased estimate of risk difference: difference in proportions of CHD in each arm

Challenge to causal inference

Ideal RCTs are often not feasible

- Too costly
- Not timely
- Unethical

Alternative: Observational studies

Observational data

No RCT had been conducted to answer the question of Lajous et al.

They used observational data to try and estimate their causal effect of interest:

- Health Professionals Follow-up Study
- **Nurses' Health Study**

Nurses' Health Study

- Prospective cohort study, enrolled 121,701 US female registered nurses aged 30-55 years in 1976.
- Participants reported via questionnaire information on health behaviors and newly diagnosed diseases every two years.
- Lajous et al. defined "baseline" as 1990 questionnaire
- Eligibility for inclusion: free of CVD, diabetes or cancer prior to 1986
- Sample of N= 53,772 women at baseline
- Censored subjects at first time failed to return questionnaire

Confounding and assumptions

- In Nurses' Health Study there is confounding (no exposure randomization at any time, no "forcing")
  - People who eat more fish at a given time may have past characteristics that make them more or less at risk for CHD
- Some subjects are also censored by incomplete follow-up (failure to return questionnaire)
- Some subjects also die of non-CHD causes…

Confounding and assumptions

If we are willing to assume "no **unmeasured** confounding or selection bias" (NUCS) we can get an unbiased estimate

- NUCS: Measured variables are sufficient to control confounding and selection by unmeasured risk factors
- NUCS is an untestable assumption – cannot test with study variables

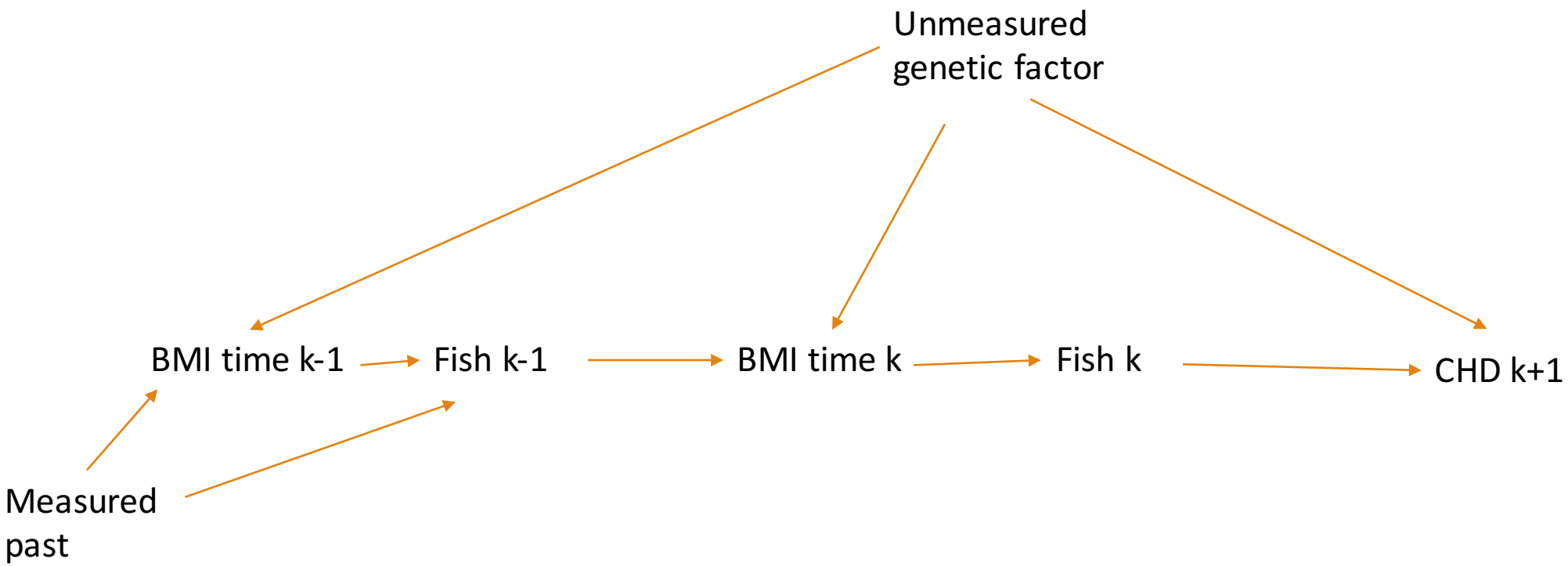
No unmeasured confounding

Key features of NUCS:

1. Allows presence of measured time-varying confounders (in addition to baseline confounding)
   - E.g. BMI measured in questionnaire interval k predicts future CHD and future fish intake.

2. Also allows that measured time-varying confounders are themselves affected by past exposure
   - E.g. BMI at k affected by Fish at k-1

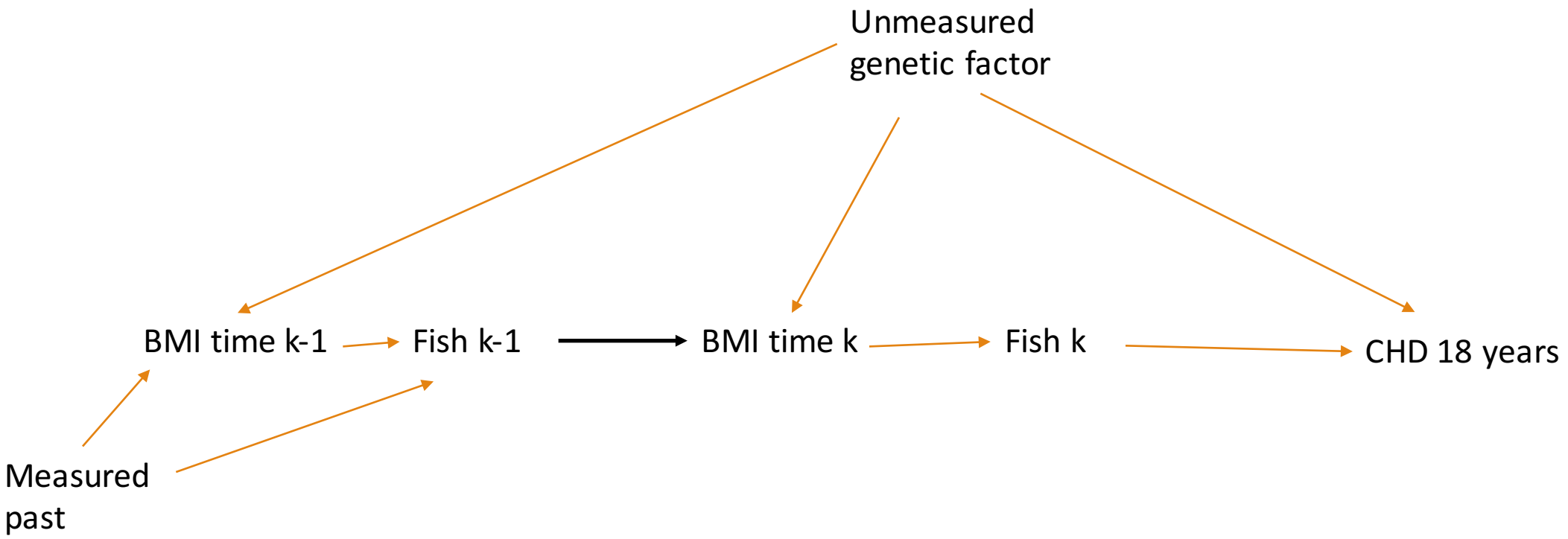
CAUSAL DAG REPRESENTING NO UNMEASURED CONFOUNDING



Absence of arrows from unmeasured risk factor into exposure guarantees no "unblocked backdoor paths" between exposure and outcome given measured past at any time.

Backdoor paths = confounded paths; Directed paths = causal paths

BMI MEASURED TIME-VARYING CONFOUNDER AFFECTED BY EXPOSURE



ALLOWS THAT MEASURED TIME-VARYING CONFOUNDER AFFECTED BY PAST EXPOSURE
("NO UNMEASURED CONFOUNDING" ALLOWS THIS STRUCTURE)

Time-varying confounding and standard regression

Turns out that even though, under this data structure, we can get an unbiased estimate,

- We cannot get it via standard regression approaches
- We need other approaches
- Why?

Standard outcome regression

Lajous et al.: might think to fit regression model for CHD hazard at a given time with independent variables:

- Function of time-varying exposure – cumulative average fish consumption through prior time k
- Function of baseline and time-varying confounders -- cumulative average of BMI through k

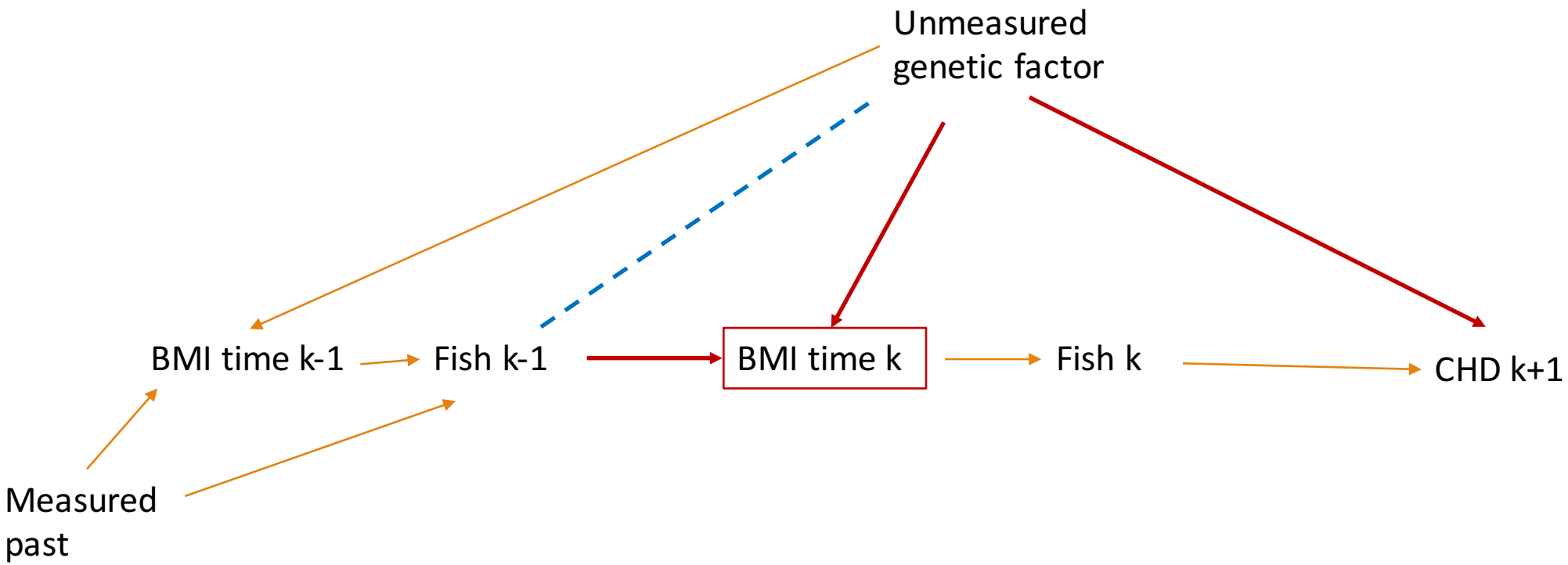
Use coefficient on cumulative average fish consumption as estimate of time-varying causal exposure effect

Standard outcome regression

Problem: even given “no unmeasured confounding” and outcome regression model correctly specified

- Coefficient is a biased estimate under our DAG

BMI MEASURED TIME-VARYING CONFOUNDER AFFECTED BY EXPOSURE



INCLUDING FUNCTION OF BMI AT k IN REGRESSION MODEL IS CONDITIONING ON IT

CONDITIONING ON COLLIDER OPENS UP NONCAUSAL PATH BETWEEN ITS CAUSES

Estimation in observational data with time-varying confounding

If not, standard regression, then how to proceed?

In this setting, methods that derive from Robins' g-formula can remain valid:

- They give unbiased estimates of time-varying causal exposure effects in the face of measured time-varying confounding affected by past exposure.

The g-formula

Robins (1986) showed that, given NUCS

- The counterfactual outcome mean/risk associated with a user-specified time-varying exposure intervention g can be written as the g-formula indexed by intervention g
- The g-formula indexed by g is a particular function of the baseline and time-varying data
- Estimated contrasts in this function for different choices of g can give unbiased estimates of causal effects
- Also requires “positivity” assumption

G-formula for Risk by end of follow-up under intervention g

Can write as weighted average of conditional risks

- Each risk conditioned on a possible treatment and confounder history observable under g and no censoring
- Weights are function of joint distribution of measured confounders at each time k conditional on past history observable under g and no censoring
  - i.e. the “chance” of observing each confounder history (under g) and no censoring

G-formula for Risk by end of follow-up under intervention g

Weights can also be a function of the observed distribution exposure at each time conditional on past history observable under g and no censoring

- This will be the case when g is defined in terms of intervention that depends on this distribution
- E.g. intervention is: “assign fish according to random draw from observed distribution of fish in Nurses’ Health Study”

How to estimate this function

In typical high-dimensional settings, we require parametric models to estimate the g-formula.

Different methods rely on different types of model assumptions

- Parametric g-formula: imposes models directly on components of weighted average
- Other methods derive from alternative representations of this weighted average which suggest constraining different quantities (e.g. IPW of MSMs, DR methods like TMLE)
- Equivalent under saturated models

Parametric g-formula Algorithm (Step 1)

First fits parametric models for

- Discrete hazard at each time conditional on past measured treatment and confounders
- Joint distribution of treatment and confounders at each time given past
- Models are generally pooled over time

Modelling joint distribution of covariates at k

Model based on arbitrary factorization of covariates at k. E.g.

$f(bmi_k, fish_k | past\ through\ k\text{-}1)$ is the same as

1. $f(bmi_k | fish_k, past\ through\ k\text{-}1) * f(fish_k | past\ through\ k\text{-}1)$ or
2. $f(fish_k | bmi_k, past\ through\ k\text{-}1) * f(bmi_k | past\ through\ k\text{-}1)$

Depending on choice of factorization, you are modelling the components of product 1 or product 2

- In absence of model misspecification, equivalent.
- Deterministic relationships may favor one factorization

Algorithm: Step 2

N times (default sample size) do the following iteratively for each k:

- Simulate exposure and confounders at each time k using estimated model parameters from Step 1. Exception: at k=0 (baseline) assign values as observed values in data set.
- Reset exposure at k according to user-defined rule *g*
  - *E.g. if simulated fish<3 reset to 3; otherwise do not intervene (threshold intervention)*
- Estimate hazard of event at time k given these generated covariate values using model in Step 1

Algorithm: Step 3

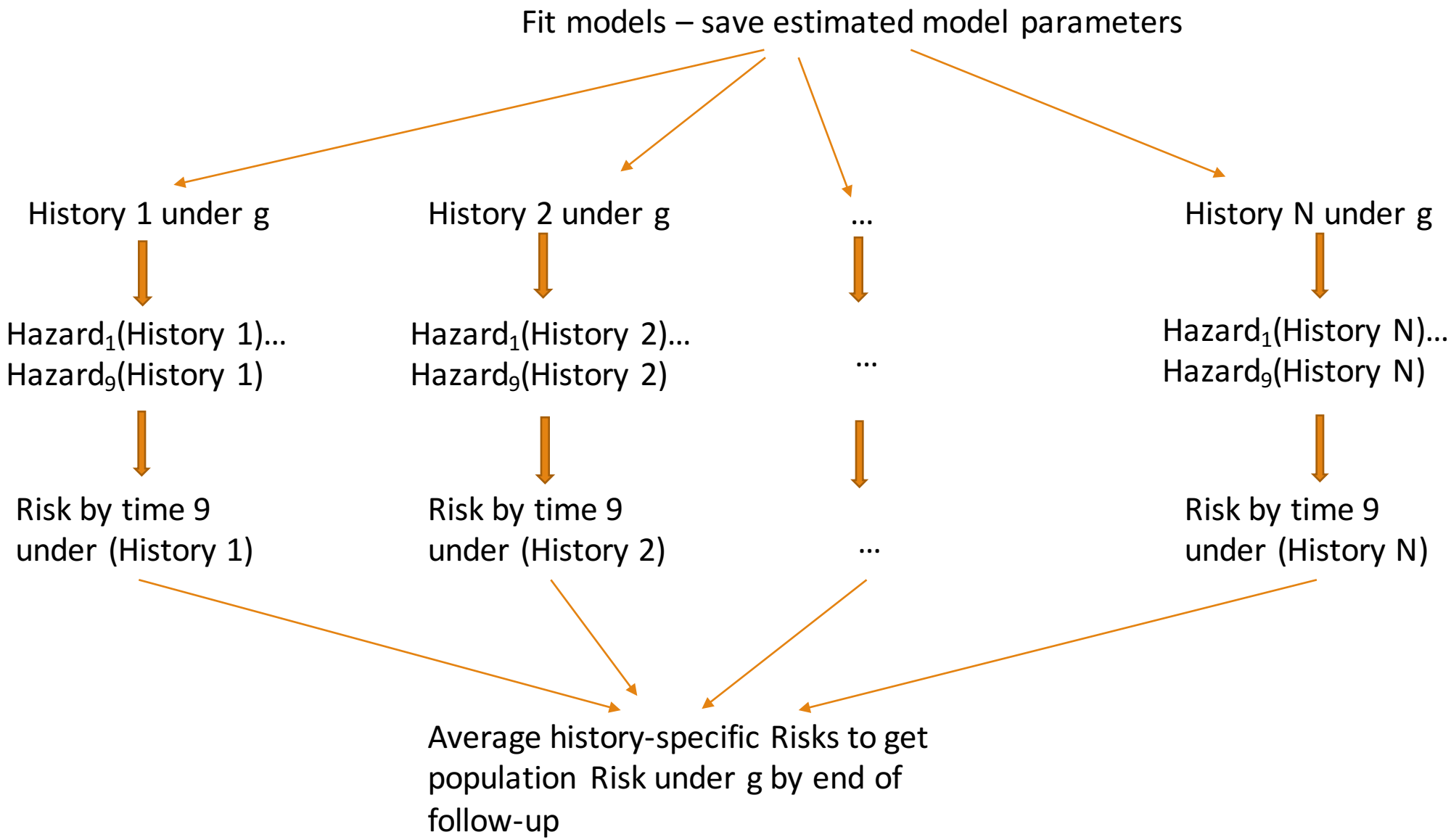
- Compute the Risk by end of follow-up for each of the N simulated histories from the N time-varying history-specific hazards.
- Average these Risks to get final estimate of Risk by end of follow-up under g

Final estimates and CIs

Repeat Steps 2 and 3 for each hypothetical intervention.

Obtain causal effect estimates from by risk differences/ratios for different g.

95% CIs obtained by repeating whole algorithm in B bootstrap samples.


Fit models – save estimated model parameters
History 1 under g
History 2 under g
…
History N under g
Hazard1(History 1)…
Hazard9(History 1)
Hazard1(History 2)…
Hazard9(History 2)
…
Hazard1(History N)…
Hazard9(History N)
Risk by time 9
under (History 1)
Risk by time 9
under (History 2)
…
Risk by time 9
under (History N)
Average history-specific Risks to get
population Risk under g by end of
follow-up

Disadvantages of parametric g-formula

- Relies heavily on parametric models and subject to related bias
- Some model misspecification can be theoretically guaranteed when null of no treatment effect is true
  - "null paradox" (Robins and Wasserman, 1997)

Advantages of parametric g-formula

- More stable than other methods for continuous exposures and given "near positivity violations"
  - Occurs when an intervention level of exposure is unlikely for certain observed confounder histories
  - Parametric g-formula handles by heavier reliance on extrapolation
- Generally, the complexity of algorithm is the same for any choice of g
  - Very little change for complex *dynamic* rules

PART II: GFORMULA SAS macro

Different types of outcomes

Macro supports 3 types of outcomes (*outctype)*

1. Continuous outcome *at* end of follow-up time
   - Choose when interest is in t-v exposure effect on *outcome mean at end of follow-up*
   - E.g. mean blood pressure *at* 5 years post-baseline
   - *outctype =conteofu*

Different types of outcomes

Macro supports 3 types of outcomes (*outctype)*

2. Binary outcome *at* end of follow-up time
   - Choose when interest is in t-v exposure effect on *probability that outcome occurs at end of follow-up*
   - E.g. Probability of obesity *at* 5 years post-baseline
   - *outctype =bineofu*

Different types of outcomes

Macro supports 3 types of outcomes (*outctype)*

3. Time-varying indicator of failure event
   - Choose when interest is in t-v exposure effect on *risk by end of follow-up*
   - E.g. Risk of CHD by 18 years post-baseline
   - ***outctype =binsurv***

Required structure for input data set

Requires a person-time data set with one record per subject and measurement time index

- Time index (*time*) must start at 0 (baseline) for each subject and increment by 1 for each subsequent time index.
- Time index represent a time interval in which covariates are measured
- Lajous et al.: each time index represents a two-year questionnaire interval

Required structure for input data set

Each person-time record will include

- Time-fixed baseline covariates (e.g. age at baseline, race)
- Current covariate measurements for that time k (bmi, fish measured in interval k)
- Time-varying indicator of censoring (e.g. failure to return interval k questionnaire)

Required structure for input data set

For *outctype*=binsurv (Lajous et al.):

- Will contain a time varying indicator of failure from event of interest for each time index k

Required structure for input data set

For *outctype*=binsurv (Lajous et al.):

- Time varying indicator of failure on line k can be coded 0, 1 or missing
  - Should be 0 if neither event nor censoring has occurred
  - Should be 1 if event has occurred
  - Should be missing if no event but censoring occurs
- First line k where outcome is 1 or missing is last line for that subject.

Required structure for input data set

Subjects who do not fail and are not censored by end of follow-up will be 0 at all times for censoring and event indicators (outctype=*binsurv*)

- Macro parameter *timepoints* encodes end of follow-up in terms of intervals
- Lajous et al.: timepoints=9 (9*2 year intervals=18 years)
- Because time index *time* starts at 0, can take maximum value of 8

SAMPLE DATA

FEATURES OF BASIC CALL

```
libname jess '/sasdata/CIMPOD_2017/Jessica_Young';
```

Define permanent libraries

```
%include '/home/jessica.gerald.young/gformula.sas';
```

Include the file with the macro

```
options notes;
*options mprint;
```

Set options for printing in log file

```
data fishdata;
set jess.fish;
run;
```

Call permanent data set

Define interventions

```
%let interv1 =intno=1,          intlabel='at least 2 servings of fish at all times',
nintvar=1,intvar1=fish, inttype1=2, intmin1=2, intpr1=1,    inttimes1=  0 1 2 3 4 5 6 7 8
;

%let interv2 =intno=2,          intlabel='at least 3 servings of fish at all times',
nintvar=1,intvar1=fish, inttype1=2, intmin1=3, intpr1=1,    inttimes1=  0 1 2 3 4 5 6 7 8
;
```

Call to GFORMULA macro

```
%gformula(
data=fishdata,
id=id,
time=time,
timeptype = conbin,
timepoints=9,
outctype=binsurv,
outc=death,
censlost=censor,
numint= 2,
fixedcov= race_1 race_2,
ncov=2,
cov1=cig,
cov1otype=3, cov1ptype = lag1bin,
cov2=fish,
cov2otype=3, cov2ptype=lag1bin,
nsamples= 0);
```

Input data set

Subject identifier from input data set

Time index from input data set

Specifies function of time for pooled over time models

End of follow-up (max value of time should be timepoints-1)

Specifies outcome is t-v binary failure indicator

Time-varying failure indicator for event of interest

Time-varying censoring indicator

Number of interventions to be simulated (minus natural course)

Time fixed baseline confounders

Number of time-varying covariates (including exposure), up to 30

Time varying covariate 1

Model specifications for t-v covariate 1

Time varying covariate 2

Model specifications for t-v covariate 2

Number of bootstrap samples

Graphical comparisons of natural course versus observed

- Set macro parameter *rungraphs=1*
- Compares “observed risk” (nonparametric estimates in censored data) versus parametric g-formula natural course estimates
- Comparison of observed covariate means versus simulated under natural course.
- Used to get a sense of presence of gross model misspecification

REVIEW OF OUTPUT

covXotypes

For each covX, X=1,…,ncov:

- The macro parameter *covXotype* selects the SAS regression fitting procedure for the conditional distribution of *covX.*
- Also determines how covX is simulated at each time *k*.
- Options available for *covXotype* are summarized in Table 1 of the documentation.

Examples of covXotype

- covXotype=1, estimates the conditional density of covX via PROC LOGISTIC. Simulation based on estimated model parameters.
  - Might be appropriate for binary variables that can take value 1 or 0 at any time with no restriction.
- covXotype=2, estimates via PROC LOGISTIC amongst records with lagged value of covX=0. Simulates from the model if last simulated value of covX was 0. If last value was 1, sets covX to 1.
  - Might be appropriate for binary variables that once they switch to 1, they stay 1 (e.g. diagnosis of diabetes *by* time k)

Examples of covXotype

- covXotype=3, estimate of conditional density of covX obtained via PROC REG. Simulation based on estimated model parameters under normality assumption.
  - Might be appropriate for continuous variables.
- covXotype=4, estimates conditional density of covX via two steps:
  - PROC LOGISTIC for whether covX>0 or not
  - PROC REG for records with covX>0
  - Simulation is in two steps
  - Might be appropriate for continuous variables with zero-heavy tails (e.g. number of cigarettes per day)

Exercise 1

Edit fishcall1.sas

1. Change the otype for time-varying covariate *cig* to cov1otype=4
2. Add additional baseline confounders: pre-baseline number of cigarettes (*cigprebl*), pre-baseline high blood pressure? (*hbpprebl*), pre-bl meat consumption (*mtprebl*), pre-bl fish consumption (*fishprebl*)

covXptypes

Determines how the “history” of covX will appear in each model

- Hazard model
- Models for covariate distributions
- Details in Table 2

covXptypes (lag1- prefix)

Prefix *lag1-* (lag1bin, lag1qdc,lag1zqdc,lag1cat,lag1spl)

- Includes function of one lagged value of covX in covariate models
- Includes function of current value of covX only in hazard models
- Function depends on choice of suffix
- Need to include lagged value of covX in input data set
  - Must be named covX_l1 (e.g. cig_l1 if covX=cig)

covXptypes (suffix options)

Suffix options that determine function:

- lag1bin: identity function (linear assumption when covX is not binary)
- lag1qdc: quadratic function
- lag1cat: categorization of variable (must also specify covXknots which give cutoffs for categories)
  - E.g. cov1=cig, cov1ptype=lag1cat,cov1knots= 5 10 15,…
- lag1spl: restricted cubic spline (must also specify covXknots)

covXptypes (lag2- prefix)

Prefix *lag2-* (lag2bin, lag2qdc,lag2zqdc,lag2cat,lag2spl)

- Includes function of two recent lagged values of covX in covariate models
- Includes function of current value of covX and covX_l1 only in hazard models
- Function depends on choice of suffix
- Need to include two lagged values of covX in input data set
  - Must be named covX_l1 and covX_l2

Other *covXptypes*

cumavg	Cumulative average	Creates and includes the cumulative average of entire history of *covX* relative to interval k beginning from time=0.
lag1cumavg	Cumulative average where the last term is pulled off the average	A variation of the cumavg ptype where the last term is pulled off of the average. In this case there are two generated predictors. At time = k these will be *covX* _l1 and the average of *covX* from time = 0 to time = k-2.
lag2cumavg	Cumulative average where the last two terms are pulled off the average	A variation of the cumavg ptype where the last two terms are pulled off of the average. In this case there are two generated predictors. At time = k these will be covX_l1, covX_l2, and the average of covX from time = 0 to time = k-3.
rcumavg	Recent cumulative average	Creates and includes the cumulative average of restricted history of *covX* relative to interval k based on two most recent values only.

Exercise 2

Edit fishcall1.sas

1. Change the ptype for time-varying covariate cig so models include indicators for categories of first lagged value of cig (can use cutoffs 1, 5 and 9).
2. Add a new time-varying covariate to the call with your choice of ptype and otype: hbp (a time-varying indicator of high-blood pressure in each interval k)
3. Add another new time-varying covariate to the call with your choice of ptype and otype: mt (number of servings of meat in each interval k)

Defining interventions

- Table 3 in documentation describes different types
- Do not need to define the natural course (by default this is run and is the default reference for causal contrasts)
- In a given intervention definition, can include up to 8 exposure variables (variables to undergo intervention)
- Interventions are defined before the call to the main GFORMULA macro in global macro variables interv1, interv2…

Code for threshold intervention on weekly Fish intake – eat *at least* 2 servings per week

```
%let interv1 = intno=1,
intlabel='at least 2 servings of fish',
nintvar=1,
intvar1=fish,
inttype1=2,
intmin1=2,
intpr1=1,
inttimes1=  0 1 2 3 4 5 6 7 8      ;
```

Intervention number

Intervention label

Number of variables to undergo intervention in this intervention

First intervention variable

Type of intervention on first intervention variable (Table 3)

When threshold intervention (inttype1=2), what is the lower threshold

Probability to perform this intervention on first intervention variable (default)

Intervention times for first intervention var

Code for “static” intervention on weekly Fish intake – eat *exactly* 2 servings per week

```
%let interv1 = intno=1,
intlabel='exactly 2 servings of fish',
nintvar=1,
intvar1=fish,
inttype1=1,
intvalue1=2,
intpr1=1,
inttimes1=  0 1 2 3 4 5 6 7 8      ;
```

Intervention number

Intervention label

Number of variables to undergo intervention in this intervention

First intervention variable

Type of intervention on first intervention variable (Table 3)

When static intervention (inttype1=1), what value to assign

Probability to perform this intervention on first intervention variable (default)

Intervention times for first intervention var

Exercise 3

Edit fishcall1.sas

1. Add a third intervention that is a threshold intervention on fish with cutoff at least 4 servings per day
2. Add a fourth intervention that is a joint intervention on mt and fish where meat is “always set to exactly 0 servings per day” and fish “at least 2 servings”

Hint: remember to change *numint* from 2 to 4 in main GFORMULA macro call.

covXptypes (skp- prefix)

Prefix *skp-* (skpbin, skpqdc,skpzqdc,skpcat,skpspl)

- Can be used when covX not measured in certain intervals for anyone and last measured value forward
- E.g. in Nurses' Health Study, certain variables not measured in certain questionnaire years
- Data coded such that last value carried forward in these years
- Assumption is that most recent measurement sufficient to ensure NUCS

covXptypes (skp- prefix)

Prefix *skp-* (skpbin, skpqdc,skpzqdc,skpcat,skpspl)

- Includes function of most recent value of covX and interaction between this value and time since last measured
- Must specify length of each time interval (*interval*)
  - E.g. interval=2 (2 year questionnaire interval)
- Specify values of *time* in which covX is not measured (*covXskip*)
  - E.g. cov1skip= 2 5

timeptype

Determines functional form of time in pooled over time models:

- *conbin, concat, conqdc, conspl*.
- Follows suffix functions for covXptype
- For *concat* and *conspl*, must define *timeknots* (the chosen category cutoffs/knots, separated by spaces).