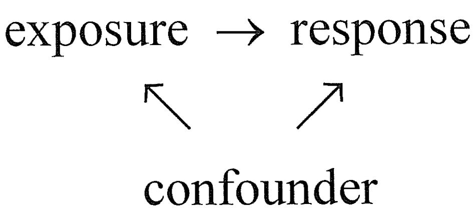
Time-dependent confounders

Niels Keiding

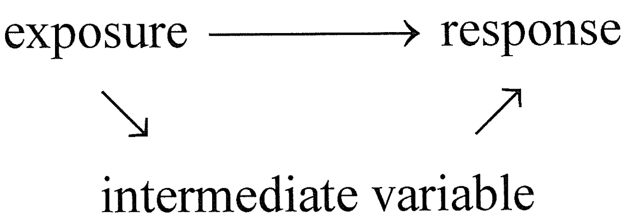
Confounders and intermediate variables



exposure → response

confounder

Control for confounder



exposure → response

intermediate variable

Do not control for intermediate variable

Time-dependent confounders

P.J. Diggle, P.J. Heagerty, K.-Y. Liang, S.L. Zeger (2002). Analysis of longitudinal data. Oxford University Press, Chapter 12.

$$\begin{tikzcd}
 & Y_i \arrow[rr] \arrow[dr] & & Y_{i+1} \\
X_{i-1} \arrow[ur] \arrow[rr] & & X_i \arrow[ur] &
\end{tikzcd}$$

Y_i intermediate between X_{i-1} and Y_{i+1}

Y_i confounder for effect of X_i on Y_{i+1}

Exogenous and endogenous

$$\begin{tikzcd}
 & \text{Time} & & 1 & 2 & \cdots & T \\
 & \text{Response} & & Y_{i1} \arrow[dd] & Y_{i2} \arrow[dd] & \cdots & Y_{iT} \arrow[dd] \\
\text{Individual } i & & & & & & \\
 & \text{Covariate} & Z_i & X_{i1} \arrow[uur] & X_{i2} \arrow[uur] & \cdots \arrow[uur] & X_{iT}
\end{tikzcd}$$

$\mathcal{H}_i^x(t) = \{X_{i1}, \cdots, X_{it}\}$ history of covariate through t

$\mathcal{H}_i^y(t) = \{Y_{i1}, \cdots, Y_{it}\}$ history of response through t

Covariate process

exogenous if

$$X_{it} \mid \mathcal{H}_i^y(t), \mathcal{H}_i^x(t-1), Z_i \quad \sim \quad X_{it} \mid \mathcal{H}_i^x(t-1), Z_i$$

covariate at time *i* given previous covariates independent of response history

endogenous if not **exogenous**

Exogeneity: comments

Exogeneity = Granger non-causality see Engle et al. 1983

Engle & Granger received Nobel prize in economics 2003

Exogeneity: $Y_{it} \perp \left(X_{it}, \cdots, X_{iT} \right) \mid \mathcal{H}_i^y(t-1), \mathcal{H}_i^x(t-1)$

Response Y_{it} at time t is

conditionally independent of all future covariates

given past outcomes and past covariates.

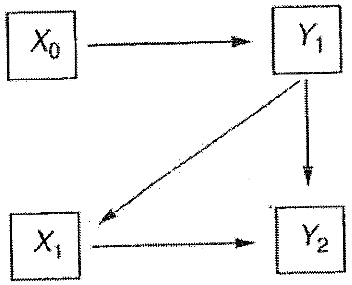
Simple feedback example

$$\operatorname{logit} \mathrm{E}(Y_1 \mid X_0 = x_0) = -0.5 - 0.5 \cdot x_0, \tag{12.5.1}$$

$$\operatorname{logit} \mathrm{E}(X_1 \mid Y_1 = y_1, X_0 = x_0) = -0.5 + 1.0 \cdot y_1, \tag{12.5.2}$$

$$\operatorname{logit} \mathrm{E}(Y_2 \mid \mathcal{H}_1^X = h_1^X, Y_1 = y_1) = -1.0 + 1.5 \cdot y_1 - 0.5 \cdot x_1, \tag{12.5.3}$$

$$X_i = \begin{cases} 1 & \text{treatment} \\ 0 & \text{not treatment} \end{cases} \qquad Y_i = \begin{cases} 1 & \text{symptoms} \\ 0 & \text{no symptoms} \end{cases}$$



Goal: See effect of X_0, X_1 on Y_2

Note: Y_1 confounder wrt X_1, Y_2
Y_1 intermediate between X_0, Y_2

Conventional analyses

1. Marginal mean $\mu_2(X_0, X_1) = P(Y_2 = 1 | X_0 = x_0, X_1 = x_1)$
   Table 12.6

Analysis does not recognize confounder Y_1

2. Condition on Y_1

Eliminates X_0, that now being blocked by Y_1 (intermediate).

Table 12.6. Expected counts when 500 subjects are initially treated, $X_0 = 1$, and 500 subjects are not treated, $X_0 = 0$, when treatment at time 2, X_1, is predicted by the outcome at time 1, Y_1, according to the model given by (12.5.1)–(12.5.3).

X_0	0								1							
n	500								500							
Y_1	0				1				0				1			
n	311				189				366				134			
X_1	0		1		0		1		0		1		0		1	
n	194		117		71		118		227		138		51		84	
Y_2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
n	142	52	96	21	27	44	59	59	166	61	113	25	19	32	42	42

$$\mathrm{E}(Y_2 \,|\, x_0 = 1, x_1 = 1) = (25 + 42)/(138 + 84) = 0.30$$

$$\mathrm{E}(Y_2 \,|\, x_0 = 1, x_1 = 0) = (61 + 32)/(227 + 51) = 0.33$$

$$\mathrm{E}(Y_2 \,|\, x_0 = 0, x_1 = 1) = (21 + 59)/(117 + 118) = 0.34$$

$$\mathrm{E}(Y_2 \,|\, x_0 = 0, x_1 = 0) = (52 + 44)/(194 + 71) = 0.36$$

See vs. do

Observational studies: we fix what we see

$$P(Y_2 = 1 \mid X_0 = 1, X_1 = 1) = 0.30$$

$$P(Y_2 = 1 \mid X_0 = 0, X_1 = 0) = 0.36$$

Intervention: we fix covariates “externally”

$$P(Y_2 = 1 \mid \mathrm{do}\, X_0 = 1, X_1 = 1) = 0.267$$

$$P(Y_2 = 1 \mid \mathrm{do}\, X_0 = 0, X_1 = 0) = 0.402$$

see Table 12.8 (“g-computation”)

Table 12.8. Expected outcomes when treatment is controlled and the causal path leading from Y_1 to X_1 is blocked.

All subjects $X_0 = X_1 = 1$

X_0	**0**								**1**							
n	0								1000							
Y_1	**0**				**1**				**0**				**1**			
n	0				0				731				269			
X_1	**0**		**1**		**0**		**1**		**0**		**1**		**0**		**1**	
n	0		0		0		0		0		731		0		269	
Y_2	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
n	0	0	0	0	0	0	0	0	0	0	598	133	0	0	134	134

$\mu^{(1)} = (133 + 134)/1000 = 0.267$

All subjects $X_0 = X_1 = 0$

X_0	**0**								**1**							
n	1000								0							
Y_1	**0**				**1**				**0**				**1**			
n	622				378				0				0			
X_1	**0**		**1**		**0**		**1**		**0**		**1**		**0**		**1**	
n	622		0		378		0		0		0		0		0	
Y_2	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
n	455	167	0	0	143	235	0	0	0	0	0	0	0	0	0	0

$\mu^{(0)} = (167 + 235)/1000 = 0.402$

Do calculus

We postulate that simple building blocks from observational studies may be lifted out of context and combined under manipulated conditions where the feedback paths are blocked. This is postulated to simulate intervention studies.

Main assumption: **No unmeasured confounders**

At time t the exposure X_t is independent of future potential outcomes given observed exposure history $\mathcal{H}^X(t-1) = (X_0, \cdots, X_{t-1})$ and observed outcome history $\mathcal{H}^Y(t) = (Y_0, \cdots, Y_t)$.

Potential outcomes

	$Y_{it}^{(x_t)}$	outcome if treatment $\boldsymbol{x}_t = (x_0, \cdots, x_{t-1})$ followed
	$Y_{it}^{(\mathbf{0})}$	outcome if treatment $\boldsymbol{x}_t = \mathbf{0} = (0, \cdots, 0)$ followed
Examples		
	$Y_{it}^{(\mathbf{1})}$	outcome if treatment $\boldsymbol{x}_t = \mathbf{1} = (1, \cdots, 1)$ followed

Outcomes other than that observed are called *counterfactual*

Causal effect of $\mathbf{1} = (\underset{0}{1}, \cdots, \underset{t-1}{1})$ vs. $\mathbf{0} = (\underset{0}{0}, \cdots, \underset{t-1}{0})$ at t

$$Y_{it}^{(\mathbf{1})} - Y_{it}^{(\mathbf{0})}$$

cannot be observed directly since we only observe one potential outcome per subject.

Estimation of causal effects in randomised studies

average response of treated subjects $E\left(Y_{it} \mid \boldsymbol{x}_t = \mathbf{1}\right)$

is an unbiased estimate of mean of $Y_{it}^{(1)}$ in entire population

and similarly for untreated subjects

$$E\left(Y_{it} \mid \boldsymbol{x}_t = \mathbf{0}\right)$$

estimates mean of $Y_{it}^{(0)}$ in entire population

Causal effect $\delta_t = \mu_t^{(1)} - \mu_t^{(0)}$ $\mu_t^{(x_t)} = E\left(Y_t^{(x_t)}\right)$

Covariates at baseline z

Causal effect $\delta_t(z) = \mu_t^{(1)}(z) - \mu_t^{(0)}(z)$ $\mu_t^{(x_t)} = E\left(Y_t^{(x_t)} \mid \boldsymbol{Z} = z\right)$

Estimation of causal effects in observational studies

We assume that we have recorded so much confounder information that the study can be considered essentially randomised given these confounders

No unmeasured confounders

At time t the exposure X_t is independent of future potential outcomes given observed exposure history $\mathcal{H}^X(t-1) = (X_0, \cdots, X_{t-1})$ and observed outcome history $\mathcal{H}^Y(t) = (Y_0, \cdots, Y_t)$.

g-computation

J.M. Robins

Likelihood decomposition

$$\mathcal{L} = \prod_{t=1}^{T} P\left[Y_{it} \mid \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-1), \boldsymbol{z}_i\right] P\left[X_{it-1} \mid \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-2), \boldsymbol{z}_i\right]$$

$$= \qquad \mathcal{L}_Y \qquad \bullet \qquad \mathcal{L}_X$$

response transitions covariate transitions

Under no unmeasured confounders the *causal effect of treatment* i.e. effect of **do** $\boldsymbol{x}_t = (x_1, \cdots, x_t)$ can be identified from $\mathcal{L}_Y$:

$$P\left[Y_{it}^{(\boldsymbol{x}_i)} \mid Y_{i1}^{(\boldsymbol{x}_t)}, \cdots, Y_{it-1}^{(\boldsymbol{x}_t)}, \boldsymbol{z}_i\right] = P\left[Y_{it} \mid \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-1) = \boldsymbol{x}_t, \boldsymbol{z}_i\right]$$

Mothers' Stress and Children's Morbidity (MSCM)

Alexander & Markowitz, Medical Care **24**, 134-147, 1986.

167 preschool children aged 1½-5 years attending inner-city paediatric clinic.

Here focus on 28 day diary containing maternal stress X_{it} (no=0, yes=1) and child illness Y_{it} (no=0, yes =1).

Baseline covariates: Table 12.1

Sample of diary data: Fig. 12.2

Table 12.1. Covariate summaries for mothers who were employed outside the home and those who were not.

	Employed = 1	Employed = 0
	$n = 55$	$n = 112$
	(%)	(%)
Married		
0 = no	42	57
1 = yes	58	43
Maternal health		
1, 2 = fair/poor	9	17
3 = good	33	34
4 = very good	47	32
5 = excellent	11	17
Child health		
1, 2 = fair/poor	7	5
3 = good	7	16
4 = very good	55	46
5 = excellent	31	33
Race		
0 = white	62	37
1 = non-white	38	63
Education		
0 $\leq$ high school	16	59
1 = HS graduate	84	41
Household size		
0 = less than 3	38	31
1 = 3 or more	62	69

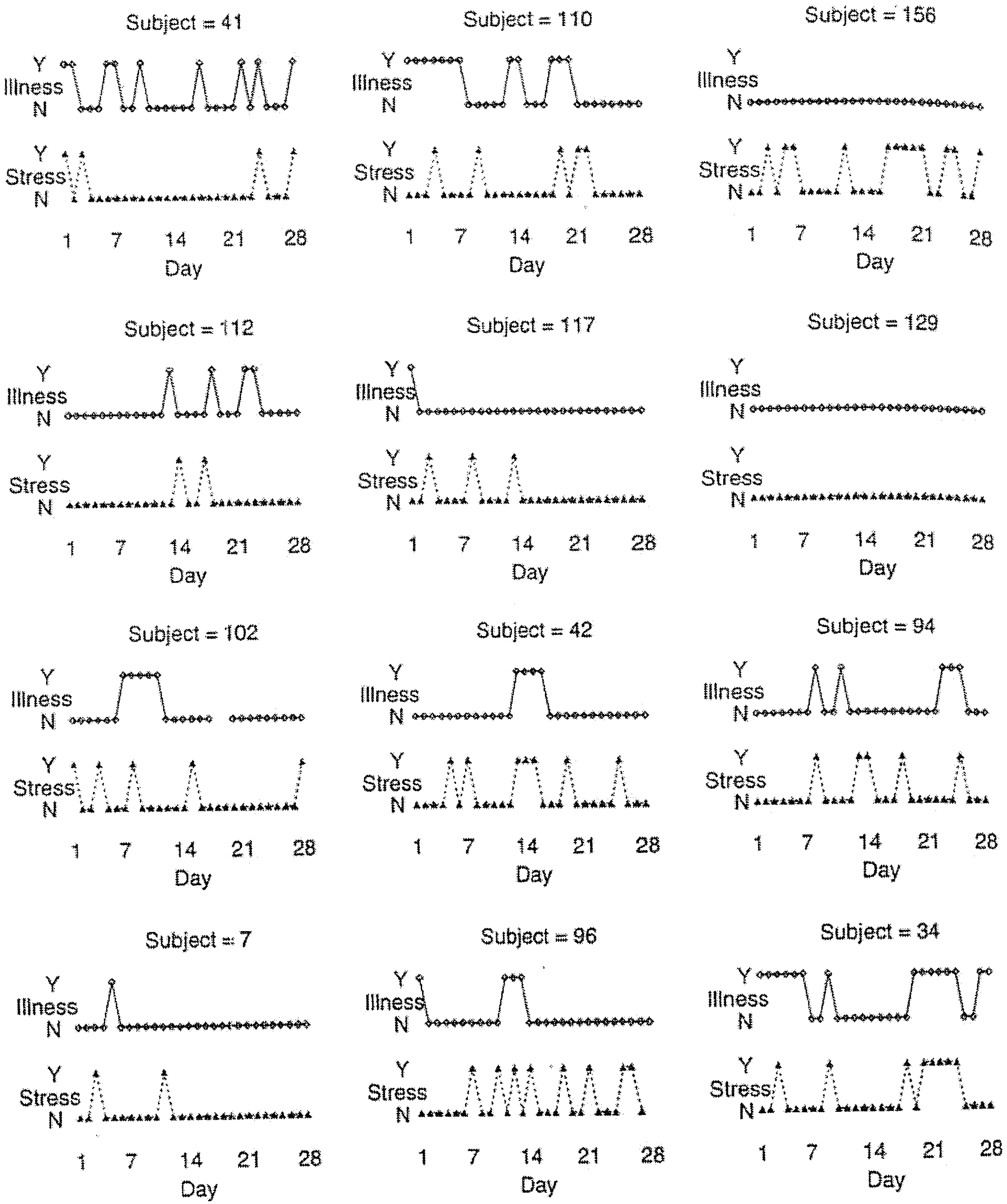


Fig. 12.2. A random sample of data from the MSCM study. The presence or absence of maternal stress and child illness is displayed for each day of follow-up.

MSCM: endogeneity (feedback)

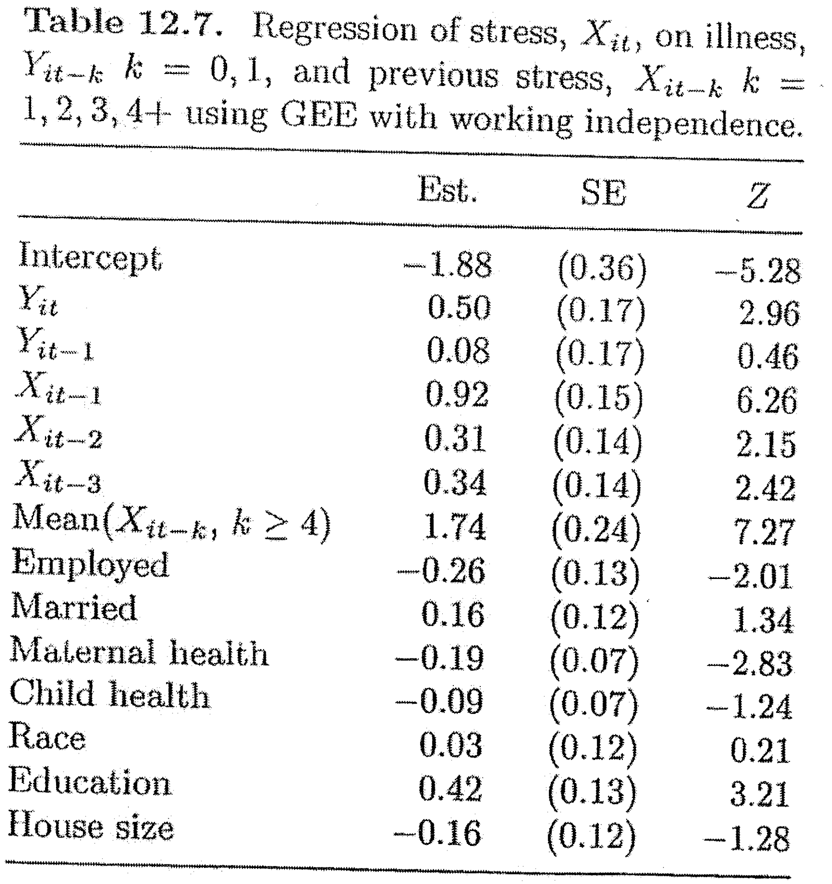


Table 12.7. Regression of stress, X_{it}, on illness, Y_{it-k} $k = 0, 1$, and previous stress, X_{it-k} $k = 1, 2, 3, 4+$ using GEE with working independence.

	Est.	SE	Z
Intercept	−1.88	(0.36)	−5.28
Y_{it}	0.50	(0.17)	2.96
Y_{it-1}	0.08	(0.17)	0.46
X_{it-1}	0.92	(0.15)	6.26
X_{it-2}	0.31	(0.14)	2.15
X_{it-3}	0.34	(0.14)	2.42
Mean(X_{it-k}, $k \geq 4$)	1.74	(0.24)	7.27
Employed	−0.26	(0.13)	−2.01
Married	0.16	(0.12)	1.34
Maternal health	−0.19	(0.07)	−2.83
Child health	−0.09	(0.07)	−1.24
Race	0.03	(0.12)	0.21
Education	0.42	(0.13)	3.21
House size	−0.16	(0.12)	−1.28

Child illness Y_{it} predicts mother's stress X_{it} even after controlling for prior stress variables

MSCM: Illness day t (Y_{it}) depending on mother's stress days $t-1, t-2, \cdots, t-7$.

Logistic regression on week, employment status, marital status, maternal and child health at baseline, race, education, household size

and

	X_{it-1}	X_{it-2}	X_{it-3}	X_{it-4}	X_{it-5}	X_{it-6}	X_{it-7}
estimate	0.34	-0.05	0.18	0.25	0.22	0.19	0.25
s.e.	0.16	0.15	0.13	0.13	0.14	0.14	0.14

so contrast between $X_{iu} = 1$ for all u and $X_{iu} = 0$ for all u gives

log odds ratio $0.34 - 0.05 + 0.18 + \cdots + 0.25 = 1.38$

but we have ignored confounder: child health $Y_{iu}, u = t-6, \cdots, t-1$.

MSCM: logistic regression of child illness on previous illness and mother's stress

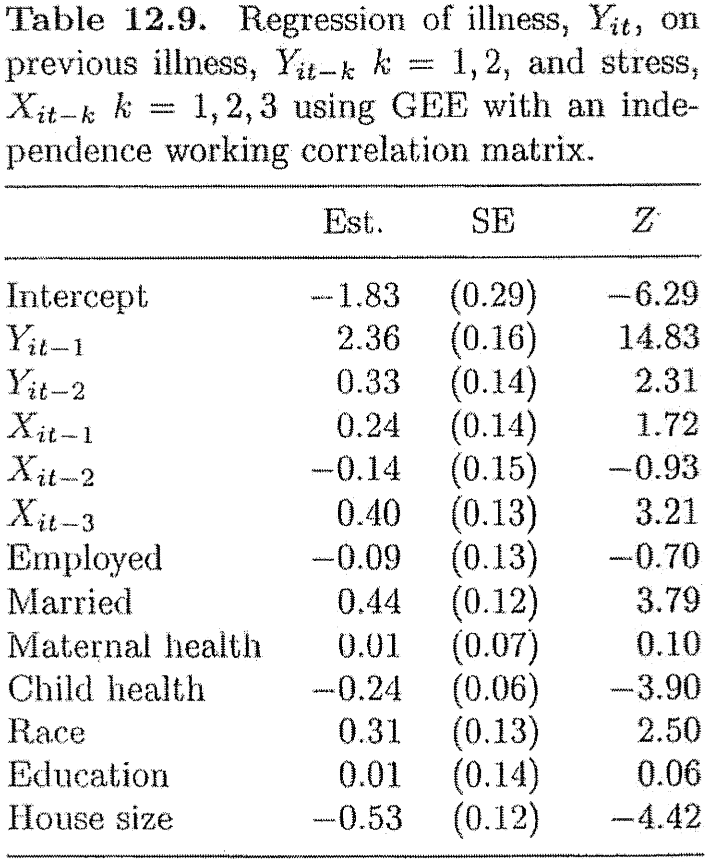


Table 12.9. Regression of illness, Y_{it}, on previous illness, Y_{it-k} $k = 1, 2$, and stress, X_{it-k} $k = 1, 2, 3$ using GEE with an independence working correlation matrix.

	Est.	SE	Z
Intercept	−1.83	(0.29)	−6.29
Y_{it-1}	2.36	(0.16)	14.83
Y_{it-2}	0.33	(0.14)	2.31
X_{it-1}	0.24	(0.14)	1.72
X_{it-2}	−0.14	(0.15)	−0.93
X_{it-3}	0.40	(0.13)	3.21
Employed	−0.09	(0.13)	−0.70
Married	0.44	(0.12)	3.79
Maternal health	0.01	(0.07)	0.10
Child health	−0.24	(0.06)	−3.90
Race	0.31	(0.13)	2.50
Education	0.01	(0.14)	0.06
House size	−0.53	(0.12)	−4.42

Compare $Y_{iu} = 1$ for all u to $X_{iu} = 0$ for all u from table:
log odds ratio
$= 0.24 - 0.14 + 0.40 = 0.50$
BUT we have conditioned on intermediate variables Y_{it-1}, Y_{it-2}.

MSCM: *g*-computation

Assume employed=0, married=0, maternal and child health=4, race=0, education=0, house size=0

$$P\left(Y_{iT}=1 \mid \boldsymbol{x}_t=\mathbf{1}\right)=0.189 \qquad P\left(Y_{iT}=1 \mid \boldsymbol{x}_t=\mathbf{0}\right)=0.095$$

$T = 28$ days

Causal log odds ratio

$$\log\left(\frac{0.189}{1-0.189}\bigg/\frac{0.095}{1-0.095}\right)=0.80$$

MSCM: comparison of three approaches

Effect of stress on illness at end *(always stress vs. never stress)*	*log odds ratio*
ignoring previous illness (ignoring confounders)	1.38
conditioning on previous illness (conditioning on intermediate variables)	0.50
g-computation	0.80

Marginal structural models

J.M. Robins

Regression model for counterfactual outcome

Example

$$X_{it}^{*} = \sum_{s<t} X_{is}$$ cumulative exposure

$$\text{logit} \ \mu^{(x_t)}(z) = \beta_0 + \beta_1 X_{it}^{*} + \beta_2^{'} \boldsymbol{Z}_i$$

Here β_1 is one single causal effect parameter.

How to estime β_1?

Construct *pseudopopulation* free of confounding using *inverse probability of treatment weights* IPTW.

Stabilized weights

$$SW_i(t) \quad = \prod_{s<t} \frac{P\left(X_{is} = x_{is} \mid \mathcal{H}_i^X(s-1) = h_i^X(s-1), z_i\right)}{P\left(X_{is} = x_{is} \mid \mathcal{H}_i^Y(s-1) = h_i^Y(s-1), \mathcal{H}_i^X(s-1) = h_i^X(s-1), z_1\right)}$$

$$= \frac{\text{Prob. treatment received until } t-1 \text{ conditional on treatment history}}{\text{Prob. treatment received until } t-1 \text{ conditional on treatment and response history}}$$

$$= 1$$

if covariate process exogenous (by definition)

Example of IPTW

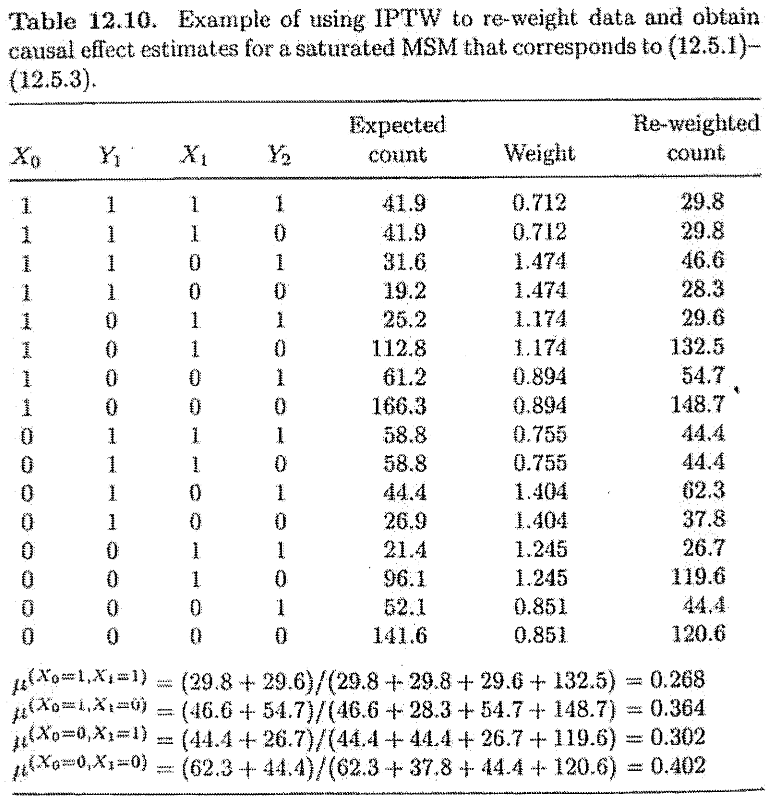
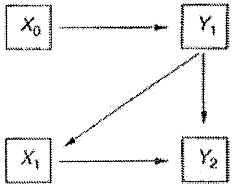


Table 12.10. Example of using IPTW to re-weight data and obtain causal effect estimates for a saturated MSM that corresponds to (12.5.1)–(12.5.3).

X_0	Y_1	X_1	Y_2	Expected count	Weight	Re-weighted count
1	1	1	1	41.9	0.712	29.8
1	1	1	0	41.9	0.712	29.8
1	1	0	1	31.6	1.474	46.6
1	1	0	0	19.2	1.474	28.3
1	0	1	1	25.2	1.174	29.6
1	0	1	0	112.8	1.174	132.5
1	0	0	1	61.2	0.894	54.7
1	0	0	0	166.3	0.894	148.7
0	1	1	1	58.8	0.755	44.4
0	1	1	0	58.8	0.755	44.4
0	1	0	1	44.4	1.404	62.3
0	1	0	0	26.9	1.404	37.8
0	0	1	1	21.4	1.245	26.7
0	0	1	0	96.1	1.245	119.6
0	0	0	1	52.1	0.851	44.4
0	0	0	0	141.6	0.851	120.6

$$\mu^{(X_0=1,X_1=1)} = (29.8 + 29.6)/(29.8 + 29.8 + 29.6 + 132.5) = 0.268$$
$$\mu^{(X_0=1,X_1=0)} = (46.6 + 54.7)/(46.6 + 28.3 + 54.7 + 148.7) = 0.364$$
$$\mu^{(X_0=0,X_1=1)} = (44.4 + 26.7)/(44.4 + 44.4 + 26.7 + 119.6) = 0.302$$
$$\mu^{(X_0=0,X_1=0)} = (62.3 + 44.4)/(62.3 + 37.8 + 44.4 + 120.6) = 0.402$$



$$SW(2) = \frac{P(X_1 \mid X_0)}{P(X_1 \mid Y_1, X_0)}$$

$X_0 = 1, Y_1 = 1, X_1 = 1$ $\quad SW(2) = 0.712$

downweighted

$X_0 = 1, Y_1 = 0, X_1 = 1$ $\quad SW(2) = 1.174$

upweighted

In reweighted population

$$P(X_1 = 1 \mid Y_1 = 1, X_0 = 1) = \frac{29.8 + 29.8}{29.8 + 29.8 + 46.6 + 28.3} = 0.443$$

$$P(X_1 = 1 \mid Y_1 = 0, X_0 = 1) = \frac{29.6 + 132.5}{29.6 + 132.5 + 54.7 + 148.7} = 0.443$$

so Y_1 no longer confounder (but still intermediate variable)

MSCM: Marginal structural models using IPTW

$$SW: \frac{\text{logistic regression as in Table 12.7 excluding } Y_{it}, Y_{it-1}}{\text{logistic regression as in Table 12.7}}$$

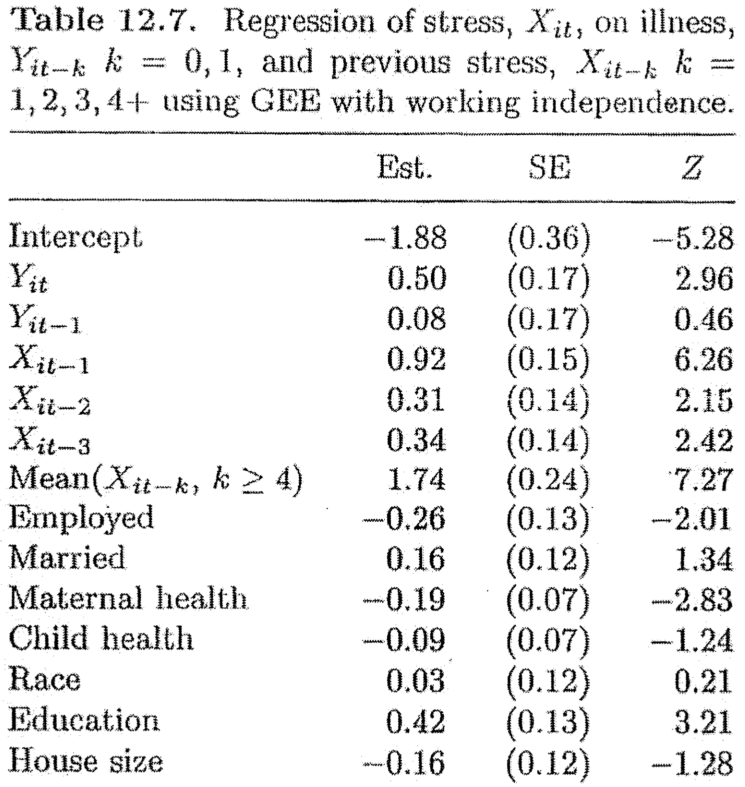


Table 12.7. Regression of stress, X_{it}, on illness, Y_{it-k} $k = 0, 1$, and previous stress, X_{it-k} $k = 1, 2, 3, 4+$ using GEE with working independence.

	Est.	SE	Z
Intercept	−1.88	(0.36)	−5.28
Y_{it}	0.50	(0.17)	2.96
Y_{it-1}	0.08	(0.17)	0.46
X_{it-1}	0.92	(0.15)	6.26
X_{it-2}	0.31	(0.14)	2.15
X_{it-3}	0.34	(0.14)	2.42
Mean(X_{it-k}, $k \geq 4$)	1.74	(0.24)	7.27
Employed	−0.26	(0.13)	−2.01
Married	0.16	(0.12)	1.34
Maternal health	−0.19	(0.07)	−2.83
Child health	−0.09	(0.07)	−1.24
Race	0.03	(0.12)	0.21
Education	0.42	(0.13)	3.21
House size	−0.16	(0.12)	−1.28

MSCM: Marginal structural models using IPTW

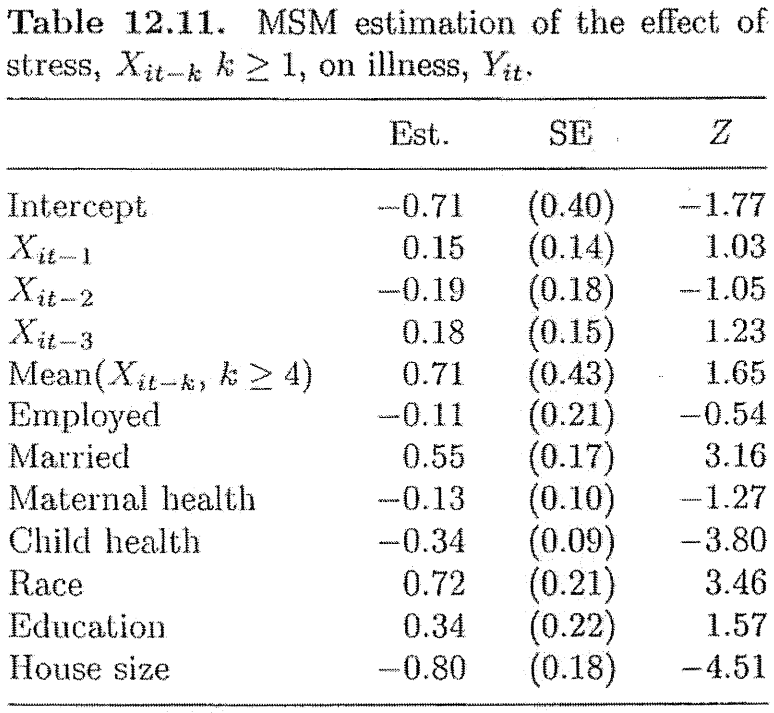


Table 12.11. MSM estimation of the effect of stress, X_{it-k} $k \geq 1$, on illness, Y_{it}.

	Est.	SE	Z
Intercept	−0.71	(0.40)	−1.77
X_{it-1}	0.15	(0.14)	1.03
X_{it-2}	−0.19	(0.18)	−1.05
X_{it-3}	0.18	(0.15)	1.23
Mean(X_{it-k}, $k \geq 4$)	0.71	(0.43)	1.65
Employed	−0.11	(0.21)	−0.54
Married	0.55	(0.17)	3.16
Maternal health	−0.13	(0.10)	−1.27
Child health	−0.34	(0.09)	−3.80
Race	0.72	(0.21)	3.46
Education	0.34	(0.22)	1.57
House size	−0.80	(0.18)	−4.51

Logistic regression
(GEE, working independence)

Causal effect of always stress vs. never stress:

Log odds ratio
$= 0.15 - 0.19 + 0.18 + 0.71 = 0.85$

s.e. = .4254 $P = 0.046$

so marginally significant causal effect of maternal stress on child health

MSCM: Comparison of four approaches

Effect of stress on illness at end *(always stress vs. never stress)*	*log odds ratio*
ignoring previous illness (ignoring confounders)	1.38
conditioning on previous illness (conditioning on intermediate variables)	0.50
g-computation	0.80
marginal structural model	0.85, s.e. 0.43