

Setup

- A common set of discrete observation times $t = 1, \dots, T$.
- Y_{it} is the response of subject i at time t .
- X_{it} is a time-varying covariate and \mathbf{z}_i a vector of baseline or time-invariant variables.
- We assume that X_{it} is measured immediately *after* Y_{it} and therefore cannot be a (direct) *cause* of the response at time t .

Exogenous and exogenous covariate

- In survival analysis: internal and external covariate
- In econometrics literature: determined by factors within/outside the system.

Defination

A covariate process is **exogenous** with respect to the outcome process if the covariate at time t is conditionally independent of all preceding response measurements, that is,

$$f(x_{it} | \mathcal{H}_i^Y(t), \mathcal{H}_i^X(t-1), \mathbf{z}_i) = f(x_{it} | \mathcal{H}_i^X(t-1), \mathbf{z}_i)$$

where

$$\begin{aligned}\mathcal{H}_i^Y(t) &= \{Y_{i1}, Y_{i2}, \dots, Y_{it}\} \\ \mathcal{H}_i^X(t-1) &= \{X_{i1}, X_{i2}, \dots, X_{it-1}\}\end{aligned}$$

Otherwise the covariate process is said to be **endogenous**.

- Scheduled observation time, t_{ij} is a time-dependent covariate and is exogenous.
- In crossover trials, the treatment sequence is prescribed and is exogenous.
- After organ transplantation, the dosage of immunosuppressant medication typically depends on the symptoms of graft rejection and is endogenous.
- In Indonesian Children Health Study, xerophthalmia status is time-varying. We can check for endogeneity by regressing X_{it} on both Y_{it}, Y_{it-1}, \dots , and X_{it-1}, \dots

ICHS: Empirically Checking Endogeneity

```
> xerop <- read.table ("../data/xerop.data",
+                       col.names = c("id", "RI", "intercept", "age",
+                                       "xero", "cos.time", "sin.time",
+                                       "sex", "height.age", "stunted", "time",
+                                       "base.age", "season", "time.time"))
> xeropw <- reshape (xerop, direction = "wide",
+                    v.names = c("RI", "xero", "age",
+                                  "height.age"),
+                    idvar = c("id", "sex"), timevar = "time",
+                    drop = c("intercept", "cos.time", "sin.time",
+                              "stunted", "base.age", "season", "time.time"))
> summary (glm (xero.2 ~ xero.1 + RI.1, family = "binomial",
+               data = xeropw))
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.3159	0.4208	-7.879	3.29e-15	***
xero.1	3.4787	0.7798	4.461	8.16e-06	***
RI.1	0.5683	0.9083	0.626	0.532	

>

```
> summary (glm (xero.3 ~ xero.2 + xero.1 + RI.2 + RI.1,
+               family = "binomial", data = xeropw))
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.5624	0.5363	-6.643	3.07e-11	***
xero.2	2.7307	1.2118	2.253	0.0242	*
xero.1	2.8932	1.5765	1.835	0.0665	.
RI.2	-14.0037	1769.2577	-0.008	0.9937	
RI.1	1.8371	0.8607	2.134	0.0328	*

```
> summary (glm (xero.3 ~ xero.2 + RI.2 + RI.1,
+               family = "binomial", data = xeropw))
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.4294	0.5093	-6.733	1.66e-11	***
xero.2	3.9255	1.0062	3.901	9.57e-05	***
RI.2	-14.1367	1769.2577	-0.008	0.9936	
RI.1	1.6184	0.8729	1.854	0.0637	.

Exogeneity

- The assumption of exogeneity allows the factorization

$$\begin{aligned}
 f(\mathbf{x}_i, \mathbf{y}_i \mid \mathbf{z}_i; \boldsymbol{\theta}) &= \left[\prod_{t=1}^T f(y_{it} \mid \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-1), \mathbf{z}_i; \boldsymbol{\theta}) \right] \\
 &\times \left[\prod_{t=1}^T f(x_{it} \mid \mathcal{H}_i^X(t-1), \mathbf{z}_i; \boldsymbol{\theta}) \right] \\
 &= \mathcal{L}_Y(\boldsymbol{\theta}) \times \mathcal{L}_X(\boldsymbol{\theta})
 \end{aligned}$$

- We further assume that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are variation independent, that is, $(\boldsymbol{\theta}_1 \in \Theta_1) \times (\boldsymbol{\theta}_2 \in \Theta_2)$. Then the process X_{it} is called *strongly exogenous* for the parameter of interest $\boldsymbol{\theta}_1$ if

$$f(\mathbf{x}_i, \mathbf{y}_i \mid \mathbf{z}_i; \boldsymbol{\theta}) = \mathcal{L}_Y(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \times \mathcal{L}_X(\boldsymbol{\theta}_2).$$

Inference for $\boldsymbol{\theta}_1$ can be based on the likelihood conditional on \mathbf{x}_i without loss of information, and therefore we do not need specify a model for the process X_{it} .

- Exogeneity also implies that

$$E(Y_{it} | x_{i1}, \dots, x_{iT}, \mathbf{z}_i) = E(Y_{it} | x_{i1}, \dots, x_{it-1}, \mathbf{z}_i).$$

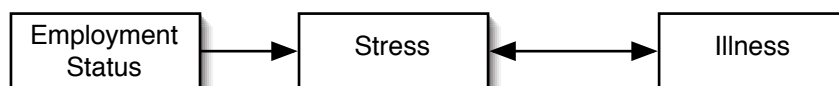
In fact, exogeneity implies a stronger conditional independence condition

$$Y_{it} \perp \{X_{it}, \dots, X_{iT}\} | \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-1).$$

- When a covariate is endogenous, we need be careful in choosing meaningful target of inference and valid methods of estimation.

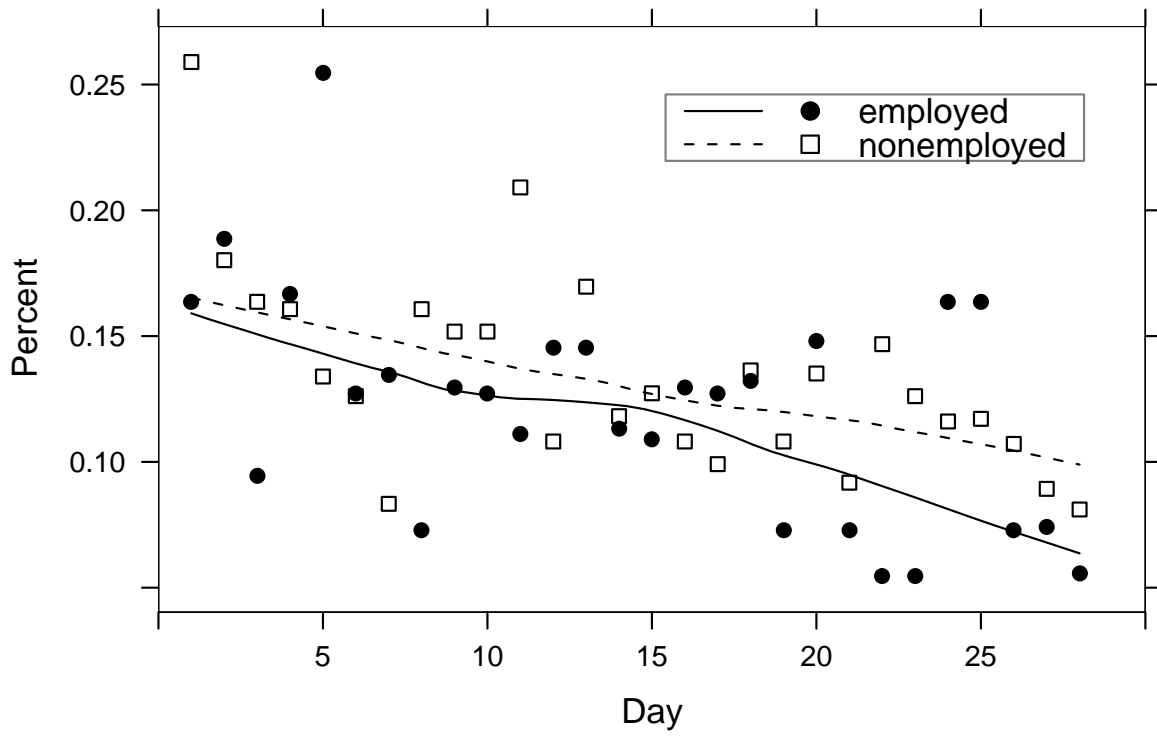
The MSCM Study

- An observational study of 167 preschool children aging 18 months to 5 years attending an inner-city pediatric clinic.
- Baseline covariates include employment status of the mothers. Four weeks of follow-up with daily measures of maternal stress and child illness.
- Questions of interest:
 1. Is there an association between maternal employment and stress?
 2. Is there an association between maternal employment and child illness?
 3. Is there evidence that maternal stress *causes* child illness?
- Stress may be in the causal pathway from employment to child illness so we do not adjust for illness in addressing Question 1. No adjustment for stress in Question 2.

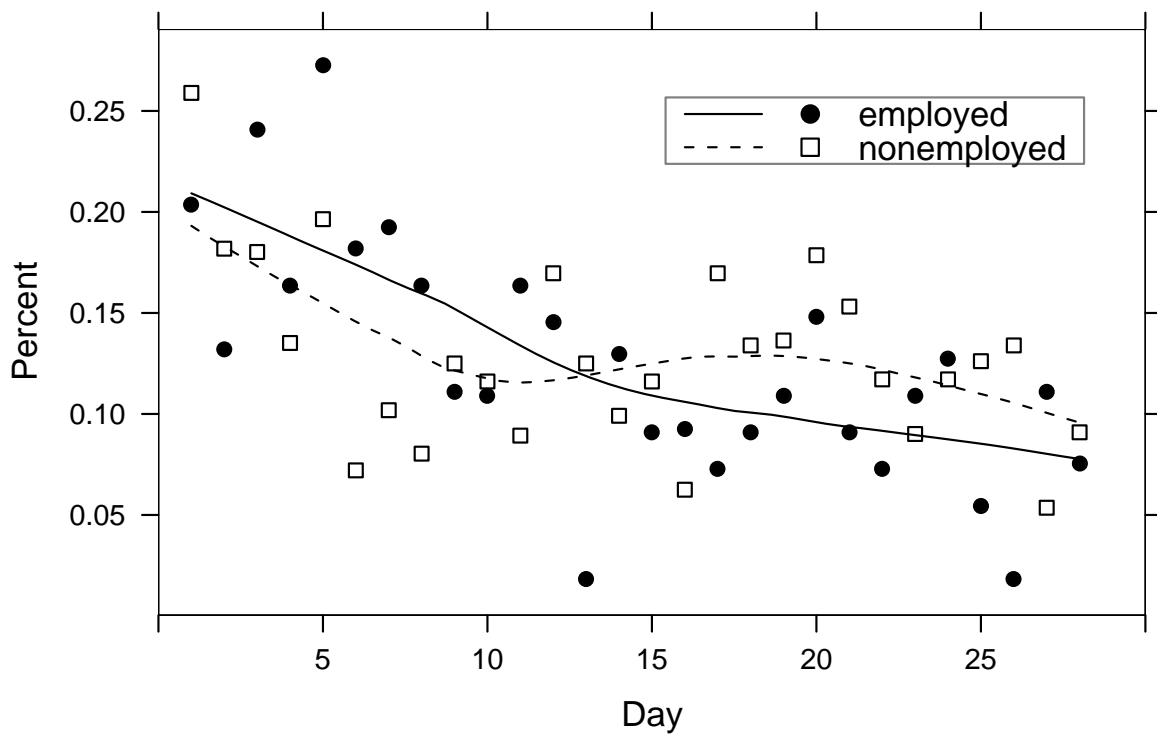


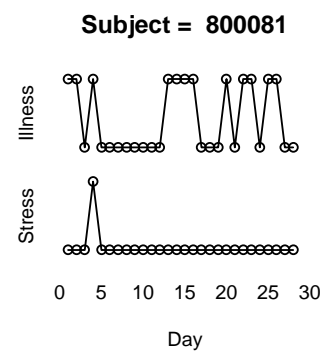
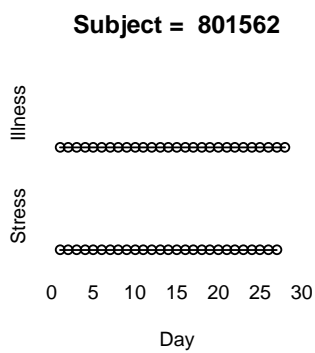
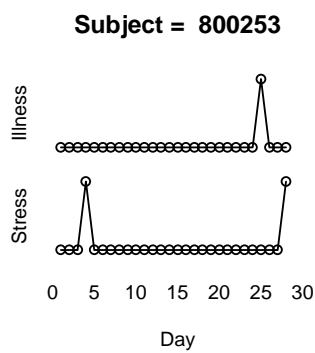
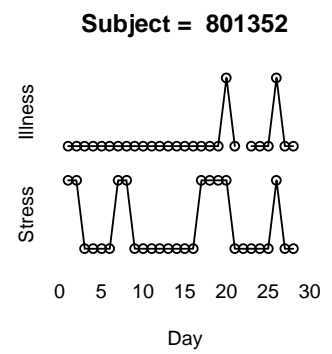
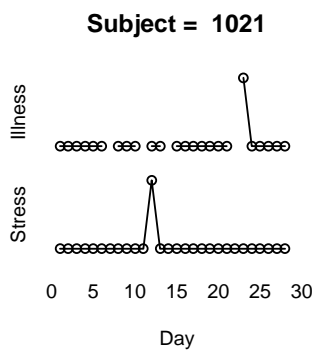
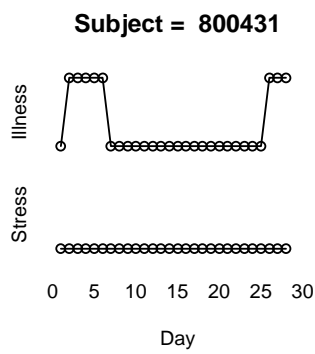
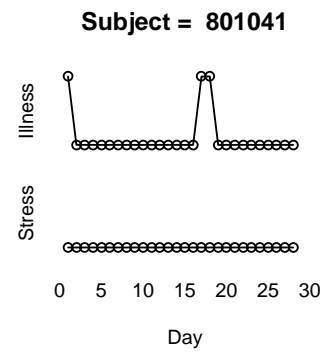
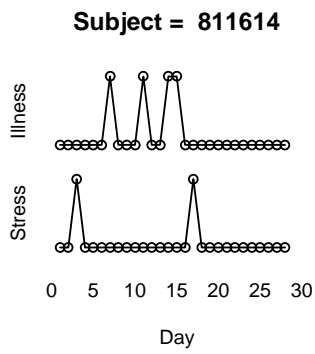
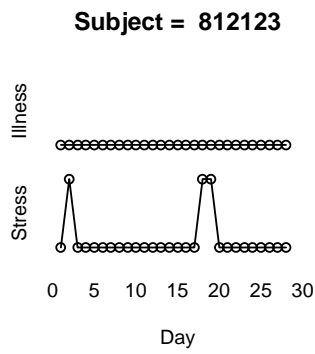
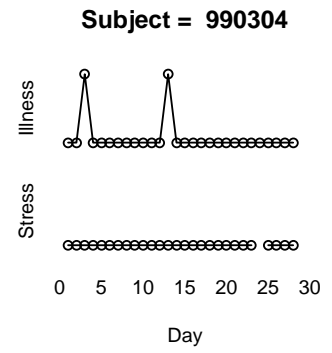
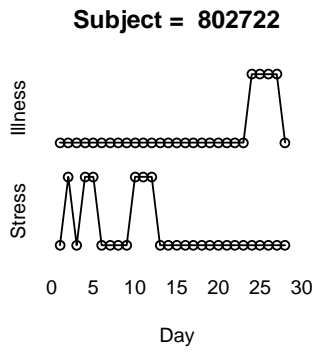
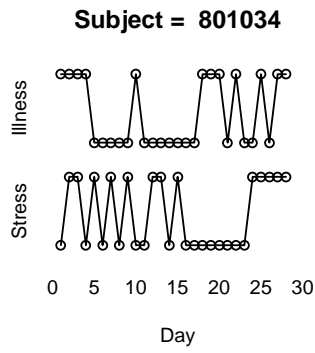
- Several aspects of Question 3:
 1. What is the cross-sectional association between stress on day t and illness on day t ? (marginal association)
 2. Does illness at day t depend on prior stress measured on day $(t - k)$?
 3. What are the factors that influence maternal stress on day t ? Does child illness on day $(t - k)$ for $k = 0, 1, 2, \dots$ predict maternal stress? (endogeneity)

Child Illness



Maternal Stress





Full and Partly Conditional Means

- For a stochastic covariate process X_{it} and an outcome process Y_{it} , we may be interested in:

Concurrent association: $E(Y_{it} | X_{it})$

Lagged association: $E(Y_{it} | X_{it-k} \text{ for some } k > 0)$

Exposure history: $E(Y_{it} | X_{i1}, X_{i2}, \dots, X_{it-1})$

Cumulative exposure: $E(Y_{it} | X_{it}^* = \sum_{s < t} X_{is})$

Entire process: $E(Y_{it} | X_{i1}, X_{i2}, \dots, X_{iT})$

- Define $E(Y_{it} | X_{i1}, X_{i2}, \dots, X_{iT})$ as the *full covariate conditional mean* and $E(Y_{it} | \text{subset}\{X_{i1}, X_{i2}, \dots, X_{iT}\})$ as a *partly conditional mean*, e.g. $E(Y_{it} | X_{it})$.
- Under the assumption of exogeneity and a finite covariate lag, the partly conditional mean $E(Y_{it} | X_{it-1}, X_{it-2}, \dots, X_{it-k})$ may equal the full conditional mean.
- When the covariate process is *endogenous*,

$$E(Y_{it} | X_{i1}, X_{i2}, \dots, X_{iT}) \neq E(Y_{it} | X_{i1}, X_{i2}, \dots, X_{it-1}),$$

the full conditional mean is not equal to the scientifically desired partly conditional mean.

- We need first identify which conditional mean is of interest, then identify of valid and efficient estimation methods.

Cross-Sectional Model and Estimation

- The GEE estimator of the regression parameter β is defined by

$$S_{\beta}(\beta, W) = \sum_{i=1}^m \left(\frac{\partial \mu_i}{\partial \beta} \right)^T W_i (Y_i - \mu_i) = \mathbf{0}.$$

- The GEE estimator is *consistent* if the estimating function is *unbiased*:

$$E(S_{\beta}(\beta, W)) = \mathbf{0}.$$

- **When is S unbiased?**

$$\begin{aligned} S_{\beta_X}(\beta, W) &= \sum_{i=1}^m \left[\begin{array}{c} \left(x_{i1} \right) \\ \vdots \\ \left(x_{in} \right) \end{array} \right]^T \begin{pmatrix} w_{i11}^* & \cdots & w_{i1n}^* \\ w_{i21}^* & \cdots & w_{i2n}^* \\ \vdots & \ddots & \vdots \\ w_{in1}^* & \cdots & w_{inn}^* \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu_{i1} \\ \vdots \\ Y_{in} - \mu_{in} \end{pmatrix} \\ &= \sum_{i=1}^m \left[\sum_{j=1}^n \sum_{k=1}^n x_{ij} w_{ijk}^* (Y_{ik} - \mu_{ik}) \right] \end{aligned}$$

where

$$w_{ijk}^* = \frac{\partial \eta_{ij}}{\partial \mu_{ij}} w_{ijk},$$

and w_{ijk} is the (j, k) element of the weight matrix W_i .

1. Consider the expectation of (dropping subscript p)

$$\begin{aligned} E(x_{ij} w_{ijk}^* (Y_{ik} - \mu_{ik})) &= E \left\{ E(x_{ij} w_{ijk}^* (Y_{ik} - \mu_{ik}) \mid \mathbf{X}_i) \right\} \\ &= E \left\{ x_{ij} w_{ijk}^* [E(Y_{ik} \mid \mathbf{X}_i) - \mu_{ik}] \right\} \end{aligned}$$

The EE is unbiased if the **full covariate conditional mean (FCCM)** assumption is satisfied, that is,

$$E(Y_{ik} \mid \mathbf{X}_i) = \mu_{ik} = E(Y_{ik} \mid X_{ik}).$$

(Pepe and Anderson, 1994)

On the other hand, if FCCM does not hold, the estimating function will likely be biased and result in inconsistent estimates for the cross-sectional mean structure.

2. However, if a diagonal weight matrix is used, then

$$\begin{aligned}
 S_{\beta_X}(\boldsymbol{\beta}, \mathbf{W}) &= \sum_{i=1}^m \left[\begin{pmatrix} x_{i1} \\ \vdots \\ x_{in} \end{pmatrix}^T \begin{pmatrix} w_{i11}^* & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{inn}^* \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu_{i1} \\ \vdots \\ Y_{in} - \mu_{in} \end{pmatrix} \right] \\
 &= \sum_{i=1}^m \left[\sum_{j=1}^n x_{ij} w_{ijj}^* (Y_{ij} - \mu_{ij}) \right]
 \end{aligned}$$

We have

$$\begin{aligned}
 \mathbf{E} (x_{ij} w_{ijj}^* (Y_{ij} - \mu_{ij})) &= \mathbf{E} \{ \mathbf{E} (x_{ij} w_{ijj}^* (Y_{ij} - \mu_{ij}) \mid X_{ij}) \} \\
 &= \mathbf{E} \{ x_{ij} w_{ijj}^* [\mathbf{E} (Y_{ij} \mid X_{ij}) - \mu_{ij}] \}
 \end{aligned}$$

which is always unbiased (FCCM is not required).

- **Conclusion:** the GEE estimator of β is consistent if
 1. The full covariate conditional mean (FCCM) assumption holds;
or
 2. A working independence weight matrix is used.
- We can always use working independence GEE to model $E(Y_{it} | X_{it})$. However we can not interpret the marginal association as causal effects. It is equally plausible that stress causes illness or that illness causes stress. In order to infer cause we need to address the temporal ordering of exposure and outcome.
- Example: Analysis of marginal association for MSCM study (DHLZ Table 12.4).

MSCM Study: Marginal Association

```

> m.ind <- gee (illness ~ stress + week + married + employed + chlth
+             + mhlth + race + education + housize,
+             data = mscm, id = id, family = binomial)
> m.ex <- gee (illness ~ stress + week + married + employed + chlth
+             + mhlth + race + education + housize,
+             data = mscm, id = id, corstr = "exchangeable",
+             family = binomial)
> m.ar1 <- gee (illness ~ stress + week + married + employed + chlth
+             + mhlth + race + education + housize,
+             data = mscm, id = id, corstr = "AR-M", Mv = 1,
+             family = binomial)
> res <- lapply (list (independent = m.ind,
+                     exchangeable = m.ex, ar1 = m.ar1),
+               function (x) (coef (summary (x))[2:3,c(1,4:5)]))
> options (digits = 3)
> res
$independent
      Estimate Robust S.E. Robust z
stress    0.668    0.1394    4.79
week    -0.178    0.0545   -3.26

$exchangeable
      Estimate Robust S.E. Robust z
stress    0.532    0.1363    3.91
week    -0.182    0.0545   -3.34

$ar1
      Estimate Robust S.E. Robust z
stress    0.379    0.1220    3.10
week    -0.206    0.0539   -3.82
> m.ex$working.correlation[1,2]
[1] 0.0615
> m.ar1$working.correlation[1,2]
[1] 0.39

```

Lagged Covariates

Scenarios of using lagged covariates:

- A single lagged covariate, X_{it-k} .
- In smoking studies, use cumulative exposure, $X_{it}^* = \sum_{s < t} X_{is}$, pack-years, as the summary of exposure history.
- Use the entire covariate history X_{i1}, \dots, X_{it} as potential predictors for Y_{it} .
- Use only a subset of recent covariates, $X_{it-1}, \dots, X_{it-k}$ when an acute effect is assumed.

In the latter two cases, the lagged covariates can be highly correlated.

A Single Lagged Covariate

In a *partly conditional* GLM

$$g(\text{E}(Y_{it} | X_{is}, \mathbf{Z}_i)) = \beta_0(t, s) + X_{is}\beta_1(t, s) + \mathbf{Z}_i^T \beta_2(t, s).$$

- It is called “partly” because only one covariate time is included as predictor.
- β_1 may depend on both the response time t and the covariate time s .
- Often β_1 is assumed to be a function of only $(t - s)$. Given a particular functional form for $\beta_1(t, s)$ we can use GEE with working independence to estimate the parameters.

Example: MSCM data

$$\text{logit}(\text{E}(Y_{it} | X_{it-k}, \mathbf{Z}_i)) = \beta_0(k) + \beta_1(k)X_{it-k} + \mathbf{Z}_i^T \beta_2(k)$$

- Fitting separate models for $k = 1, 2, \dots, 7$.
- Assume β_2 constant for k , and some parametric function (of $k = 1, \dots, 29$), for $\beta_0(k)$ and $\beta_1(k)$.

```
> mscm$s7 <- shift (mscm$s6, id = mscm$id)
> il.lags <- list ()
> for (ii in paste ("s", 1:7, sep = "")) {
+   il.lags[[ii]] <- gee (illness ~ mscm[,ii] + week + married +
+                       employed + chlth + mhlth + race +
+                       education + housize, data = mscm,
+                       id = id, family = binomial)
+ }
> il.lags.coef <- do.call ("rbind",
+                          lapply (il.lags, function (x) {
+                            coef (summary (x))[2,c(1,4)]
+                          }))
> il.lags.coef <- data.frame (il.lags.coef)
> names (il.lags.coef) <- c("beta", "se")
> with (il.lags.coef, {
+   plot (beta, xlim = c(0, 30), ylim = c(-0.6, 0.8),
+        xlab = "Time lag (days)",
```

```
+         ylab = "Coefficient (log odds ratio)",  
+         main = "MSCM Study: Lag coefficient function")  
+   abline (h = 0, lty = 2)  
+   errbar (1:7, beta, beta - se * 1.96, beta + se * 1.96,  
+         add = TRUE)  
+ })
```

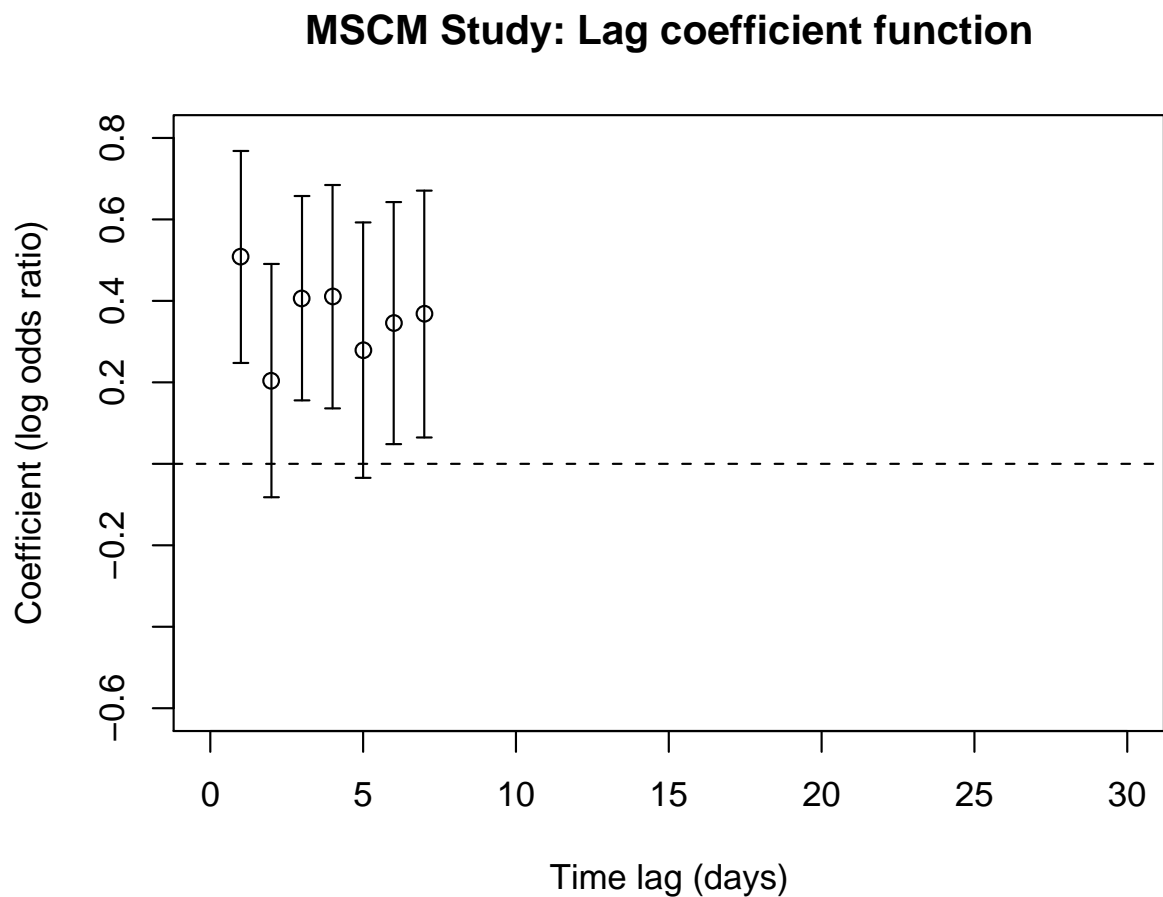


Figure 1: DHLZ Figures 12.3

Multiple Lagged Covariates

Sometimes we may need models with multiple lagged variables:

$$g(\mathbb{E}(Y_{it} | \mathcal{H}_{it-1}^X, \mathbf{Z}_i)) = \beta_0 + \beta_1 X_{it-1} + \cdots + \beta_L X_{it-L} + \mathbf{Z}_i^T \boldsymbol{\alpha},$$

for a finite lag L . The above model may be over-specified, however and lead to estimates of lagged covariate effects that are difficult to interpret.

We can further assume that the lagged coefficients $(\beta_1, \beta_2, \dots, \beta_L)$ follow a lower order smooth parametric function (**distributed lag models**):

$$\beta_j = \gamma_0 + \gamma_1 B_1(j) + \cdots + \gamma_p B_p(j)$$

where $\mathbf{B}(j)$ is an appropriate basis vector. For polynomial model $B_l(j) = j^l, l = 0, \dots, p$.

- We have

$$\boldsymbol{\beta}_{(L+1) \times 1} = \mathbf{B}_{(p+1) \times (L+1)}^T \boldsymbol{\gamma}_{(p+1) \times 1}.$$

and the regression can be done using transformed variables $\mathbf{X}_i \mathbf{B}^T$

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\alpha} = (\mathbf{X}_i \mathbf{B}^T) \boldsymbol{\gamma} + \mathbf{Z}_i \boldsymbol{\alpha}.$$

- The coefficient model allows parsimonious modeling.
- We need choose the appropriate L (number of lagged covariates) and p (order of the coefficient model).

MSCM data (DHLZ Figures 12.4 and Table 12.5)

Saturated model

```
> iT5.sat <- gee (illness ~ s1 + s2 + s3 + s4 + s5 + s6 + s7 +
+               week + married + employed +
+               chlth + mhlth + race + education + housize,
+               data = mscm, id = id,
+               family = binomial)
> iT5sat.e <- coef (summary (iT5.sat))[2:8, c(1,4)]
> colnames (iT5sat.e) <- c("Beta", "SE")
> rownames (iT5sat.e) <- c("s1","s2","s3","s4","s5","s6","s7")
> iT5sat.e
```

	Beta	SE
s1	0.34329239	0.1561614
s2	-0.05151545	0.1509723
s3	0.17934857	0.1363666
s4	0.26664393	0.1373006
s5	0.23557409	0.1463803
s6	0.20839598	0.1421672
s7	0.26424049	0.1367730

Natural spline with 4 degrees of freedom

```
B <- ns (1:7, knots = c(3, 5), intercept = TRUE)
Xstar <- as.matrix (mscm[,paste("s",1:7,sep="")]) %*% B
iT5.ns <- gee (illness ~ Xstar +
              week + married + employed +
              chlth + mhlth + race + education + housize,
              data = mscm, id = id,
              family = binomial)
tmp.c <- iT5.ns$coef[2:5]
tmp.v <- iT5.ns$robust[2:5, 2:5]
iT5ns.e <- cbind (B %*% tmp.c,
                 sqrt (diag (B %*% tmp.v %*% t(B))))
> colnames (iT5ns.e) <- c("Beta", "SE")
> rownames (iT5ns.e) <- c("s1","s2","s3","s4","s5","s6","s7")
> iT5ns.e
```

	Beta	SE
s1	0.2517556	0.15824540
s2	0.1450647	0.11694727
s3	0.1088346	0.12277381
s4	0.1808002	0.09622847
s5	0.2677936	0.11855602
s6	0.2741420	0.11505092
s7	0.2250573	0.13545063

Monotone model, $\beta_j = \gamma_0 + \gamma_1 \cdot (1/j)$

```
> M <- matrix(c(rep(1,7),1,1/2,1/3,1/4,1/5,1/6,1/7),ncol=2)
> M
      [,1]      [,2]
[1,]    1 1.0000000
[2,]    1 0.5000000
[3,]    1 0.3333333
[4,]    1 0.2500000
[5,]    1 0.2000000
[6,]    1 0.1666667
[7,]    1 0.1428571
> Xstar.mono <- as.matrix (mscm[,paste("s",1:7,sep="")]) %*% M
> iT5.mono <- gee (illness ~ Xstar.mono +
+                 week + married + employed +
+                 chlth + mhlth + race + education + housize,
+                 data = mscm, id = id,
+                 family = binomial)
> tmp.c <- iT5.mono$coef[2:3]
> tmp.v <- iT5.mono$robust[2:3, 2:3]
> iT5mono.e <- cbind (M %*% tmp.c,
+                    sqrt (diag (M %*% tmp.v %*% t(M))))
> colnames (iT5mono.e) <- c("Beta", "SE")
> rownames (iT5mono.e) <- c("s1","s2","s3","s4","s5","s6","s7")
> iT5mono.e
      Beta      SE
s1 0.2224116 0.17224088
s2 0.2110720 0.07948940
s3 0.2072921 0.07137674
s4 0.2054021 0.07558084
s5 0.2042682 0.08047788
s6 0.2035122 0.08455853
s7 0.2029722 0.08781513
```

```

iT5sat.e <- data.frame (iT5sat.e)
with (iT5sat.e, {
  plot (Beta, xlim = c(0.5, 7.5), ylim = c(-0.5, 0.8), xaxt="n", pch=1,
        xlab = "Time lag (days)", ylab = "Coefficient (log odds ratio)",
        main = "MSCM Study: Lag coefficient function")
  axis(side=1,at=1:7,lab=c("1","2","3","4","5","6","7"))
  abline (h = 0, lty = 2)
  errbar (1:7, Beta, Beta - SE * 1.96, Beta + SE * 1.96,
        add = TRUE)})
iT5ns.e <- data.frame (iT5ns.e)
with (iT5ns.e, {
  points (1:7+0.1,Beta, pch=12)
  lines (1:7+0.1,Beta, lty = 3)
  abline (h = 0, lty = 2)
  errbar (1:7+0.1, Beta, Beta - SE * 1.96, Beta + SE * 1.96,
        add = TRUE)})
iT5mono.e <- data.frame (iT5mono.e)
with (iT5mono.e, {
  points (1:7+0.2,Beta, pch=6)
  lines (1:7+0.2,Beta, lty = 4)
  abline (h = 0, lty = 2)
  errbar (1:7+0.2, Beta, Beta - SE * 1.96, Beta + SE * 1.96,
        add = TRUE)})
legend(5,-0.2,c("Saturated","Spline","Monotone"),pch=c(1,12,6),cex=1.5)

```

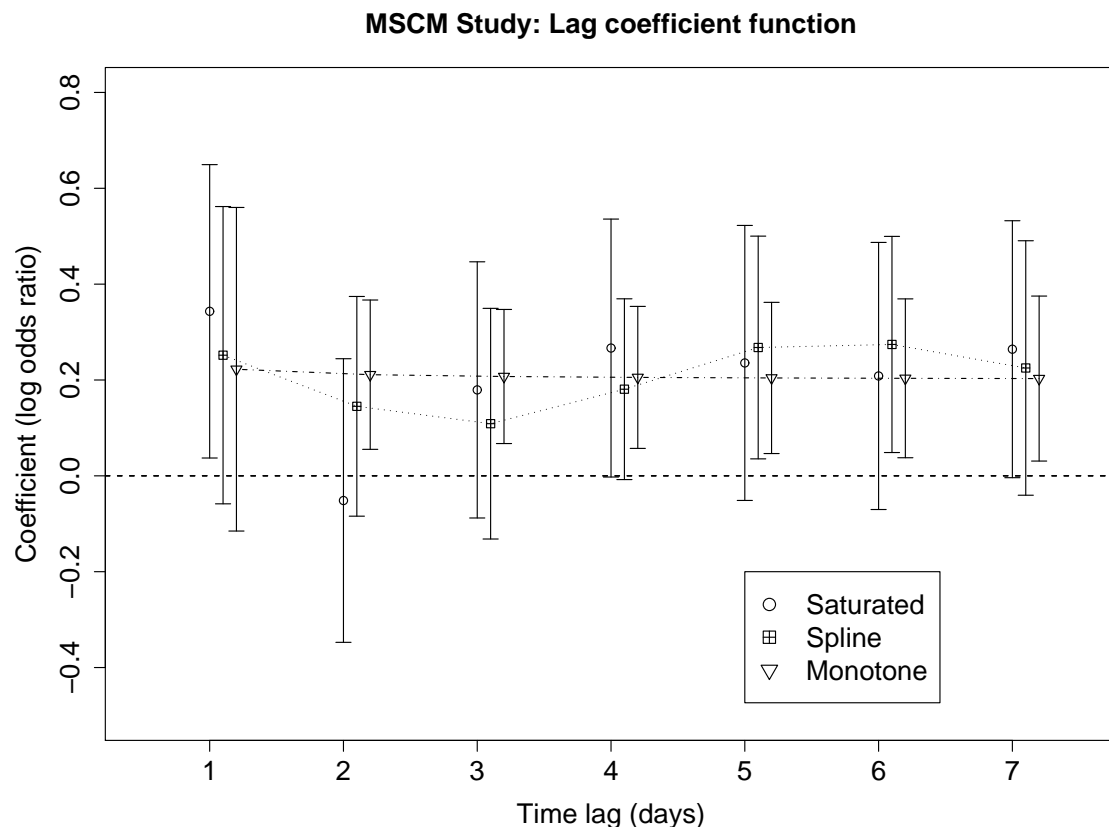
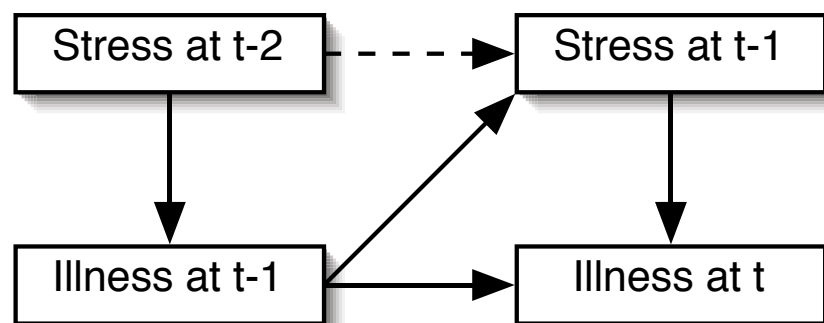


Figure 2: DHLZ Figures 12.4

- Each of these models shows an association between maternal stress in the previous 7 days and current child illness.
- We cannot use AIC or BIC to compare the adequacy of different distributed lag models.
- However, we can evaluate the **predictive accuracy** of each model (Harrell et al., 1984).
 1. Delete individual subjects
 2. Refit the model and compare observed and fitted outcome vectors
 3. Use c-index (area under the ROC curve) as a global summary of model accuracy.
- The c-index for the saturated model: 64.1%, the spline model: 63.8%, and the monotone model: 64.2%. They provide almost identical predictive accuracy.

Time-Dependent Confounders

- When a covariate is associated with both the exposure of interest and the outcome of interest:
 - a *confounder* if ignoring it leads to biased exposure effect estimates.
 - an *intermediate variable* if it is in the causal pathway from exposure to outcome.
- We should adjust for the confounders but not the intermediate variables.
- In longitudinal studies, the time-dependent variable can be both a confounder and an intermediate variable.
- MSMC data: endogeneity of stress can be tested by regressing stress at time t on stress at time $t - 1$, $t - 2$, \dots , and on child illness at time t and $t - 1$ (DHLZ Table 12.7).



Causal Inference

- **Counterfactual/potential outcome**

- Define $Y_{it}^{(\mathbf{x}_t)}$ as the outcome for subject i at time t that would be observed if a given covariate sequence $\mathbf{x}_t = (x_1, x_2, \dots, x_{t-1})$ was in force.
- For a given subject we can only observed a single outcome at time t for a specific covariate sequence \mathbf{x}_t^* . All other possible outcomes $Y_{it}^{(\mathbf{x}_t)}$ for $\mathbf{x}_t \neq \mathbf{x}_t^*$ are not observed.

- The **causal effect** of $\mathbf{x}_t = (1, 1, \dots, 1)$ versus $\mathbf{x}_t = (0, 0, \dots, 0)$ at time t is defined as

$$\delta_t = E [Y_{it}^{(1)} - Y_{it}^{(0)}] = E [Y_{it}^{(1)}] - E [Y_{it}^{(0)}].$$

- For each i , at most only one of $Y_{it}^{(0)}$ or $Y_{it}^{(1)}$ is observed, or neither is observed.
- The causal effect refer to the effect of interventions in the entire population rather than possibly selected, observed subjects.
- For a randomized trial with full compliance, $E[Y_{it} | \mathbf{x}_t = \mathbf{1}]$ is an unbiased estimate of $E [Y_{it}^{(1)}]$. However, in other cases, $E[Y_{it} | \mathbf{x}_t = \mathbf{1}]$ is not necessarily an unbiased estimate of $E [Y_{it}^{(1)}]$. Note that $E[Y_{it} | \mathbf{x}_t]$ is the conditional expectation given the observed treatment, while $E [Y_{it}^{(1)}]$ is the population average if the treatment is enforced to the whole population.
- With additional baseline covariates \mathbf{Z} , we define

$$\mu_t^{(\mathbf{x}_t)}(\mathbf{z}) = E (Y_{it}^{(\mathbf{x}_t)} | \mathbf{Z} = \mathbf{z})$$

as the average outcome that would be observed at time t within the sub-population defined by $\mathbf{Z} = \mathbf{z}$ if all subjects following the treatment/exposure path \mathbf{x}_t .

- The cause effect of continuous exposure is defined as

$$\delta_T(\mathbf{z}) = \mu_T^{(1)}(\mathbf{z}) - \mu_T^{(0)}(\mathbf{z})$$

where T represents the end of the study.

- Robins et al (1999) formalized the definition of **no unmeasured confounders**.

- In the absence of unmeasured confounders, association implies (defines) causation.
- Formally, the assumption of “no unmeasured confounders” or **sequential randomization assumption** for longitudinal data is

$$\left\{ (Y_s^{(1)}, Y_s^{(0)}); s = t + 1, t + 2, \dots, T \right\} \perp X_t \mid \mathcal{H}^X(t-1), \mathcal{H}^Y(t)$$

- That is, the exposure at time t is independent of any future potential outcomes at time $s > t$, given information through time t .
 - One example where this assumption would be violated: a physician prescribes treatment to the sickest patients.
 - This assumption (and assumptions needed in causal inference in general) are not empirically verifiable.
- Pearl (2000) argued that causal statements require new notation

$$E[Y_2 | \text{do}(X_0 = 1, X_1 = 1)]$$

to denote the outcome for the entire population (which is the same as $\mu_2^{(1)}$).

- His notation emphasizes the fact the we are interested in an average outcome after assignment of treatment rather than the average of outcomes in subgroups after simply observing the treatment status.

- Estimation approaches for causal inference (mostly due to Jim Robins):
 - g -computation algorithm.
 - Semiparametric g -estimation of structured nested mean models (SNMM).
 - Marginal structural models (MSM) using inverse probability of treatment weights (IPTW).

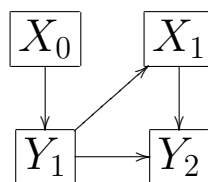
Artificial Example

DHLZ considered an artificial data with two time points and binary covariate X and outcome Y . Data is generated from the following sequential conditional models:

$$\text{logit } E(Y_1 | X_0 = x_0) = -0.5 - 0.5x_0, \quad (1)$$

$$\text{logit } E(X_1 | Y_1 = y_1, X_0 = x_0) = -0.5 + 1.0y_1 \quad (2)$$

$$\text{logit } E(Y_2 | X_1 = x_1, X_0 = x_0, Y_1 = y_1) = -1.0 + 1.5y_1 - 0.5x_1 \quad (3)$$



- Marginal expectations

$$\Pr(Y_1 = 1 | x_0 = 0) = \frac{e^{-0.5}}{1 + e^{-0.5}} = 0.38$$

$$\Pr(Y_1 = 1 | x_0 = 1) = \frac{e^{-1}}{1 + e^{-1}} = 0.27$$

$$\Pr(X_1 = 1 | Y_1 = 0, x_0) = \frac{e^{-0.5}}{1 + e^{-0.5}} = 0.38$$

$$\Pr(X_1 = 1 | Y_1 = 1, x_0) = \frac{e^{0.5}}{1 + e^{0.5}} = 0.62$$

$$\Pr(Y_2 = 1 | X_0 = 1, X_1 = 1) = 0.30$$

$$\Pr(Y_2 = 1 | X_0 = 0, X_0 = 0) = 0.36.$$

- The marginal distribution of Y_2 conditional on (X_0, X_1) is

$$\text{logit } E(Y_2 | x_0, x_1) = -0.56 - 0.13x_0 - 0.10x_1 - 0.04x_0x_1. \quad (4)$$

- The marginal model above reflects an effect of X_0 in addition to that of X_1 on Y_2 .
 - Since Y_1 is correlated with both X_1 and X_2 , the observed marginal effects do not account for the **confounder** Y_1 , and hence do not reflect the causal effect of treatment.
 - At the same time, Y_1 is also an **intermediate variable**. If we adjust for Y_1 (like in model (3)) the effect of X_0 will be masked (no effect).
 - In summary, no standard regression methods can be used to obtain causal statements.
- The potential outcomes are (see DHLZ Table 12.8 for details)

$$\mu_2^{(1)} = 0.27$$

$$\mu_2^{(0)} = 0.40$$

- The *causal* odds ratio is 0.54 (a 46% decrease in odds) versus the *observed* odds ratio of 0.76 (a smaller, 24% decrease in odds).

g-Computation

- In this approach, causal effect estimates are obtained from estimates of the observed response transition model under the assumption of no unmeasured confounders.
- The likelihood for a binary response Y_{i1}, \dots, Y_{iT} and a binary time-dependent covariate X_{i0}, \dots, X_{iT-1} can be decomposed into

$$\begin{aligned} \mathcal{L} &= \prod_{t=1}^T \Pr(Y_{it} | \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-1), \mathbf{Z}_i) \\ &\quad \times \prod_{t=1}^T \Pr(X_{it-1} | \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-2), \mathbf{Z}_i) \\ &= \mathcal{L}_Y \times \mathcal{L}_X \end{aligned}$$

- Unknown parameters in the response transition model \mathcal{L}_Y and covariate transition model \mathcal{L}_X can be estimated using maximum likelihood.
- Under the assumption of no unmeasured confounders

$$\begin{aligned} \Pr(Y_{it}^{(\mathbf{x}_t)} = 1 | \mathbf{Y}_{it-1}^{(\mathbf{x}_t)}, \mathbf{Z}_i) &= \\ &= \Pr(Y_{it} = 1 | \mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t-1) = \mathbf{x}_t, \mathbf{Z}_i). \end{aligned}$$

(*Proposition 2* in Robins et al, 1999)

- We obtain the marginal distribution of the outcome at the final time t , $\Pr(Y_{it}^{(\mathbf{x}_t)} | \mathbf{Z}_i)$, by using the conditional probabilities to obtain the joint probability for $Y_{i1}^{(\mathbf{x}_t)}, \dots, Y_{it}^{(\mathbf{x}_t)}$ and the summing over all possible intermediate paths for the first $t - 1$ outcomes.

$$\begin{aligned} \mu^{(\mathbf{x}_t)}(\mathbf{z}) &= \Pr[Y_{it}^{(\mathbf{x}_t)} = 1 | \mathbf{Z}_i = \mathbf{z}], \\ &\Pr[Y_{it}^{(\mathbf{x}_t)} | \mathbf{Z}_i = \mathbf{z}] \\ &= \sum_{\mathcal{Y}_{t-1}} \Pr[Y_{it}^{(\mathbf{x}_t)}, Y_{it-1}^{(\mathbf{x}_t)}, \dots, Y_{i1}^{(\mathbf{x}_t)} | \mathbf{Z}_i = \mathbf{z}] \\ &= \sum_{\mathcal{Y}_{t-1}} \prod_{s=1}^t \Pr[Y_{is}^{(\mathbf{x}_t)} | Y_{is-1}^{(\mathbf{x}_t)}, \dots, Y_{i1}^{(\mathbf{x}_t)}; \mathbf{Z}_i = \mathbf{z}] \\ &= \sum_{\mathcal{Y}_{t-1}} \prod_{s=1}^t \Pr[Y_{is} | \mathcal{H}^Y(s-1) = \mathbf{y}_{s-1}, \mathcal{H}^X(s-1) = \mathbf{x}_{s-1}; \mathbf{Z}_i = \mathbf{z}] \end{aligned}$$

- This is a special case of the *g-computation algorithm*. In general, this *g-computation* formula can be evaluated using Monte Carlo.
- One advantage of *g-computation* is that a model for the treatment is not required.
- One limitation is that no direct regression model parameter represents the null hypothesis of no causal effect, thus does not facilitate formal testing.

Inverse Probability of Treatment Weights (IPTW)

- The **marginal structured models** (MSMs) (Robins, 1998; Hernan et al., 2001) allows direct regression modeling of causal effects but requires a model for the covariate process.
- The MSM models the expectation of the counterfactual outcome directly.

$$\begin{aligned}\mu_t^{(\mathbf{x}_t)}(\mathbf{z}) &= \text{E} (Y_{it}^{(\mathbf{x}_t)} \mid \mathbf{Z}_i = \mathbf{z}) \\ g(\mu_t^{(\mathbf{x}_t)}) &= \beta_0 + \boldsymbol{\beta}_1^T \mathbf{X}_{it}^* + \boldsymbol{\beta}_2^T \mathbf{Z}_i\end{aligned}$$

where \mathbf{X}_{it}^* is a function of the covariate history.

- The parameter $\boldsymbol{\beta}_1$ can be used to quantify and test the causal effect of exposure.
- We cannot use GEE directly because the prior responses are confounder. We also do not want to control for prior responses because they are also intermediate variables.

- Estimation is done using IPTW where weights are used to construct a **pseudo-population** the previous outcomes are no longer confounders.

- The **stabilizing weights** are defined by

$$SW_i(t) = \prod_{s < t} \frac{\Pr(X_{is} = x_{is} \mid \mathcal{H}_i^X(s-1), \mathbf{Z}_i)}{\Pr(X_{is} = x_{is} \mid \mathcal{H}_i^Y(s-1), \mathcal{H}_i^X(s-1), \mathbf{Z}_i)}$$

- In our hypothetical example,

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

- The weights compare the probability of the treatment received through time $t - 1$ conditional only on the treatment histories to the probability of treatment received conditional on both treatment and outcome histories.
- The weights would be 1 if exogeneity is satisfied and also serves as a measure of endogeneity.
- Weights need to be estimated by choosing and fitting models for both the numerator and denominator. Correct modelling for the denominator is needed for consistent parameter estimation, while the model choice for the numerator only impacts efficiency not validity.
- GEE with working independence and the estimated weights $\widehat{SW}_i(t)$ can be used to estimate β_1 with empirical sandwich variance estimates.

MSCM Data

- Marginal association model (i.e. the saturated model with multiple lag covariates, DHLZ Table 12.5)
- Transition model with lagged illness (response, order = 2: Y_{it-k} , $k = 1, 2$) and lagged maternal stress (order = 3: X_{it-k} , $k = 1, 2, 3$). GEE with independence working correlation matrix (Table 12.9, DHLZ).
- g -computation: use the previous transition model for Y_{it} , choose a final end-point time, identify a covariate value of interest (for Z_i), and generate Markov chains with stress controlled at 1 for all times, and controlled at 0 for all times.
- MSM with IPTW: estimate $SW_i(t)$ based on models for the exposure process X_{it} , then use weighted GEE with independence correlation (Table 12.11, DHLZ).

	log odds ratio ¹
Marginal association	1.38
Transition model	0.50
g -computation	0.80
MSM	0.85 (p=0.046)

¹: for those employed=0, married=0, maternal and child health=4, race=0, education=0, and house size=0.

Further Reading

- Chapter 12 of DHLZ.