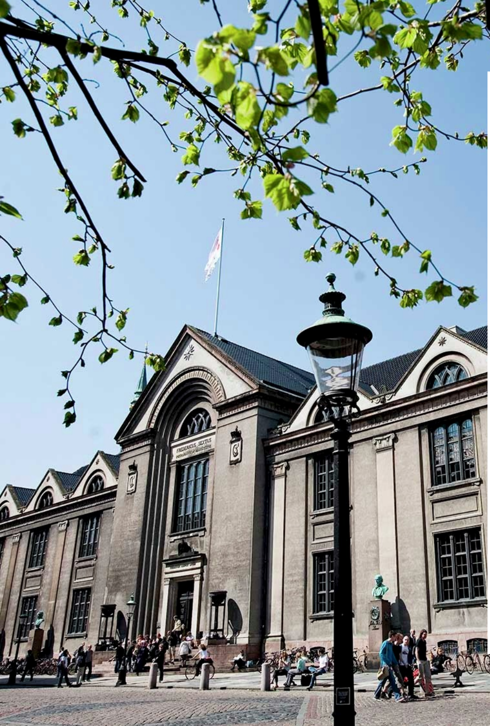

Sequential outcome regression

TMLE for breakfast May 23

Emilie Wessel
Section of Biostatistics

UNIVERSITY OF COPENHAGEN

Disclaimer: This talk is *not* about TMLE.

Disclaimer: This talk is *not* about TMLE.

Aim: Provide you with the idea of sequential regression.

Disclaimer: This talk is *not* about TMLE.

Aim: Provide you with the idea of sequential regression.

$\rightarrow$ **Maximum likelihood based G-computation estimation**

Notation

Consider the random vector, $\mathbf{X} = (L, A, Y)$. Let P_X denote the joint probability measure. We assume that $\mathsf{P}_X \in \mathcal{P}$ where $\mathcal{P}$ is an otherwise unspecified model which is dominated by a σ-finite measure, μ. With the notation,

$$
\begin{aligned}
Q_Y(\mathrm{d}y \mid a, l) &= \mathsf{P}_{Y|A,L}(\mathrm{d}y \mid A = a, L = l) \\
G_A(\mathrm{d}a \mid l) &= \mathsf{P}_{A|L}(\mathrm{d}a \mid L = l), \\
Q_L(\mathrm{d}l) &= \mathsf{P}_L(\mathrm{d}l),
\end{aligned}
$$

the joint distribution P_X factorizes

$$
\mathsf{P}_X(dx) = Q_Y(\mathrm{d}y \mid a, l) G_A(\mathrm{d}a \mid l) Q_L(\mathrm{d}l).
$$

Notation

For any measurable function f of the data we use the notation for the expected value of f under P_X:

$$\mathsf{P}_X f = \int f(x)\, \mathsf{P}_X(\mathrm{d}x).$$

Using the factorization,

$$\mathsf{P}_X(dx) = Q_Y(\mathrm{d}y \mid a, l) G_A(\mathrm{d}a \mid l) Q_L(\mathrm{d}l),$$

we extend this compact operator notation to conditional distributions and write

$$\mathsf{P}_X f = Q_L G_A Q_Y f = \iiint f(x) Q_Y(\mathrm{d}y \mid a, l) G_A(\mathrm{d}a \mid l) Q_L(\mathrm{d}l).$$

Longitudinal data structures

For K time points, $\{0 = t_0 < t_1 < \cdots < t_K\}$, we consider longitudinal data given by

$$\mathbf{X} = (L(0), A(0), Y(1), L(1), A(1), \ldots, Y(K)),$$

s.t. $Y(k)$ is the value of a stochastic outcome process $Y(t)$ at time t_k.

The state space of the process Y is $\{0, 1\}$, i.e.,

$$\begin{aligned} Y(t) &= 0 : \text{event has not yet occurred at time } t \\ Y(t) &= 1 : \text{event has occurred in the time interval } [0, t] \end{aligned}$$

Ex. 1: One time point, $K = 1$, no censoring

Consider the data, $\mathbf{X} = (L(0), A(0), Y(1)) \sim \mathsf{P}_X \in \mathcal{P}$.

Our target parameter is the intervention-specific ATE defined by

$$\psi(\tilde{\mathsf{P}}) = \tilde{\mathsf{P}}(Y^{(1)}(1) = 1) - \tilde{\mathsf{P}}(Y^{(0)}(1) = 1),$$

where $(L(0), Y^{(1)}(1), Y^{(0)}(1)) \sim \tilde{\mathsf{P}} \in \tilde{\mathcal{P}}$ is the counterfactual data under interventions, where the treatment is set to either one or zero.

Want: To identify the target parameter through observed data, i.e.,

$$\psi(\tilde{\mathsf{P}}) = \theta(\mathsf{P}_X) \text{ for some functional, } \theta, \text{ of } \mathsf{P}_X$$

Ex. 1: One time point, $K = 1$, no censoring

Want: To identify the target parameter through observed data.

$$
\begin{aligned}
\tilde{\mathrm{P}}(Y^{(1)}(1) = 1) &= \int \tilde{\mathrm{P}}(Y^{(1)}(1) = 1 \mid A(0) = 1, L(0) = l) Q_{L(0)}(\mathrm{d}l) \\
&\overset{!}{=} \int \mathrm{P}(Y(1) = 1 \mid A(0) = 1, L(0) = l) Q_{L(0)}(\mathrm{d}l) \\
&= \int \mathrm{P}(Y(1) = 1 \mid A(0) = a, L(0) = l) G^*_{A(0)}(\mathrm{d}a) Q_{L(0)}(\mathrm{d}l) \\
&= \int Q_{Y(1)}(1 \mid a, l) G^*_{A(0)}(\mathrm{d}a) Q_{L(0)}(\mathrm{d}l) \\
&= Q_{L(0)} G^*_{A(0)} Q_{Y(1)} f, \qquad \text{(G-formula)}
\end{aligned}
$$

for $f(\mathbf{X}) = \mathbb{1}\{Y(1) = 1\}$. This leads to the estimator,

$$
\hat{Q}_{L(0)} \hat{G}^*_{A(0)} \hat{Q}_{Y(1)} f = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_{Y(1)}(1 \mid 1, L_i(0)).
$$

Ex. 2: Multiple time points, $K = 2$, with censoring

Consider data, $\mathbf{X} = (L(0), A(0), C(1), Y(1), L(1), A(1), C(2), Y(2))$, where C censoring variable taking values $\{0, 1\}$ with $C(t) = 0$ to mean that the event has been censored. Our target parameter is given by

$$\psi(\tilde{\mathsf{P}}) = \tilde{\mathsf{P}}(Y^{(1)}(2) = 1) - \tilde{\mathsf{P}}(Y^{(0)}(2) = 1).$$

Want: Identifiability, i.e., $\psi(\tilde{\mathsf{P}}) = \theta(\mathsf{P}_X)$. Need some more notation!

Ex. 2: Multiple time points, $K = 2$, with censoring

Consider data, $\mathbf{X} = (L(0), A(0), C(1), Y(1), L(1), A(1), C(2), Y(2))$, where C censoring variable taking values $\{0, 1\}$ with $C(t) = 0$ to mean that the event has been censored. Our target parameter is given by

$$\psi(\tilde{\mathsf{P}}) = \tilde{\mathsf{P}}(Y^{(1)}(2) = 1) - \tilde{\mathsf{P}}(Y^{(0)}(2) = 1).$$

Want: Identifiability, i.e., $\psi(\tilde{\mathsf{P}}) = \theta(\mathsf{P}_X)$. Need some more notation!

$$\begin{aligned} \text{History: } \bar{A}(1) =& (A(0), A(1)) \\ \text{Factorization: } dP_{\mathbf{X}} =& Q_{Y(2)} G_{C(2)} G_{A(1)} Q_{L(1)} Q_{Y(1)} G_{C(1)} G_{A(0)} Q_{L(0)} \end{aligned}$$

Then for some function of the data, $f(\mathbf{X})$, we can write

$$P_{\mathbf{X}} f = \int f(x) dP_{\mathbf{X}}(x) = Q_{L(0)} G_{A(0)} G_{C(1)} Q_{Y(1)} Q_{L(1)} G_{A(1)} G_{C(2)} Q_{Y(2)} f.$$

Ex. 2: Multiple time points, $K = 2$, with censoring

Consider the counterfactual distribution defined by

$$\begin{aligned} G^*_{\bar{A}(1)}(\mathrm{d}(a_0, a_1)) =& \mathbb{1}\{a_0 = 1, a_1 = 1\} \\ G^*_{\bar{C}(2)}(\mathrm{d}(c_1, c_2)) =& \mathbb{1}\{c_1 = 1, c_2 = 1\}. \end{aligned}$$

The G-formula in this case is obtained by

$$\begin{aligned} \tilde{\mathsf{P}}(Y^{(1)}(2) = 1) \stackrel{!}{=} \int & \mathsf{P}(Y(2) = 1 \mid \bar{C}(2) = c, \bar{A}(1) = a, \bar{L}(1) = l, Y(1) = y) \\ & \times G^*_{\bar{C}(2)} G^*_{\bar{A}(1)} Q_{L(1)} Q_{Y(1)} G^*_{C(1)} G^*_{A(0)} Q_{L(0)} \\ =& Q_{L(0)} G^*_{A(0)} G^*_{C(1)} Q_{Y(1)} Q_{L(1)} G^*_{\bar{A}(1)} G^*_{\bar{C}(2)} Q_{Y(2)} f, \end{aligned} \tag{G-formula}$$

for $f(\mathbf{X}) = \mathbb{1}\{Y(2) = 1\}$.

Sequential outcome regression

$$\text{G-formula:} \quad Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, \boldsymbol{G^*_{\bar{C}(2)}}\, \boldsymbol{Q_{Y(2)}}\, \boldsymbol{f}$$

Step 1. Regress $Y(2)$ on past covariates, $\bar{A}(1)$ and $\bar{L}(1)$, for subjects at risk

$$\rightsquigarrow \quad G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f = \mathrm{P}(Y(2) = 1 \mid \bar{C}(2) = 1, Y(1) = y, \bar{A}(1), \bar{L}(1))$$

Deterministic info: $Y(1) = 1 \Rightarrow Y(2) = 1$ and $C(1) = 0$ or $C(2) = 0 \Rightarrow Y(2) = 0$

```
head(d)
```

	L_0	A_0	C_1	Y_1	L_1	A_1	C_2	Y_2
1:	1	1	1	0	1	0	1	0
2:	0	0	1	1	1	0	1	1
3:	1	0	1	0	1	1	0	0

```
fit2 <- glm(Y_2 ~ L_0 + A_0 + L_1 + A_1,
            data = d[Y_1 == 0 & C_1 == 1 & C_2 == 1],
            family = binomial(link = "logit"))
```

Sequential outcome regression

G-formula: $Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, \boldsymbol{G^*_{\bar{C}(2)}}\, \boldsymbol{Q_{Y(2)}}\, \boldsymbol{f}$

Step 1. Regress $Y(2)$ on past covariates, $\bar{A}(1)$ and $\bar{L}(1)$, for subjects at risk

$$\rightsquigarrow \; G^*_{\bar{C}(2)} Q_{Y(2)} f = \mathrm{P}(Y(2) = 1 \mid \bar{C}(2) = 1, Y(1) = y, \bar{A}(1), \bar{L}(1))$$

$$Y(1) + (1 - Y(1))\, \mathrm{P}(Y(2) = 1 \mid \bar{C}(2) = 1, Y(1) = 0, \bar{A}(1), \bar{L}(1))$$

Deterministic info: $Y(1) = 1 \Rightarrow Y(2) = 1$ and $C(1) = 0$ or $C(2) = 0 \Rightarrow Y(2) = 0$

```
head(d)
```

	L_0	A_0	C_1	Y_1	L_1	A_1	C_2	Y_2
1:	1	1	1	0	1	0	1	0
2:	0	0	1	1	1	0	1	1
3:	1	0	1	0	1	1	0	0

```
fit2 <- glm(Y_2 ~ L_0 + A_0 + L_1 + A_1,
            data = d[Y_1 == 0 & C_1 == 1 & C_2 == 1],
            family = binomial(link = "logit"))
```

Sequential outcome regression

$$\text{G-formula:}\quad Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, \boldsymbol{G^*_{\bar{A}(1)}}\, \boldsymbol{G^*_{\bar{C}(2)}}\, \boldsymbol{Q_{Y(2)}}\, \boldsymbol{f}$$

Step 2. Predict according to counterfactual distribution

$$\rightsquigarrow\quad G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)} f = \mathrm{P}(Y(2) = 1 \mid \bar{C}(2) = 1, Y(1) = y, \bar{A}(1) = 1, \bar{L}(1))$$

```
d1 <- copy(d); d1[,A_0 := 1]; d1[,A_1 := 1]
head(d1)
```

	L_0	A_0	C_1	Y_1	L_1	A_1	C_2	Y_2
1:	1	1	1	0	1	1	1	0
2:	0	1	1	1	1	1	1	1
3:	1	1	1	0	1	1	0	0

```
GQhat2_0 <- predict(fit1, newdata = d1, type = "response")
d[,GQhat2 := (Y_1 + (1 - Y_1)*GQhat2_0)] # 1 if Y_1 = 1
```

Sequential outcome regression

$$\text{G-formula:} \quad Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, \boldsymbol{G^*_{\bar{A}(1)}}\, \boldsymbol{G^*_{\bar{C}(2)}}\, \boldsymbol{Q_{Y(2)}}\, \boldsymbol{f}$$

Step 2. Predict according to counterfactual distribution

$$\rightsquigarrow \quad G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f = \mathrm{P}(Y(2) = 1 \mid \bar{C}(2) = 1, Y(1), \bar{A}(1) = 1, \bar{L}(1)) = \boldsymbol{\bar{Q}^{d,2}_{L(2)}}$$

```
d1 <- copy(d); d1[,A_0 := 1]; d1[,A_1 := 1]
head(d1)
```

	L_0	A_0	C_1	Y_1	L_1	A_1	C_2	Y_2
1:	1	1	1	0	1	1	1	0
2:	0	1	1	1	1	1	1	1
3:	1	1	1	0	1	1	0	0

```
GQhat2_0 <- predict(fit1, newdata = d1, type = "response")
d[,GQhat2 := (Y_1 + (1 - Y_1)*GQhat2_0)] # 1 if Y_1 = 1
```

Sequential outcome regression

$$\text{G-formula:} \quad Q_{L(0)}\, G^*_{A(0)}\, \boldsymbol{G^*_{C(1)}}\, \boldsymbol{Q_{Y(1)}}\, \boldsymbol{Q_{L(1)}}\, \boldsymbol{G^*_{\bar{A}(1)}}\, \boldsymbol{G^*_{\bar{C}(2)}}\, \boldsymbol{Q_{Y(2)}}\, \boldsymbol{f}$$

Step 3. Regress $G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)} f$ on $A(0)$ and $L(0)$ for subjects at risk

$$\begin{aligned} &\rightsquigarrow \; G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)} f \\ &= \mathbb{E}\left[G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)} f \mid C(1) = 1, A(0), L(0) \right] \end{aligned}$$

Note: $G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)} f \in (0,1)$ not binary. We use quasi-binomial logistic regression with an extra dispersion parameter to describe additional variation in data – in R with glm, coefficient estimates are the same, but the SE's differ.

```
fit1 <- glm(GQhat2 ~ L_0 + A_0,
            data = d[C_1 == 1],
            family = quasibinomial(link = "logit"))
```

Sequential outcome regression

$$\text{G-formula:} \quad Q_{L(0)}\, \boldsymbol{G^*_{A(0)}}\, \boldsymbol{G^*_{C(1)}}\, \boldsymbol{Q_{Y(1)}}\, \boldsymbol{Q_{L(1)}}\, \boldsymbol{G^*_{\bar{A}(1)}}\, \boldsymbol{G^*_{\bar{C}(2)}}\, \boldsymbol{Q_{Y(2)}}\, \boldsymbol{f}$$

Step 4. Predict according to counterfactual distribution

$$\begin{aligned} \rightsquigarrow \quad & G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f \\ =& \mathbb{E}\left[G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f \mid C(1) = 1, A(0) = 1, L(0) \right] \end{aligned}$$

```
GQhat1 <- predict(fit1, newdata = d1, type = "response"))
```

Step 5. Take sample average over $L(0)$

$$\rightsquigarrow \quad Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f$$

```
mean(GQhat1)
[1] 0.03377855
```

Sequential outcome regression

$$\text{G-formula:}\quad Q_{L(0)}\,\boldsymbol{G^*_{A(0)}}\,\boldsymbol{G^*_{C(1)}}\,\boldsymbol{Q_{Y(1)}}\,\boldsymbol{Q_{L(1)}}\,\boldsymbol{G^*_{\bar{A}(1)}}\,\boldsymbol{G^*_{\bar{C}(2)}}\,\boldsymbol{Q_{Y(2)}}\,\boldsymbol{f}$$

Step 4. Predict according to counterfactual distribution

$$\begin{aligned}\rightsquigarrow\ & G^*_{A(0)}\,G^*_{C(1)}\,Q_{Y(1)}\,Q_{L(1)}\,G^*_{\bar{A}(1)}\,G^*_{\bar{C}(2)}\,Q_{Y(2)}\,f\\ =&\mathbb{E}\Big[\underbrace{G^*_{\bar{A}(1)}\,G^*_{\bar{C}(2)}\,Q_{Y(2)}\,f}_{\bar{\boldsymbol{Q}}^{d,2}_{L(2)}} \mid C(1)=1, A(0)=1, L(0)\Big] = \bar{\boldsymbol{Q}}^{d,2}_{L(1)}\end{aligned}$$

```
GQhat1 <- predict(fit1, newdata = d1, type = "response"))
```

Step 5. Take sample average over $L(0)$

$$\rightsquigarrow\quad Q_{L(0)}\,G^*_{A(0)}\,G^*_{C(1)}\,Q_{Y(1)}\,Q_{L(1)}\,G^*_{\bar{A}(1)}\,G^*_{\bar{C}(2)}\,Q_{Y(2)}\,f$$

```
mean(GQhat1)
[1] 0.03377855
```

Sequential outcome regression

G-formula: $\boldsymbol{Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f}$

Step 4. Predict according to counterfactual distribution

$$\begin{aligned}\rightsquigarrow\ & G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f \\ =& \mathbb{E}\Big[\underbrace{G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f}_{\bar{Q}^{d,2}_{L(2)}} \mid C(1)=1, A(0)=1, L(0)\Big] = \boldsymbol{\bar{Q}^{d,2}_{L(1)}}\end{aligned}$$

```
GQhat1 <- predict(fit1, newdata = d1, type = "response"))
```

Step 5. Take sample average over $L(0)$

$$\rightsquigarrow\quad Q_{L(0)}\, G^*_{A(0)}\, G^*_{C(1)}\, Q_{Y(1)}\, Q_{L(1)}\, G^*_{\bar{A}(1)}\, G^*_{\bar{C}(2)}\, Q_{Y(2)}\, f = \boldsymbol{\bar{Q}^{d,2}_{L(0)}}$$

```
mean(GQhat1)
[1] 0.03377855
```

Sequential outcome regression with the Ltmle package

Maximum likelihood based G-computation estimate with Ltmle

Step 0. Prepare data.

```
head(data)
```

	L_0	A_0	C_1	Y_1	L_1	A_1	C_2	Y_2
1:	1	1	uncensored	0	1	0	uncensored	0
2:	0	0	uncensored	1	NA	NA	NA	1
3:	1	0	uncensored	0	1	1	censored	NA

```
gform
[1] "A_0 ~ L_0"          "C_1 ~ L_0 + A_0 "
[2] "A_1 ~ L_0 + A_0"    "C_2 ~ L_0 + L_1 + A_0 + A_1"

Qform
                  Y_1                                    Y_2
"Q.plus1 ~ L_0 + A_0"   "Q.plus1 ~ L_0 + L_1 + A_0 + A_1"
```

Sequential outcome regression with the Ltmle package

Maximum likelihood based G-computation estimate with Ltmle

Use setting: gcomp = TRUE (default is FALSE)

```
fit_ltmle <- Ltmle(data = data,
                   Anodes = c("A_0","A_1"),
                   Cnodes = c("C_1","C_2"),
                   Lnodes = c("L_0", "L_1"),
                   Ynodes = c("Y_0", "Y_1"),
                   survivalOutcome = TRUE,
                   Qform = Qform,
                   gform = gform,
                   abar = list(c(1,1),c(0,0)),
                   gcomp = TRUE,
                   SL.library = "glm")

summary(fit_ltmle)$effect.measures$treatment$estimate
[1] 0.03377855
```

Discussion

Extension to $K > 2$ time points is straightforward

Discussion

Extension to $K > 2$ time points is straightforward

Step 1 + 2. Regress $Y(K)$ on past cov. (until time $K-1$) for subjects at risk and predict according to intervention rule $\rightsquigarrow G^*_{\bar{A}(K-1)} G^*_{\bar{C}(K)} Q_{Y(K)} f$

Step 3 + 4. Regress $G^*_{\bar{A}(K-1)} G^*_{\bar{C}(K)} Q_{Y(K)} f$ on past cov. (until time $K-2$) for subjects at risk and predict according to intervention rule

$\vdots$

Step $2K+1$. Take sample average over $L(0)$

Discussion

Extension to $K > 2$ time points is straightforward

Extension to competing risk:

* Suppose $D(1)$ competing risk, e.g., a value of one means death
* At risk: d[Y_1 == 0 & C_1 == 1 & C_2 == 1 & D_1 == 0]
* Deterministic info about $Y(2)$, e.g., $D(1) = 1 \Rightarrow Y(2) = 0$
* In Ltmle: deterministic.Q.function

Discussion

Extension to $K > 2$ time points is straightforward

Extension to competing risk:

* Suppose $D(1)$ competing risk, e.g., a value of one means death
* At risk: d[Y_1 == 0 & C_1 == 1 & C_2 == 1 & D_1 == 0]
* Deterministic info about $Y(2)$, e.g., $D(1) = 1 \Rightarrow Y(2) = 0$
* In Ltmle: deterministic.Q.function

References

[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973, 2005. (Hard to read!)

[2] Samuel D Lendle, Joshua Schwab, Maya L Petersen, and Mark J van der Laan. ltmle: an r package implementing targeted minimum loss-based estimation for longitudinal data. Journal of Statistical Software, 81:1–21, 2017. (Ltmle doc.)