FACULTY OF SCIENCES

European Causal Inference Meeting, 2026

Debiased Machine Learning for Risk Minimization

with an application to Generalized Partially Linear Models

Federico Bertoia, Achille Demares, Georgi Baklicharov and Stijn Vansteelandt, Ghent University

GHENT UNIVERSITY

Funded by the European Union

European Research Council
Established by the European Commission

SETTING

We observe N iid data on $O_i \equiv \big(Y_i, A_i, L_i\big)$ such that:

- Y_i is the outcome of interest
- A_i is a (continuous) treatment
- L_i a vector of baseline covariates to adjust for confounding

Focus on the conditional average treatment effect of shifting the exposure from A to $A + a$

$$E\big[Y^{A+a} - Y \mid A, L\big]$$

under standard causal assumptions. More generally we consider contrasts

$$g\big\{E\big[Y^{A+a} \mid A, L\big]\big\} - g\big\{E\big[Y \mid A, L\big]\big\} = \int_{A}^{A+a} h(s, L)ds$$

Parametrization of the CATE

The function $h(a, L)$ belongs to an unknown function class $\mathcal{H}$. Estimating it fully nonparametrically poses different challenges:

1. $\sqrt{n}$-inference
2. Uncertainty quantification
3. Can be difficult to communicate

To remedy these problems, we approximate $h(a, L)$ via parametric working models of the form:

$$h(a, L) = \psi(a, L)^\top \theta$$

Such that the (generalized) CATE takes the form:

$$g\{E[Y^{A+a} \mid A, L]\} - g\{E[Y \mid A, L]\} = \int_A^{A+a} \psi(s, L)^\top \theta ds$$

Different approximations lead to different working models

$\psi(a, L)$	$g\{E[Y \mid A, L]\} =$	Model
1	$\alpha(L) + A\theta$	Linear GPLM
$(1,\ a)^\top$	$\alpha(L) + A\theta_1 + A^2\theta_2$	Quadratic GPLM
$(1,\ a,\ a^2)^\top$	$\alpha(L) + A\theta_1 + A^2\theta_2 + A^3\theta_3$	Cubic GPLM
$(1,\ L_1)^\top$	$\alpha(L) + A\theta_1 + AL_1\theta_2$	Linear + effect modification
$(1,\ a,\ L_1,\ aL_1)^\top$	$\alpha(L) + A\theta_1 + A^2\theta_2 + AL_1\theta_3 + A^2L_1\theta_4$	Quadratic + effect modification

PREDICTING THE SHIFT EFFECT

For simplicity, consider the linear GPLM approximation such that:

$$g\left\{E\left[Y^{A+a} \mid A, L\right]\right\} - g\left\{E\left[Y \mid A, L\right]\right\} = \theta a$$

Under this approximation, we can predict the effect of shifting the exposure from A to $A + a$ in two ways:

1. Predict on the link-scale g:

$$g\left\{E\left[Y^{A+a} \mid A, L\right]\right\} = g\left\{E\left[Y \mid A, L\right]\right\} + \theta a$$

2. Predict on the response-scale:

$$E\left[Y^{A+a} \mid A, L\right] = g^{-1}\left[g\left\{E\left[Y \mid A, L\right]\right\} + \theta a\right]$$

From Estimands to Risk Functions

We define our parameter of interest θ as the minimizer of a population risk function that quantifies (squared) biased in the approximation of the true causal effect.

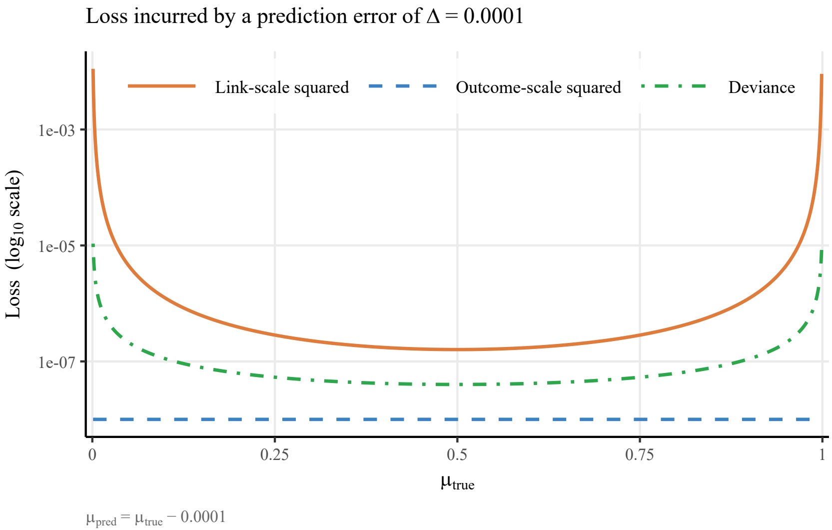
1. Link-scale risk function (Vansteelandt, 2025)

$$\mathcal{L}^{\text{link}}(\theta) = E\left[\int \left(g\left\{E\left[Y^{A+a} \mid A, L\right]\right\} - g\left\{E\left[Y \mid A, L\right]\right\} - \theta a\right)^2 f(A+a \mid L)da\right]$$

2. Response-scale risk function

$$\mathcal{L}^{\text{response}}(\theta) = E\left[\int \left(E\left[Y^{A+a} \mid A, L\right] - g^{-1}\left[g\left\{E\left[Y \mid A, L\right]\right\} + \theta a\right]\right)^2 f(A+a \mid L)da\right]$$

Comparison of Risk Functions



A Plug-In Estimator for Continuous Exposures

Recall the response-scale risk function with a linear GPLM approximation:

$$\mathcal{L}^{\text{response}}(\theta) = E\left[\int \left(E\left[Y^{A+a} \mid A, L\right] - g^{-1}\left[g\left\{E\left[Y \mid A, L\right]\right\} + \theta a\right]\right)^2 f(A+a \mid L)da\right]$$

Let $A + a = A^*$:

$$\mathcal{L}^{\text{response}}(\theta) = E\left[\int \left(E\left[Y \mid A^*, L\right] - g^{-1}\left[g\left\{E\left[Y \mid A, L\right]\right\} + \theta\left(A^* - A\right)\right]\right)^2 f(A^* \mid L)dA^*\right]$$

Compute the integral via Monte Carlo integration:

$$\mathcal{L}^{\text{response}}(\theta) = E\left[\frac{1}{J}\sum_{j=1}^{J}\left(\widehat{E}(Y \mid A_i^{*(j)}, L_i) - g^{-1}\left[g\left\{\widehat{E}(Y \mid A_i, L_i)\right\} + \theta\left(A_i^{*(j)} - A_i\right)\right]\right)^2\right],$$

where $A_i^{*(1)}, \ldots, A_i^{*(J)}$ are J independent draws from $\widehat{f}(\cdot \mid L_i)$, so that $a = A_i^{*(j)} - A_i$ plays the role of the shift.

A Plug-In Estimator for Continuous Exposures

Finally, replace outer expectation with the sample average. A plug-in estimate can be computed as

$$\widehat{\mathcal{L}}^{\text{plug-in}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J}\sum_{j=1}^{J}\left(\widehat{E}(Y \mid A_i^{*(j)}, L_i) - g^{-1}\left[g\left\{\widehat{E}(Y \mid A_i, L_i)\right\} + \theta\left(A_i^{*(j)} - A_i\right)\right]\right)^2,$$

It requires estimating two nuisance parameters:

1. $E\left(Y \mid A, L\right)$, which we model using Super Learner (van der Laan et al., 2007).
2. $f\left(A \mid L\right)$, which we model using Engression (Shen and Meinshausen, 2025).

Debiased Machine Learning for a General Risk Function

1. Plug-in Estimator is biased

$$\widehat{\theta}^{\text{plug-in}} = \arg\min_{\theta \in \mathbb{R}} \widehat{\mathcal{L}}(\theta)$$

2. Construct a Debiased Estimator for the Risk

$$\widehat{\mathcal{L}}^{\text{debiased}}(\theta) = \widehat{\mathcal{L}}(\theta) + \mathbb{P}_n \Phi\big(\widehat{\mathcal{L}}(\theta)\big)$$

3. Estimate θ by minimizing the debiased risk

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}} \widehat{\mathcal{L}}^{\text{debiased}}(\theta)$$

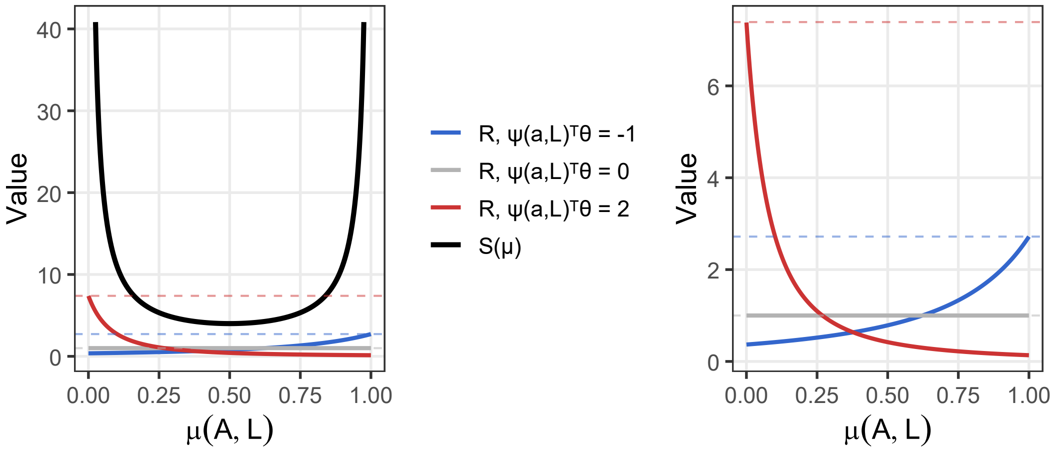
Asymptotic distribution

Under cross-fitting and standard rate conditions for the nuisances,

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\ \mathrm{Var}\big(\Phi(\theta_0)\big)\big)$$

(In)Stability of the Efficient Influence Function

Stability factors under the logit link



Simulations — Overview

Two DGPs under the logit link with a binary outcome and a continuous exposure following the (linear) generalized partially linear model:

1. Linear DGP

$$g\{E[Y \mid A, L]\} = \alpha^\top L + \theta A$$

2. Nonlinear DGP

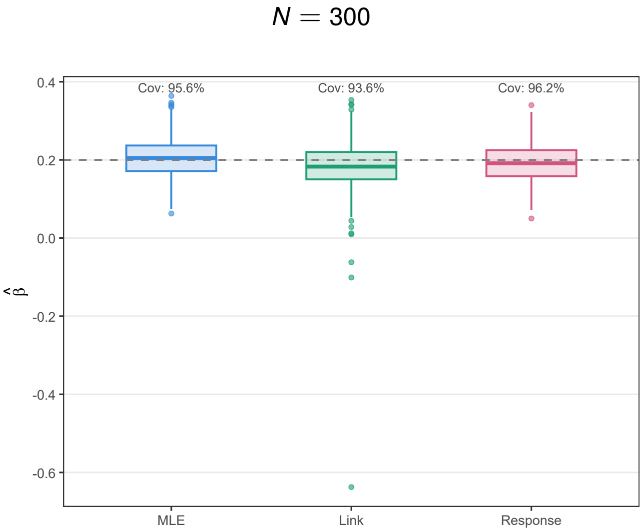
$$g\{E[Y \mid A, L]\} = \alpha(L) + \theta A$$

DGP	Setting	Q01	Q05	Q25	Q50	Q75	Q90	Q95	Q99
Linear	Non-extreme	0.117	0.151	0.227	0.313	0.435	0.580	0.675	0.842
Nonlinear	Extreme	0.413	0.594	0.791	0.875	0.937	0.974	0.986	0.997

Three estimators, $N = \{300, 1000, 2500\}$, 500 simulations and 5 folds cross fitting:

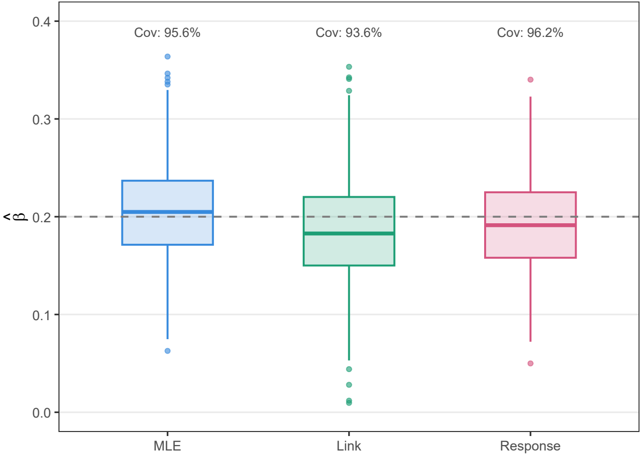
- **MLE**: logistic regression.
- **Link** / **Response**: DML estimators based on link/response risk function.

Simulations — Linear DGP



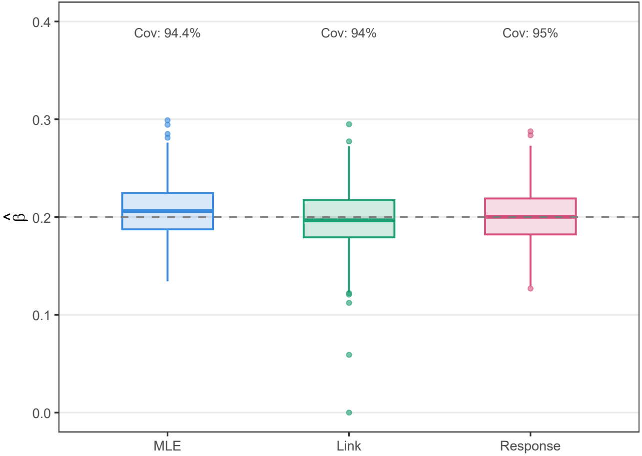
Simulations — Linear DGP

$N = 300$ (Zoomed In)



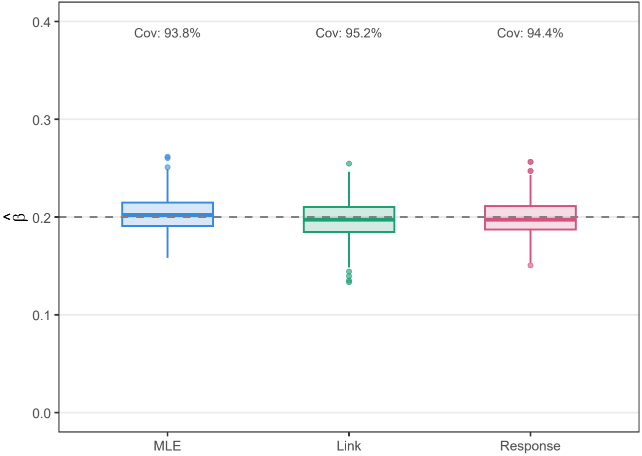
SIMULATIONS — LINEAR DGP

$N = 1000$



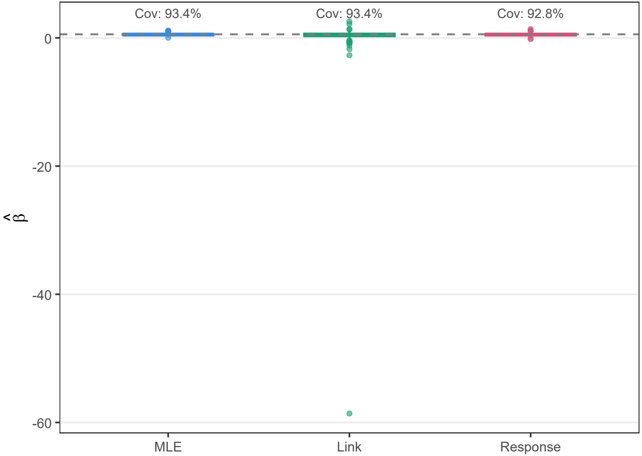
Simulations — Linear DGP

$$N = 2500$$



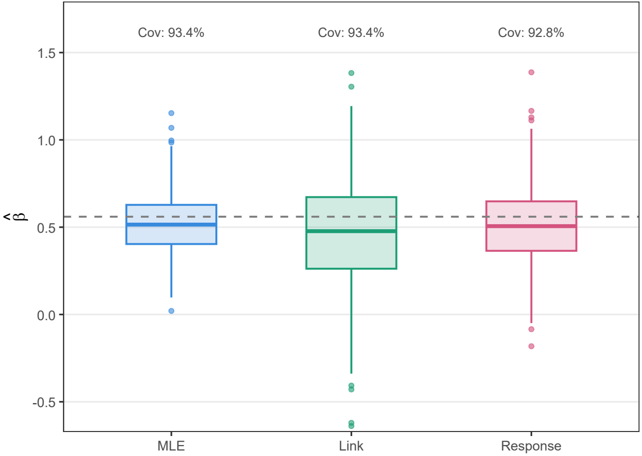
Simulations — Nonlinear DGP

$N = 300$



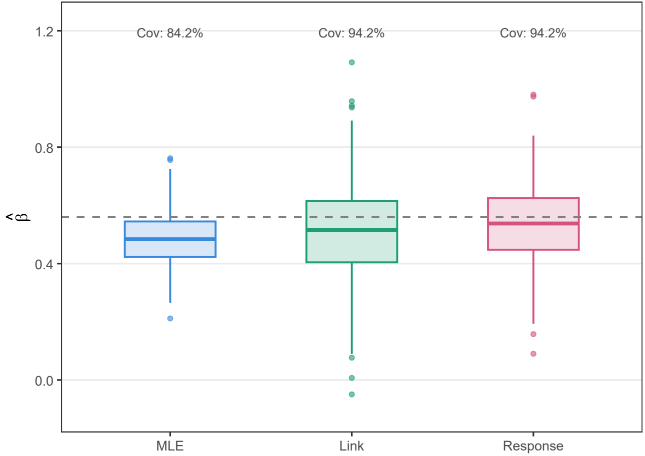
SIMULATIONS — NONLINEAR DGP

$N = 300$ (Zoomed In)



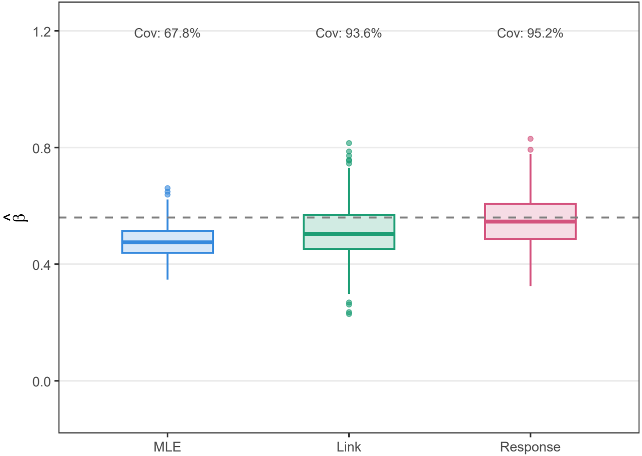
Simulations — Nonlinear DGP

$N = 1000$



Simulations — Nonlinear DGP

$N = 2500$



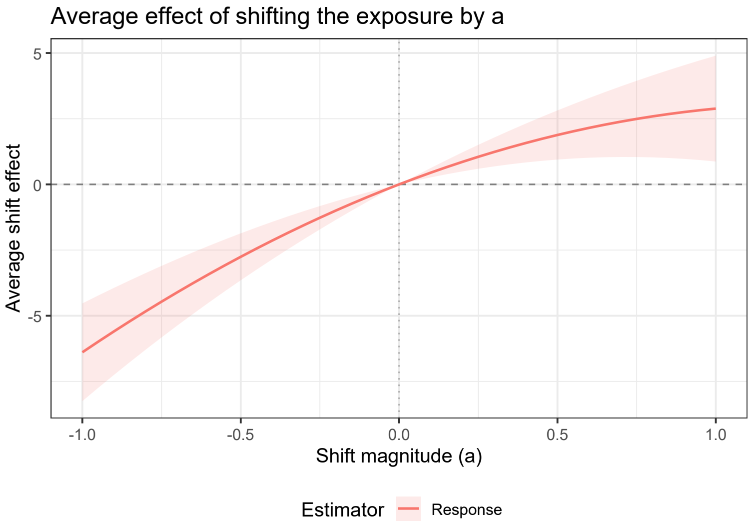
Data Analysis — Overview

We study the effect of $PM_{2.5}$ particle level on cardiovascular mortality rate (CMR) across 2,132 US counties (2010).

- **Exposure:** $PM_{2.5}$ particle level ($\mu g/m^3$)
- **Outcome:** annual cardiovascular deaths per 100,000 people
- **Confounders:** 10 county-level variables (poverty rate, population, household income, ...)

Data from the National Studies on Air Pollution and Health (Wyatt et al., 2020).

Data Analysis - Average Shift Effect



Summary and Conclusions

Outside simple settings, causal estimands are often hard to define (e.g., continuous exposures) or communicate (e.g., CATEs).

- We instead use models as approximations, shifting the question from which estimand to which risk function.
- Our risk functions are distinct in that they purely quantify causal effect approximation error - not covariate effects - in a practice-relevant way.
- Debiased machine learning then directly minimizes these approximation errors.

Thank you for your attention!

Feel free to reach out at

`federico.bertoia@ugent.be`

References I

Shen, X. and Meinshausen, N. (2025). Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677.

van der Laan, M. J., Polley, E., and Hubbard, A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25).

Vansteelandt, S. (2025). Towards efficient and interpretable assumption-lean generalized linear modeling of continuous exposure effects. *Biometrics*, 81(2):ujaf071.

Wyatt, L. H., Peterson, G. C. L., Wade, T. J., Neas, L. M., and Rappold, A. G. (2020). Annual pm2. 5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 us counties. *Data in brief*, 30:105318.

Appendix

$h(a, L)$ as Average Derivative Effect

We consider contrasts

$$g\{E[Y^{A+a} \mid A, L]\} - g\{E[Y \mid A, L]\} = \int_A^{A+a} h(s, L)\, ds$$

where $g(\cdot)$ is a user-specified link function (e.g., identity, logit or log). By the fundamental theorem of calculus, this implies

$$h(a, L) = \frac{d}{ds} g\{E[Y^s \mid A, L]\}\Big|_{s=a}$$

is the **average derivative effect** on the $g(\cdot)$ scale.

Efficient Influence Function of Response Scale

The efficient influence function of $\mathcal{L}(\theta, \mathcal{P})$, under the nonparametric model is

$$\begin{aligned}
\Phi(\mathcal{L}(\theta,\mathcal{P})) = & \int_a \left(\mu(A+a,L) - g^{-1}\left[g\{\mu(A,L)\} + \psi(a,A,L)^\top \theta\right]\right)^2 f(A+a \mid L)\, da - \mathcal{L}(\theta,\mathcal{P}) \\
& + 2\{Y - \mu(A,L)\} \int_a \left(\mu(A,L) - g^{-1}\left[g\{\mu(A-a,L)\} + \psi(a,A,L)^\top \theta\right]\right) f(A-a \mid L)\, da \\
& - 2g'\{\mu(A,L)\}\{Y - \mu(A,L)\} \int_a \left(\mu(A+a,L) - g^{-1}\left[g\{\mu(A,L)\} + \psi(a,A,L)^\top \theta\right]\right) \\
& \times g^{-1\prime}\left[g\{\mu(A,L)\} + \psi(a,A,L)^\top \theta\right] f(A+a \mid L)\, da \\
& + \int_a \left(\mu(A,L) - g^{-1}\left[g\{\mu(A-a,L)\} + \psi(a,A,L)^\top \theta\right]\right)^2 f(A-a \mid L)\, da \\
& - \int_A \int_a \left(\mu(A+a,L) - g^{-1}\left[g\{\mu(A,L)\} + \psi(a,A,L)^\top \theta\right]\right)^2 f(A+a \mid L) f(A \mid L)\, da\, dA.
\end{aligned}$$

where $\mu(A,L) = E(Y \mid A,L)$.

Stability factor: response-scale vs. link-scale

Response-scale factor R

$$R(\mu, a, A, L, \theta) = \frac{g'\{\mu(A, L)\}}{g'\{\eta(\mu, a)\}}$$

Uniformly bounded:

$$e^{-|\psi^\top \theta|} \le R \le e^{|\psi^\top \theta|}$$

Under the log link: $R = e^{\psi^\top \theta}$ exactly, independent of $\mu(A, L)$.

$R = 1$ when $\theta = 0$ or under the identity link.

Link-scale factor S

$$S(\mu) = g'\{\mu(A, L)\}$$

Unbounded at boundary:

$$\sup_{\mu \in \mathcal{M}} S(\mu) = \infty$$

- Logit link: $g'(\mu) = \frac{1}{\mu(1-\mu)} \to \infty$
- Log link: $g'(\mu) = \frac{1}{\mu} \to \infty$ as $\mu \to 0^+$

where $\eta(\mu, a) = g\{\mu(A, L)\} + \psi(a, A, L)^\top \theta$.

Average Shift Effect

Average Shift Effect estimand:

$$\mathrm{ATE}(a;\,\theta) = E\left[g^{-1}\left(g(\mu) + \psi(a, L)^{\top}\theta\right) - \mu\right]$$

Its efficient influence function under the nonparametric model is

$$\Phi\left(\mathrm{ATE}(a;\,\theta_0)\right) = g^{-1}\left(g(\mu) + \psi^{\top}\theta_0\right) - \mu - \mathrm{ATE}(a;\,\theta_0) + \left(\frac{g'(\mu)}{g'(\eta)} - 1\right)(Y - \mu) + \nabla_{\theta}\mathrm{ATE}^{\top}\,\Phi(\theta_0),$$

where $\eta = g^{-1}(g(\mu) + \psi^{\top}\theta_0)$ and $\Phi(\theta_0)$ is the influence function of $\widehat{\theta}$.

Data Analysis - Average Shift Effect (MLE)