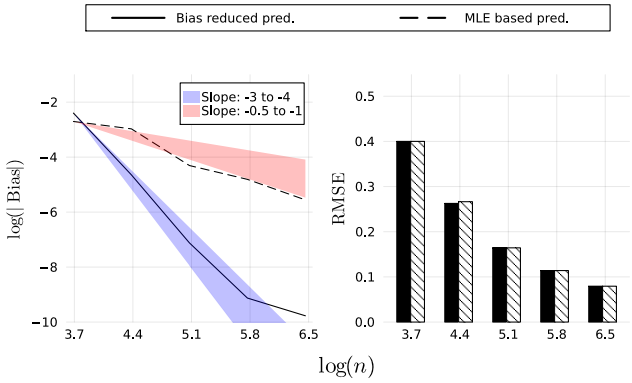
WARWICK
THE UNIVERSITY OF WARWICK

Bias reduced prediction
for black-box models

Philipp Sterzinger
philipp.sterzinger@warwick.ac.uk

Bias reduced prediction for black-box models

We develop a novel prediction approach that yields first order **unbiased predictions** for a broad class of statistical models whose training can be framed as a **M-estimation** problem.



Setup
Fixed design

- Response vectors: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ic_i})^\top \in \mathcal{Y} \subseteq \Re^{c_i}$, realizations of random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$
- Covariates: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, with $\boldsymbol{x}_i \in \mathcal{X} \subseteq \Re^{q_i}$
- Conditional distribution: $G(\boldsymbol{Y}|\boldsymbol{X})$ (unknown) distribution of $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_k^\top)^\top$ given the set of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, denoted by $\boldsymbol{X}$
- Rate of information accumulation: $n = n(k, \{c_i\}, \{q_i\})$

For simplicity, $y_i \in \Re, \boldsymbol{x}_i \in \Re^p, n = k$ but more general setting possible

Setup
Fixed design

- Response vectors: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ic_i})^\top \in \mathcal{Y} \subseteq \Re^{c_i}$, realizations of random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$
- Covariates: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, with $\boldsymbol{x}_i \in \mathcal{X} \subseteq \Re^{q_i}$
- Conditional distribution: $G(\boldsymbol{Y}|\boldsymbol{X})$ (unknown) distribution of $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_k^\top)^\top$ given the set of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, denoted by $\boldsymbol{X}$
- Rate of information accumulation: $n = n(k, \{c_i\}, \{q_i\})$

For simplicity, $y_i \in \Re, \boldsymbol{x}_i \in \Re^p, n = k$ but more general setting possible

Given a new data point $\boldsymbol{x} \in \mathcal{X}$, we want to make a prediction about some aspect of the corresponding unobserved response $\boldsymbol{y} \in \mathcal{Y}$.

Setup

Prediction function

Known function $g(\boldsymbol{x}; \boldsymbol{\theta})$, for prediction point $\boldsymbol{x}$ parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \Re^p$ (e.g. weights and biases in NNs), i.e.

$$\begin{aligned} g: \quad & \mathcal{X} \times \boldsymbol{\Theta} \to \Re \\ & \boldsymbol{x} \mapsto g(\boldsymbol{x}; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}. \end{aligned}$$

M-estimation

Train the prediction function by solving the *M*-estimation problem

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \mathbf{0}_p \,,$$

for $\boldsymbol{\psi}^i(\boldsymbol{\theta}) = (\psi_1^i(\boldsymbol{\theta}), \ldots, \psi_p^i(\boldsymbol{\theta}))^\top$, $\psi_r^i(\boldsymbol{\theta}) = \psi_r^i(\boldsymbol{\theta}, \boldsymbol{y}_i, \boldsymbol{x}_i) : \boldsymbol{\Theta} \times \mathcal{Y} \times \mathcal{X} \to \Re$, $r \in \{1, \ldots, p\}$. e.g. FOC for models where training equates to optimizing some loss function

M-estimation perspective
Target prediction

As we observe more data ($n \to \infty$), training should stabilize in the sense that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ for some $\boldsymbol{\theta}_0 \in \Re^p$.

The prediction of interest is thus

$$\pi_0(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\theta}_0), \quad \text{for } \boldsymbol{\theta}_0 = \operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\theta}}, \quad \text{and } \hat{\boldsymbol{\theta}} : \sum_{i=1}^{n} \psi^i(\boldsymbol{\theta}) = \mathbf{0}_p \,,$$

which we approximate by the plug-in prediction $g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$.

M-estimation perspective
Target prediction

As we observe more data ($n \to \infty$), training should stabilize in the sense that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ for some $\boldsymbol{\theta}_0 \in \Re^p$.

The prediction of interest is thus

$$\pi_0(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\theta}_0), \quad \text{for } \boldsymbol{\theta}_0 = \operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\theta}}, \quad \text{and } \hat{\boldsymbol{\theta}} : \sum_{i=1}^{n} \psi^i(\boldsymbol{\theta}) = \mathbf{0}_p \,,$$

which we approximate by the plug-in prediction $g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$.

Can we tweak the *M*-estimation model training to get more accurate (low bias) predictions?

M-estimation perspective

RB*M*-estimation

Kosmidis and Lunardon (2020) show that under some regularity conditions,

$$E_G(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathcal{O}(n^{-1})\,.$$

They develop a class of adjustment functions

$$A: \quad \boldsymbol{\Theta} \times \mathcal{Y}^{\otimes n} \times \mathcal{X}^{\otimes n} \to \Re^p\,,$$

which solely depend on the estimating functions ψ^i and its derivatives, such that the adjusted system of estimating equations

$$\sum_{i=1}^{n} \psi^i(\boldsymbol{\theta}) + A(\boldsymbol{\theta}) = \mathbf{0}_p\,,$$

yields a first-order unbiased $\bar{\boldsymbol{\theta}}$, i.e.

$$E_G(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathcal{O}(n^{-3/2})\,.$$

M-estimation perspective

RB*M*-estimation

Kosmidis and Lunardon (2020) show that under some regularity conditions,

$$E_G(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathcal{O}(n^{-1})\,.$$

They develop a class of adjustment functions

$$A: \quad \boldsymbol{\Theta} \times \mathcal{Y}^{\otimes n} \times \mathcal{X}^{\otimes n} \to \Re^p\,,$$

which solely depend on the estimating functions ψ^i and its derivatives, such that the adjusted system of estimating equations

$$\sum_{i=1}^{n} \psi^i(\boldsymbol{\theta}) + A(\boldsymbol{\theta}) = \mathbf{0}_p\,,$$

yields a first-order unbiased $\bar{\boldsymbol{\theta}}$, i.e.

$$E_G(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathcal{O}(n^{-3/2})\,.$$

Naive plug-in $g(\boldsymbol{x}; \bar{\boldsymbol{\theta}})$ ↯ (Order of bias not preserved under nonlinear transformations)

Bias reduced prediction
Redundant est. equations

Introduce a redundant estimating equation to reframe the prediction task as a M-estimation problem (see Stefanski and Boos (2002))

Bias reduced prediction

Redundant est. equations

Introduce a redundant estimating equation to reframe the prediction task as a M-estimation problem (see Stefanski and Boos (2002))

2-step prediction

- *Training* $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \mathbf{0}_p$$

- *Prediction*: $\hat{\pi}(\boldsymbol{x}) = g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$

Bias reduced prediction
Redundant est. equations

Introduce a redundant estimating equation to reframe the prediction task as a M-estimation problem (see Stefanski and Boos (2002))

2-step prediction

- *Training* $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \mathbf{0}_p$$

- *Prediction*: $\hat{\pi}(\boldsymbol{x}) = g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$

Redundant estimating eq.

- *Training* $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^\top, \pi)^\top$:

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \mathbf{0}_p$$
$$\pi - g(\boldsymbol{x}; \boldsymbol{\theta}) = 0$$

- *Prediction*: $\hat{\pi} = g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$

Bias reduced prediction
Redundant est. equations

Introduce a redundant estimating equation to reframe the prediction task as a M-estimation problem (see Stefanski and Boos (2002))

2-step prediction

- *Training* $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \mathbf{0}_p$$

- *Prediction*: $\hat{\pi}(\boldsymbol{x}) = g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$

Redundant estimating eq.

- *Training* $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^\top, \pi)^\top$:

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) = \mathbf{0}_p$$

$$\pi - g(\boldsymbol{x}; \boldsymbol{\theta}) = 0$$

- *Prediction*: $\hat{\pi} = g(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$

Bias reduced prediction

- *Training* $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^\top, \pi)^\top$:

$$\sum_{i=1}^{n} \boldsymbol{\psi}^i(\boldsymbol{\theta}) + A_{1:p}(\boldsymbol{\vartheta}) = \mathbf{0}_p$$

$$\pi - g(\boldsymbol{x}; \boldsymbol{\theta}) + A_p(\boldsymbol{\vartheta}) = 0$$

- *Prediction*: $\bar{\pi}$ solves above

Bias reduced prediction
Asymptotics

Under M-estimation regularity conditions, and smooth $g(\boldsymbol{x};\boldsymbol{\theta})$ (cont., $4\times$ diff.) it holds that

$$\mathsf{E}_G(\bar{\vartheta} - \vartheta_0) = \mathsf{E}_G\left(\begin{bmatrix}\bar{\boldsymbol{\theta}}\\ \bar{\pi}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\theta}_0\\ \pi_0\end{bmatrix}\right) = \mathcal{O}(n^{-3/2}),$$

and $\boldsymbol{Q}(\vartheta_0)^{1/2}(\bar{\vartheta} - \vartheta_0) \xrightarrow{d} N(0, \boldsymbol{I}_{p+1})$ for some matrix $\boldsymbol{Q}$ with consistent plug-in estimator.

Logistic regression
Model

- Responses: $y_1, \ldots, y_n$, $y_i \in \{0, 1\}$, realizations of random variables $Y_1, \ldots, Y_n$
- Covariates: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i \in \mathcal{X} \subseteq \Re^p$
- Conditional distribution: $Y_i\big|_{\boldsymbol{x}_i} \overset{\text{ind.}}{\sim} \text{Ber}(\mu(\boldsymbol{x}_i^\top \boldsymbol{\theta}_0))$, $\log\left(\frac{\mu(z)}{1-\mu(z)}\right) = z$

Given new data point $\boldsymbol{x} \in \Re^p$, we want to predict the conditional mean $\mu(\boldsymbol{x}^\top \boldsymbol{\theta}_0)$.

Logistic regression
Model

- Responses: $y_1, \ldots, y_n$, $y_i \in \{0, 1\}$, realizations of random variables $Y_1, \ldots, Y_n$
- Covariates: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i \in \mathcal{X} \subseteq \Re^p$
- Conditional distribution: $Y_i\big|_{\boldsymbol{x}_i} \overset{\text{ind.}}{\sim} \text{Ber}(\mu(\boldsymbol{x}_i^\top \boldsymbol{\theta}_0))$, $\log\left(\frac{\mu(z)}{1-\mu(z)}\right) = z$

Given new data point $\boldsymbol{x} \in \Re^p$, we want to predict the conditional mean $\mu(\boldsymbol{x}^\top \boldsymbol{\theta}_0)$.

Regularity conditions satisfied for conditions that give consistency and asymptotic normality of the MLE (e.g Gourieroux and Monfort (1981), Amemiya (1985))

Logistic regression

Empirical adjustments

Agnostic about modelling assumption, solely depend on estimating function & derivatives thereof.

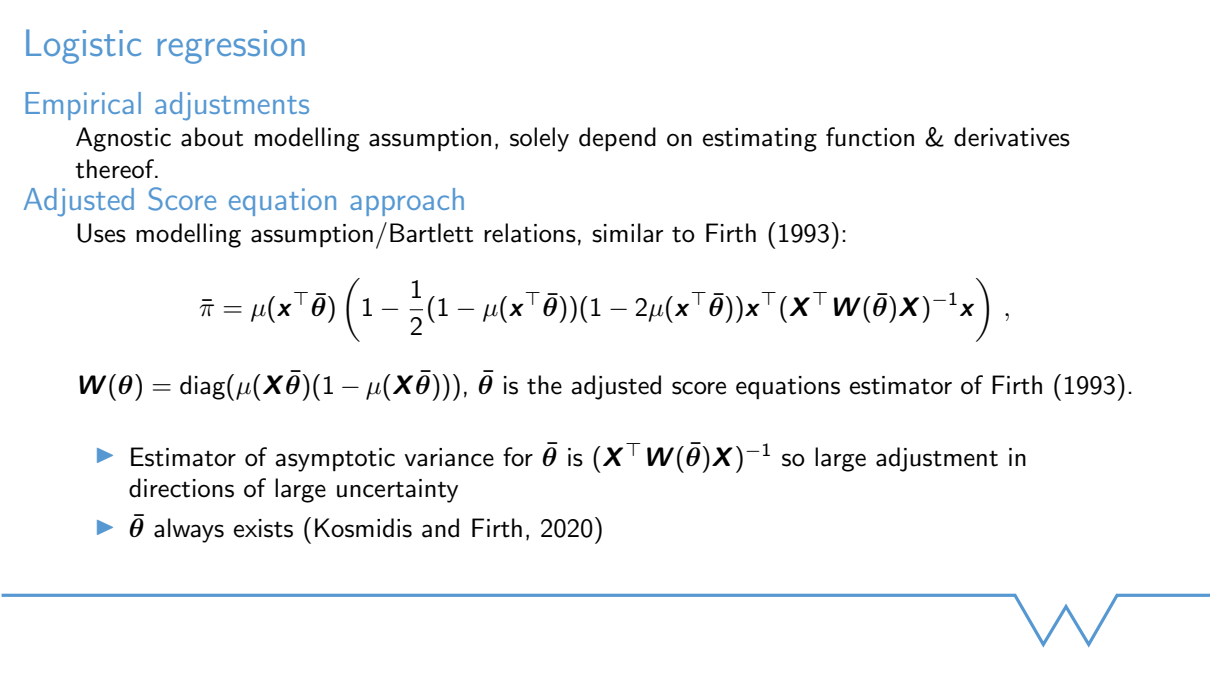
Adjusted Score equation approach

Uses modelling assumption/Bartlett relations, similar to Firth (1993):

$$\bar{\pi} = \mu(\boldsymbol{x}^\top \bar{\boldsymbol{\theta}}) \left(1 - \frac{1}{2}(1 - \mu(\boldsymbol{x}^\top \bar{\boldsymbol{\theta}}))(1 - 2\mu(\boldsymbol{x}^\top \bar{\boldsymbol{\theta}}))\boldsymbol{x}^\top (\boldsymbol{X}^\top \boldsymbol{W}(\bar{\boldsymbol{\theta}})\boldsymbol{X})^{-1}\boldsymbol{x}\right),$$

$\boldsymbol{W}(\boldsymbol{\theta}) = \text{diag}(\mu(\boldsymbol{X}\bar{\boldsymbol{\theta}})(1 - \mu(\boldsymbol{X}\bar{\boldsymbol{\theta}})))$, $\bar{\boldsymbol{\theta}}$ is the adjusted score equations estimator of Firth (1993).

- Estimator of asymptotic variance for $\bar{\boldsymbol{\theta}}$ is $(\boldsymbol{X}^\top \boldsymbol{W}(\bar{\boldsymbol{\theta}})\boldsymbol{X})^{-1}$ so large adjustment in directions of large uncertainty
- $\bar{\boldsymbol{\theta}}$ always exists (Kosmidis and Firth, 2020)

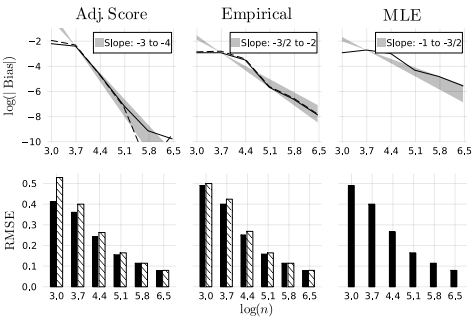


Logistic regression
Simulation study

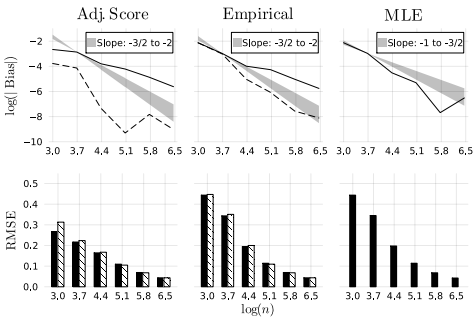
Following the DGP of Puhr et al. (2017), which was designed to give nontrivial simulation designs for logistic regression, consider

- $\boldsymbol{x} \in \Re^p, \boldsymbol{X} \in \Re^{n \times p}, \boldsymbol{\theta}_0 \in \Re^p$, $n \in \{20, 40, \ldots, 2560\}, p = 11$
- $\boldsymbol{x}, \boldsymbol{X}$ drawn once, repeated draws of $\boldsymbol{y}$
- Predictions obtained using (i) MLE plug-in and (ii) bias reduced prediction with two empirical and two adjusted score equation adjustments

Logistic regression
Simulation study



(a) $\boldsymbol{x} : \mu(\boldsymbol{x}^\top \boldsymbol{\theta}_0) = 0.5$

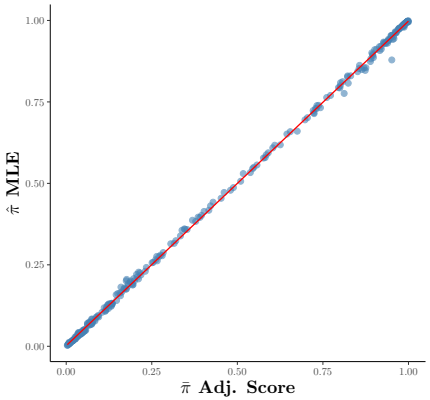


(b) $\boldsymbol{x} : \mu(\boldsymbol{x}^\top \boldsymbol{\theta}_0) = 0.1$

Real data example

Statlog (Heart) Data ($n = 270, p = 14$)

Predict probability of heart disease given covariates with leave-one-out sample split



Conclusions & Extensions

We developed a novel improved prediction approach applicable to a large class of models. The developed theory is backed by empirical results

Future research

- Consider highly nonlinear models where prediction bias is more pronounced
- Extend to discontinuous prediction functions $g(\boldsymbol{x}, \boldsymbol{\theta})$, e.g. to predict categorical responses
- Higher dimensional asymptotic regimes ($p \to \infty$)

References

Statlog (Heart). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C57303.

Amemiya, T. (1986, c1985). *Advanced econometrics*. Oxford: Basil Blackwell.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Gourieroux, C. and A. Monfort (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics 17*(1), 83–97.

Kosmidis, I. and D. Firth (2020, 08). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika 108*(1), 71–82.

Kosmidis, I. and N. Lunardon (2020). Empirical bias-reducing adjustments to estimating functions. *arXiv preprint arXiv:2001.03786*.

Puhr, R., G. Heinze, M. Nold, L. Lusa, and A. Geroldinger (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in medicine 36*(14), 2302–2317.

Stefanski, L. A. and D. D. Boos (2002). The calculus of m-estimation. *The American Statistician 56*(1), 29–38.

M-estimation perspective

RB*M*-estimation (Kosmidis and Lunardon, 2020)

If the following M-estimation regularity conditions hold

- Consistency: $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathcal{O}(n^{-1/2})$
- Unbiasedness: $\mathrm{E}_G(\boldsymbol{\psi}^i(\boldsymbol{\theta}_0)) = \mathbf{0}_p$
- Smoothness: Derivatives of $\boldsymbol{\psi}^i(\boldsymbol{\theta})$ exist up to fourth order in a neighbourhood of $\boldsymbol{\theta}_0$
- Boundedness:
  - Central moments of derivatives estimating equations: $\mathcal{O}(n^{1/2})$,
  - Joint central moments of derivatives estimating equations: $\mathcal{O}(n^{b/2})$, b is the number of terms

then

$$E_G(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathcal{O}(n^{-3/2}), \quad \boldsymbol{Q}(\boldsymbol{\theta}_0)^{1/2}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{I}_p)\,,$$

for some matrix $\boldsymbol{Q}(\boldsymbol{\theta}_0)$ for which a consistent estimator exists.