Bias correction in GLMs

Bicko, Jonathan & Ben

2021 Jun 21 (Mon)

Introduction

Generating predictions from regression models can be challenging and depends on whether the relationship between the response variable and the predictors is linear or nonlinear. For example, if response, Y, changes nonlinearly with the predictor variable, X, averaged response variable with respect to the predictor, $E(Y(X))$, does not necessarily equal the response at the mean of the predictor, $Y(E(X)$. This can be understood in relation to Jensen's inequality which states that, for a nonlinear function, $Y(X)$, then $E(Y(X)) > Y(E(X))$, if $Y(X)$ is positive second derivative; and $E(Y(X)) < Y(E(X))$ if $Y(X)$ is negative second derivative. Currently, existing R packages for predicting responses make distribution assumptions about the non-focal predictors, (for example, averaging the non-focal predictors at various levels of focal predictors), leading to potential biasness if the particular predictor is not well represented, or as a result of nonlinear averaging. To understand what we are trying to achieve, we first define the following:

Prediction plots

Notation :

- x_f: a value of the focal predictor
- $x_{\{n\}}$: a vector of values of the non-focal predictors for a particular observation
- $\eta(x_f, x_{\{n\}}) = \beta_f x_f + \sum \beta_{\{n\}} x_{\{n\}}$ = linear predictor (e.g. prediction on the log-odds scale)
- $g^{-1}()$: inverse-link function (e.g. logistic)
- $D(x_{\{n\}}|x_f)$: distribution of the non-focal predictors conditional on a particular value of the focal predictor
- β_{fi}: the coefficient describing the interaction(s) of the focal and non-focal parameters

The purpose and goal of a *prediction plot* seems fairly straightforward; for specified values of (a) focal predictor(s), we want to give a point estimate and confidence intervals for the prediction of the model for a "typical" (= random sample over the multivariate distribution of non-focal parameters) individual with those values of the predictors.

BB's

("what is the expected probability of having clean water for a 25-year-old male"?). These are most of the problems we have actually been addressing with our bias correction stuff above. To do these computations, we need to take the values of the non-focal predictors (or their means and covariances) from the *conditional distribution.* (If the focal predictors are discrete, we condition on exact values; if they are continuous/have mostly unique values, we condition on appropriately sized bins.) In other words,

$$\underset{D(x_{\{n\}}|x_f)}{\text{mean}} g^{-1}\left(\eta(x_{\{n\}}, x_f)\right)$$

Suppose we consider the values of non-focal predictors, to implement this:

- compute linear predictor of the non-focal predictors, $\eta_{\{n\}} = \sum \beta_{\{n\}} x_{\{n\}}$

- find a list of vectors of observations of $\eta_{\{n\}}$ associated with each value (bin) of the focal predictor, $\eta_{j\{n\}}$, $j = 1, 2, \cdots$
- for each $\eta_{j\{n\}}$:
  - compute $\hat{y}_j = \text{mean } g^{-1}\left(\beta_f x_{j_f} + \eta_{j\{n\}}\right)$

What's the expected probability of having clean water for all x-year-old with the ***corresponding*** *wealth index??*

JD's and **SC**'s

- compute linear predictor of the non-focal predictors, $\eta_{\{n\}} = \sum \beta_{\{n\}} x_{\{n\}}$
- for every value of the focal predictor, x_{j_f}:
  - compute $\hat{y}_j = \text{mean } g^{-1}\left(\beta_f x_{j_f} + \eta_{\{n\}}\right)$

The major difference between **BB**'s and **JD**'s approach is that in the former case, we consider non-focal linear predictor corresponding to specific value (bin) of the focal predictor, while in the later case, for each value of the focal predictor, we consider linear predictor associated with entire population.

Questions

- What does the second case (JD's) represent?
- Which is the correct approach?

Simulation

We implement and apply these methods in the context of both simple generalized linear and mixed effect models, using simulated data sets – for univariate and multivariate models.

We start with a univariate case where we have only one predictor.

Simple fixed effect model

$$
\begin{aligned}
\text{logit}(\text{status} = 1) &= \eta \\
\eta &= \beta_0 + \beta_{\text{A}}\text{Age} + \beta_{\text{W}}\text{Wealthindex} \\
\text{Age} &\sim \text{Normal}(0.2, 1) \\
\text{Wealthindex} &\sim \text{Normal}(0, 1) \\
\beta_0 &= 1.5 \\
\beta_{\text{A}} &= 1.0 \\
\beta_{\text{W}} &= 2
\end{aligned}
$$

Considering $N = 1e4$

```r
simplesim <- function(N=1e4, beta0=1.5, betaA=1.0, betaW=2
        , age_sd=1, age_mean=0.2
        , wealth_sd=1, wealth_mean=0
    ) {
    age <- rnorm(N, age_mean, age_sd)
    wealthindex <- rnorm(N, wealth_mean, wealth_sd)
    eta <- beta0 + betaA * age + betaW * wealthindex
    sim_df <- (data.frame(age=age, wealthindex=wealthindex, eta=eta)
        %>% mutate(status = rbinom(N, 1, plogis(eta)))
        %>% select(-eta)
    )
    return(sim_df)
}
```

```
sim_df <- simplesim()
true_prop <- mean(sim_df$status)
print(true_prop)
```

```
## [1] 0.7185
```

```
head(sim_df)
```

```
##           age wealthindex status
## 1 -0.59429693   1.2414996      1
## 2 -0.62138924   0.3860569      1
## 3  0.26417924   0.6529173      1
## 4 -0.07500945   0.3314702      1
## 5  0.34401336  -0.2215855      1
## 6 -0.31897466  -0.1013064      1
```

Simple logistic model

```
simple_mod <- glm(status ~ age + wealthindex, data = sim_df, family="binomial")
```

Variable predictions

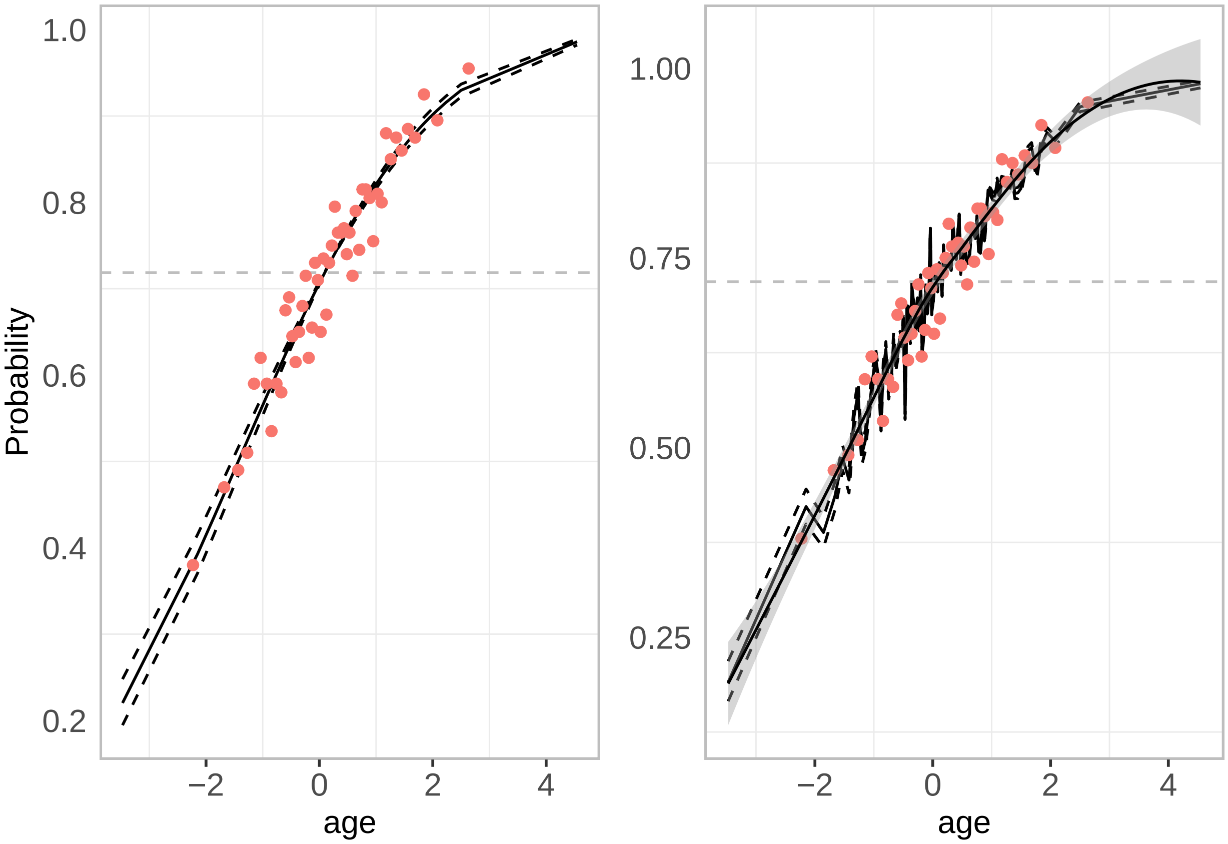

Figure 1: Prediction plots with N=1e4: JD's approach on the left; BB's on the right.

Considering $N = 1e5$ and then refit the model

```
sim_df <- simplesim(N=1e5)
true_prop <- mean(sim_df$status)
simple_mod <- glm(status ~ age + wealthindex, data = sim_df, family="binomial")
```

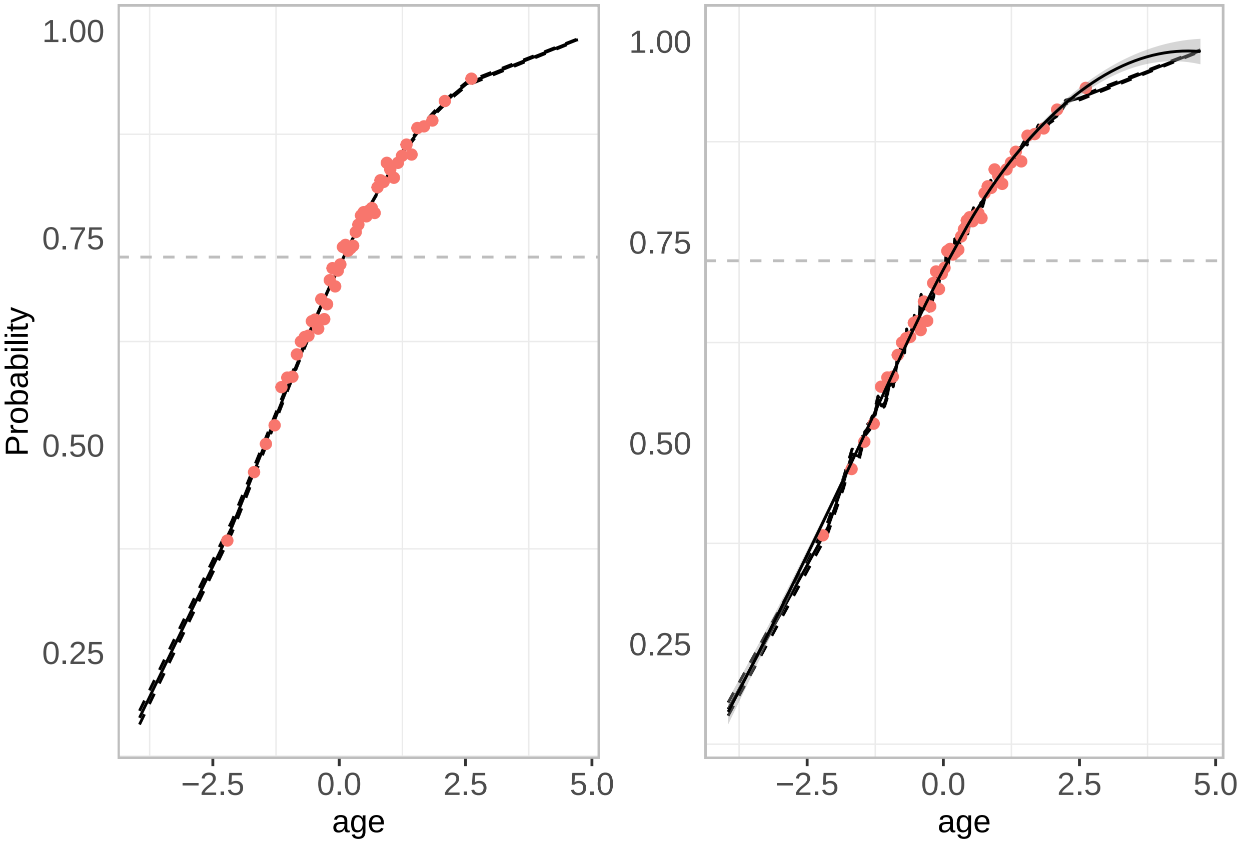

Figure 2: Prediction plots with N=1e5: JD's approach on the left; BB's on the right.