Bias correction in GLMs

Bicko, Jonathan & Ben

2021 Jun 21 (Mon)

Introduction

Prediction of the variation in the response variable depends on whether the relationship between the response variable and the predictors is linear or nonlinear. For example, when 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. Most packages for predicting responses make distribution assumptions about the non-focal predictors, (for example, conditioning at the mean value of the predictor), leading to potential biasness if the particular predictor is not well represented, or as a result of nonlinear averaging. We consider the following approaches for bias correction:

- Population averaging
  - Whole population
  - Quantiles
- Second-order correction

and provide a comparison with the uncorrected *Distributional conditioning* (varpred or emmeans approach).

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

$$
\begin{aligned}
\mathrm{logit}(\mathrm{status} = 1) &= \eta \\
\eta &= \beta_0 + \beta_{\mathrm{A}}\mathrm{Age} + \beta_{\mathrm{W}}\mathrm{Wealthindex} \\
\mathrm{Age} &\sim \mathrm{Normal}(0.2, 1) \\
\mathrm{Wealthindex} &\sim \mathrm{Normal}(0, 1) \\
\beta_0 &= 0.7 \\
\beta_{\mathrm{A}} &= 0.2 \\
\beta_{\mathrm{W}} &= 0.5
\end{aligned}
$$

```
N <- 1e4
beta0 <- 0.7
betaA <- 0.2
betaW <- 0.5
```

```r
age_max <- 1
age_min <- 0.2
age <- runif(N, age_min, age_max)
# age <- rnorm(N, age_max, age_max)

wealthindex <- rnorm(N, 0, 1)

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

```
## [1] 0.6916
```

```r
head(sim_df)
```

```
##         age wealthindex status
## 1 0.8452918   1.1198420      1
## 2 0.2563395  -0.6219684      0
## 3 0.4192913  -1.5949657      1
## 4 0.7882493  -1.2565989      1
## 5 0.3893671   1.7148530      1
## 6 0.8806260  -0.1938844      1
```

Simple logistic model

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

Variable predictions

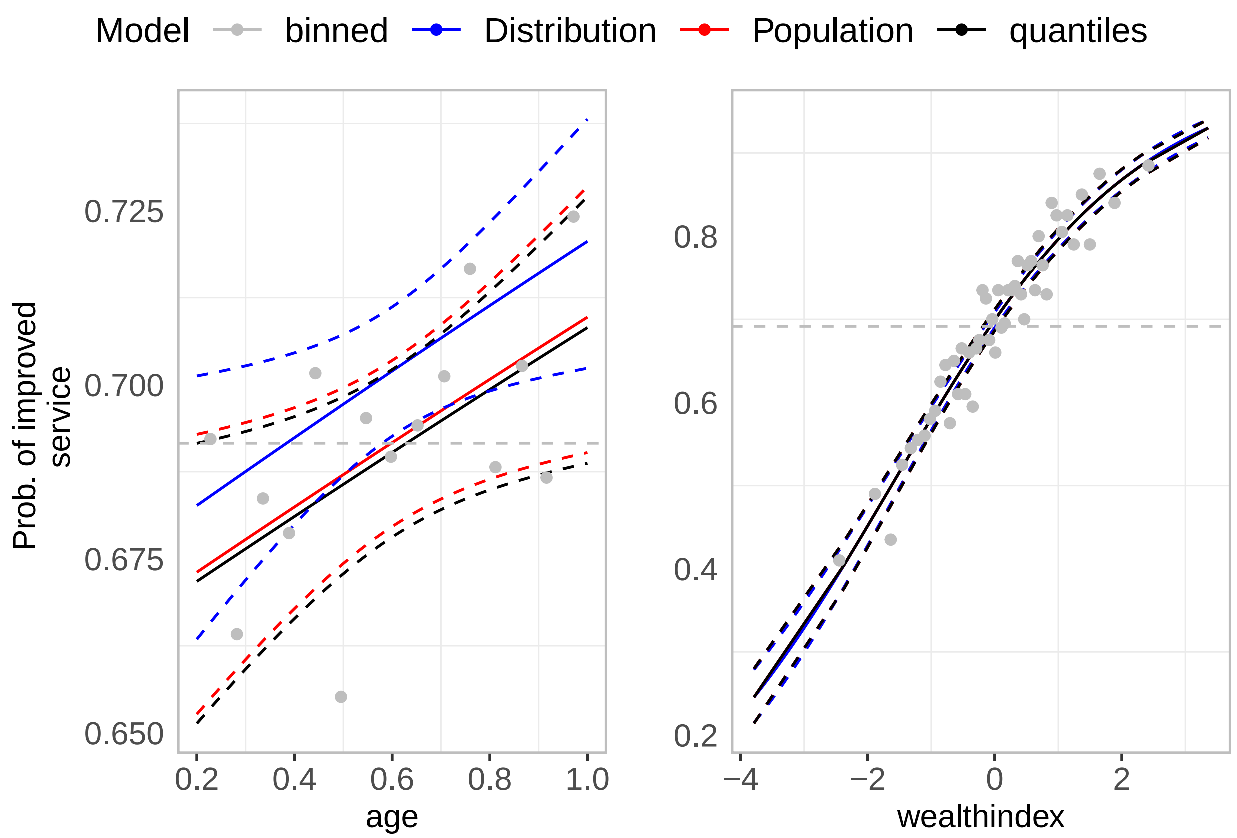

Figure 1: A comparison of variable predictions with uncorrected and corrected estimates. The uncorrected Distribution approach is based on averaging of non-focal predictors; Population approach involves using the observed values of non-focal predictors together with the quantiles of the focal predictors; while quantiles approach involves sampling (or picking quantiles of) the observed values of both focal and non-focal predictors. The uncorrected Distribution approach seems to over-predict and is slightly higher then the marginal estimates.

The marginal mean is 0.6916 while the estimated are:

- age: quantiles 0.690167; population 0.6915805; distribution 0.7018437
- wealthindex: quantiles 0.690167; population 0.6901678; distribution 0.6404765

Random effect model

```
# Simulation parameters
nHH <- 100  # Number of HH (primary units) per year

nyrs <- 50  # Number of years to simulate
yrs <- 2000 + c(1:nyrs) # Years to simulate
N <- nyrs * nHH

## HH random effect sd
hhSD <- 0.5

# Generate dataset template
```

```
temp_df <- (data.frame(hhid = rep(c(1:nHH), each = nyrs)
        , years = rep(yrs, nHH)
        , age = runif(n=N, age_min, age_max)
        , wealthindex = rnorm(n = N, 0, 1)
    )
)

# Simulate HH-level random effects (residual error)
hhRE <- rnorm(nHH, hhSD)
temp_df$hhRE <- hhRE[temp_df$hhid]

sim_df <- (temp_df
    %>% mutate(eta = beta0 + betaA * age + betaW * wealthindex + hhRE
        , status = rbinom(N, 1, plogis(eta))
    )
    %>% select(-eta)
)
true_prop_reff <- mean(sim_df$status)
print(true_prop)
```

```
## [1] 0.6916
```

```
print(head(sim_df, 10))
```

```
##    hhid years       age wealthindex     hhRE status
## 1     1  2001 0.9018686 -0.99696269 0.514887      1
## 2     1  2002 0.3506389  0.68378134 0.514887      1
## 3     1  2003 0.4490254 -0.18218552 0.514887      1
## 4     1  2004 0.9666503  0.01820703 0.514887      1
## 5     1  2005 0.5382138 -1.48422812 0.514887      1
## 6     1  2006 0.2568336  1.06422463 0.514887      0
## 7     1  2007 0.9850446  0.08233927 0.514887      1
## 8     1  2008 0.2927210 -1.02046500 0.514887      0
## 9     1  2009 0.8826939  2.11619121 0.514887      1
## 10    1  2010 0.2453546 -0.31856268 0.514887      1
```

Fit model

```
reff_mod <- glmmTMB(status ~ age + wealthindex + (1|hhid)
    , data = sim_df
    , family = binomial(link = "logit")
)
```

Variable predictions

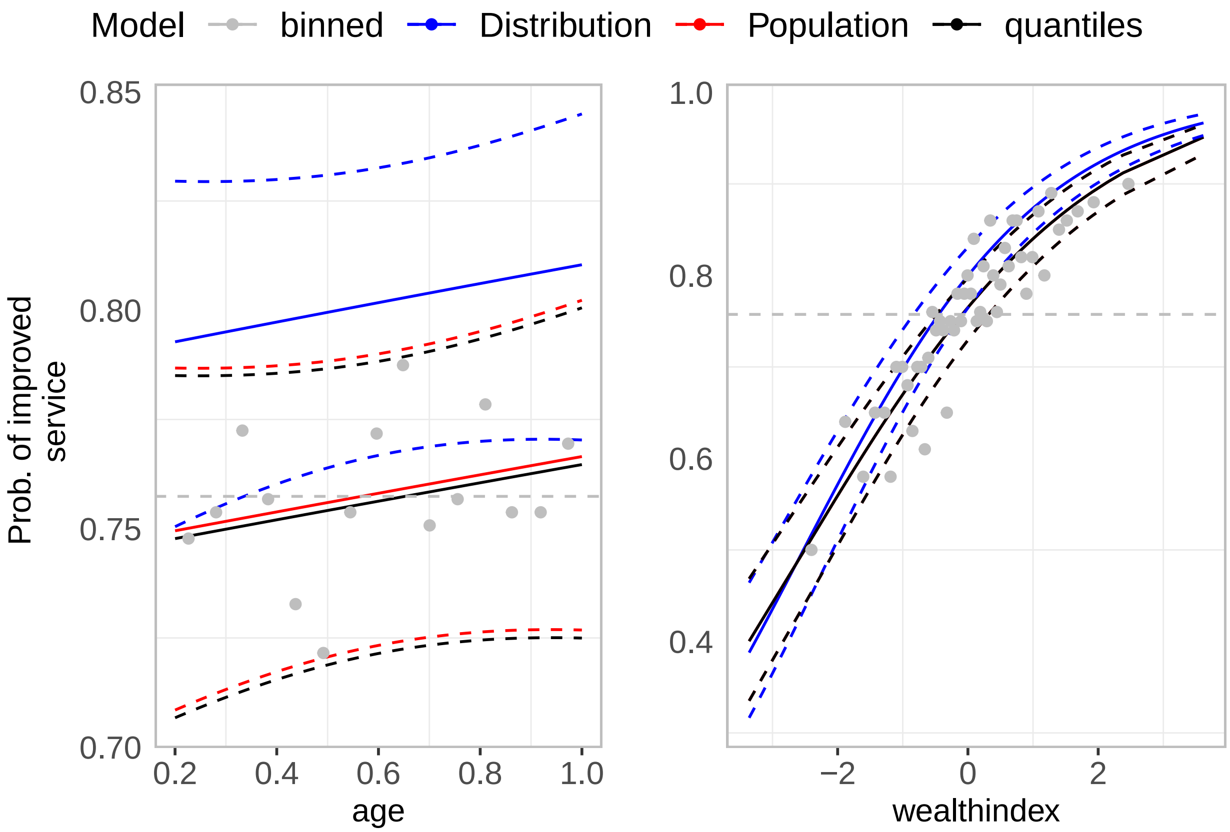

Figure 2: Similar to the previous simple model, the quantile and population-based averaging gives very close estimates as opposed to the uncorrected distribution-based which appears to over-predict.

We can also add centered predictions as in the case of *varpred*:

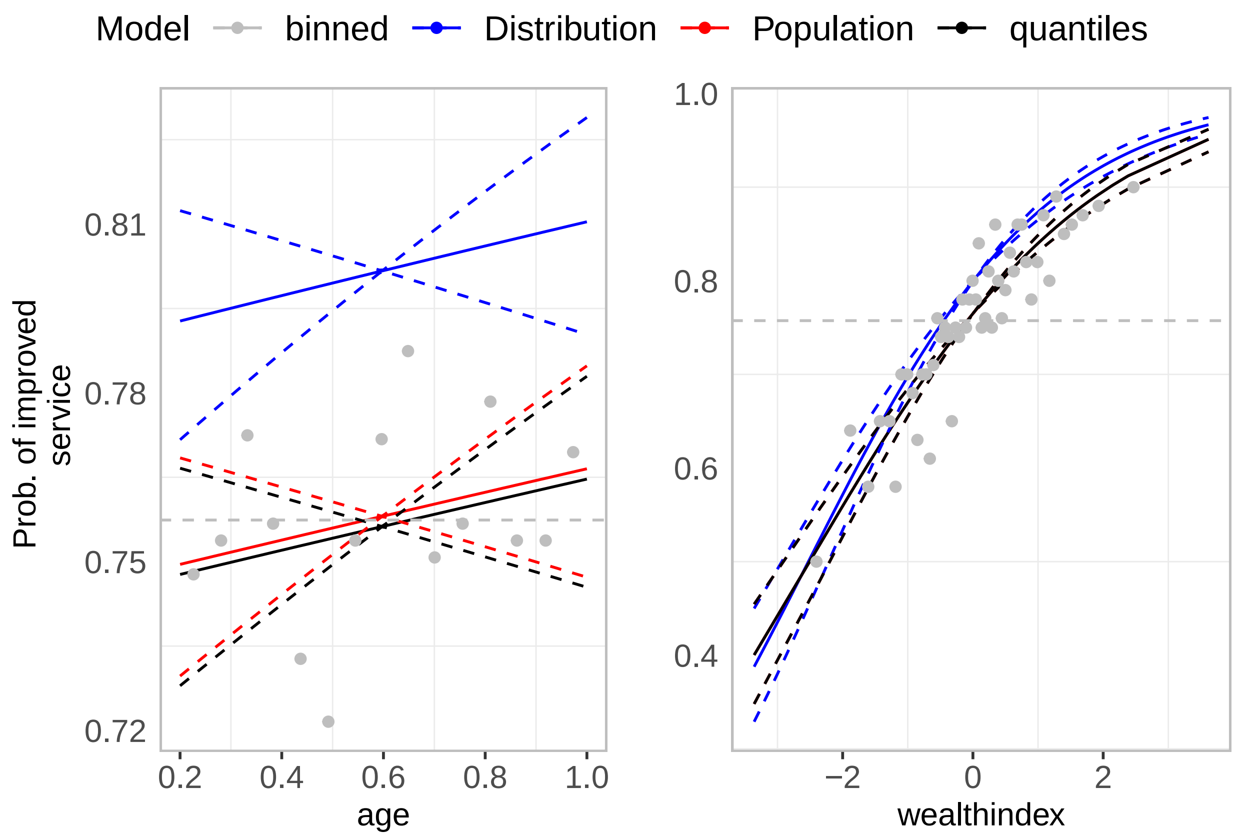

Figure 3: Same plots above but with centered confidence bands...

Multivariate model

We now consider a model with multiple predictors.

Simple logistic model

```
N <- 1e4
beta0 <- 0.7
betaA <- 0.2
betaE <- -0.7
betaW <- 0.5
age_max <- 1
age_min <- 0.2
expend_mean <- 0.2
expend_sd <- 1
age <- runif(N, age_min, age_max)
expenditure <- rnorm(N, expend_mean, expend_sd)
wealthindex <- rnorm(N, 0, 1)
eta <- beta0 + betaA * age + betaW * wealthindex + betaE*expenditure
sim_df <- (data.frame(age=age, expenditure=expenditure, wealthindex=wealthindex, eta=eta)
    %>% mutate(status = rbinom(N, 1, plogis(eta)))
    %>% select(-eta)
)
true_prop <- mean(sim_df$status)
```

```
print(true_prop)
```

```
## [1] 0.6431
```

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

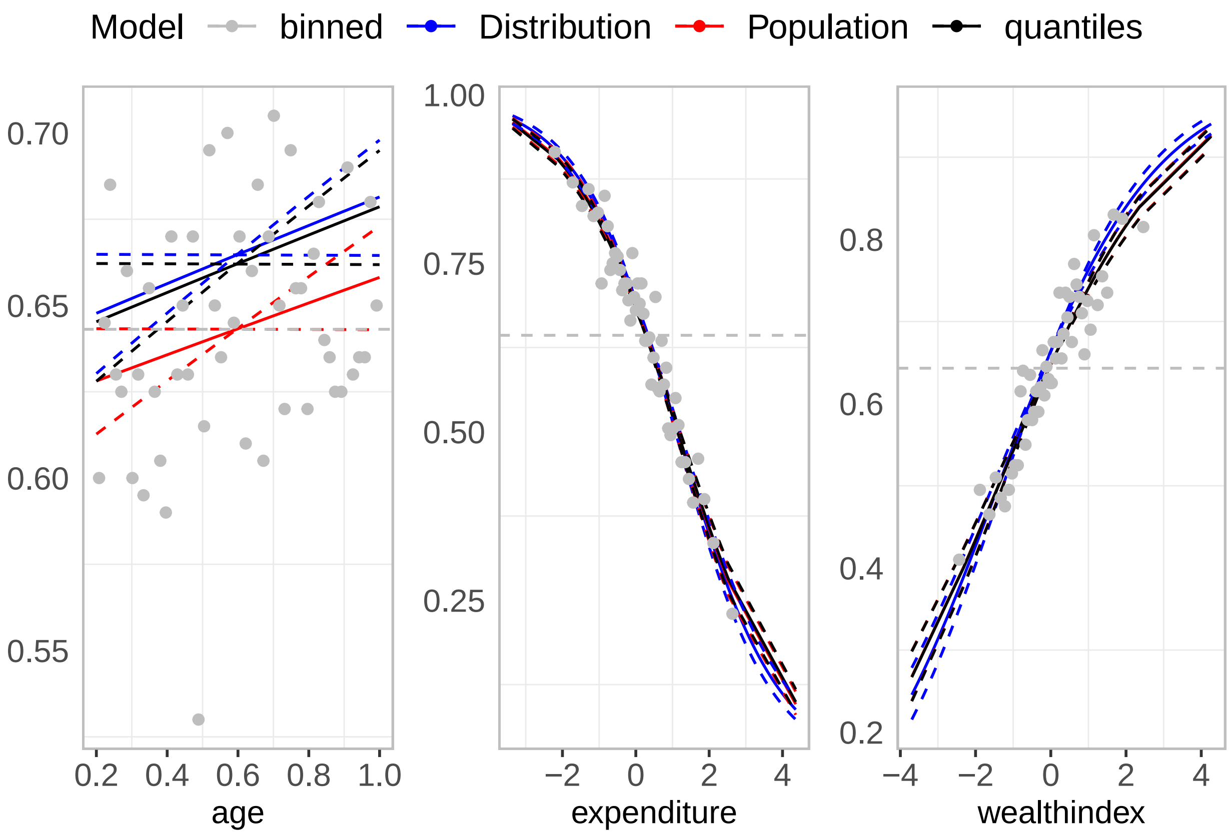

Figure 4: A comparison of bias correction approaches for a simple multivariate logistic model.

Mixed model

```
beta0 <- 0.7
betaA <- 0.2
betaW <- 0.5
betaE <- -0.7
age_max <- 1
age_min <- 0.2
expend_mean <- 0.2
expend_sd <- 1

# Simulation parameters
nHH <- 100 # Number of HH (primary units) per year
nyrs <- 50 # Number of years to simulate
yrs <- 2000 + c(1:nyrs) # Years to simulate
N <- nyrs * nHH
## HH random effect sd
hhSD <- 0.5
# Generate dataset template
```

```
temp_df <- (data.frame(hhid = rep(c(1:nHH), each = nyrs)
        , years = rep(yrs, nHH)
        , age = runif(n=N, age_min, age_max)
        , expenditure = rnorm(N, expend_mean, expend_sd)
        , wealthindex = rnorm(n = N, 0, 1)
    )
)
# Simulate HH-level random effects (residual error)
hhRE <- rnorm(nHH, hhSD)
temp_df$hhRE <- hhRE[temp_df$hhid]
sim_df <- (temp_df
    %>% mutate(eta = beta0 + betaA * age + betaE * expenditure + betaW * wealthindex + hhRE
    , status = rbinom(N, 1, plogis(eta))
    )
    %>% select(-eta)
)
true_prop_reff <- mean(sim_df$status)
print(true_prop_reff)
```

```
## [1] 0.7282
```

```
reff_mod_multi <- glmmTMB(status ~ age + expenditure + wealthindex + (1|hhid)
    , data = sim_df
    , family = binomial(link = "logit")
)
```

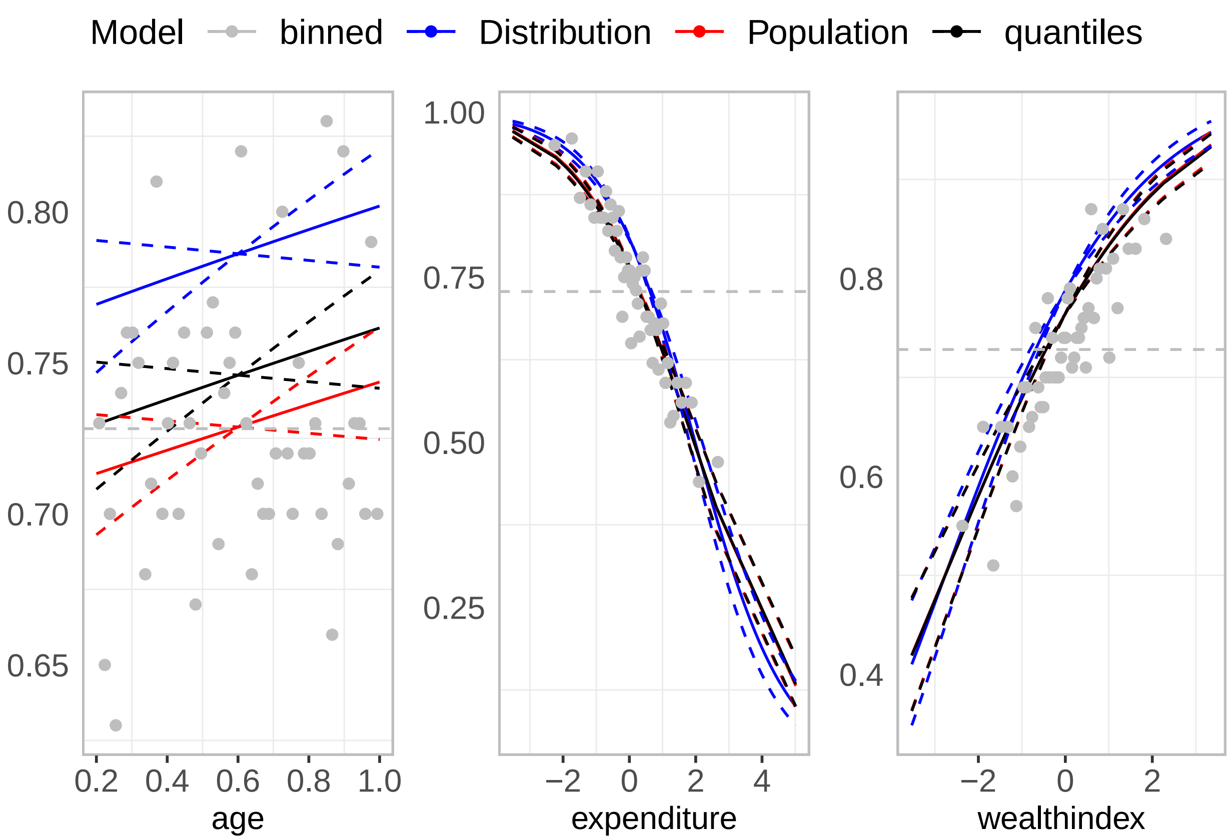

Figure 5: A comparison of bias correction approaches for a reff multivariate logistic model.