Bias correction in GLMs

Bicko, Jonathan & Ben

2021 Jun 21 (Mon)

Introduction

We intend to investigate our prediction based on known truth and any bias potentially introduced by non-linear averaging, conditioning or random effect. We'll start with a simple case of a only fixed effect model and then consider a mixed effect model.

Simulation

We perform a simple simulation for a fixed effect model

$$\begin{aligned}
\text{logit}(\text{status} = 1) &= \eta \\
\eta &= \beta_0 + \beta_{\text{A}}\text{Age} + \beta_{\text{W}}\text{Wealthindex} \\
\text{Age} &\sim \text{Uniform}(0.2, 1) \\
\text{Wealthindex} &\sim \text{Normal}(0, 1) \\
\beta_0 &= 0.7 \\
\beta_{\text{A}} &= 0.3 \\
\beta_{\text{W}} &= 0.6
\end{aligned}$$

```
N <- 1000
beta0 <- 0.7
betaA <- 0.2
betaW <- 0.5

age_max <- 1
age_min <- 0.2
age <- runif(N, age_min, 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] 0.694
```

```r
head(sim_df)
```

```
##         age wealthindex status
## 1 0.8452918   1.5829806      1
## 2 0.2563395  -0.7920202      1
## 3 0.4192913   0.8592506      1
## 4 0.7882493   1.1605790      1
## 5 0.3893671   1.1220068      1
## 6 0.8806260   0.7425947      1
```

Simple logistic model

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

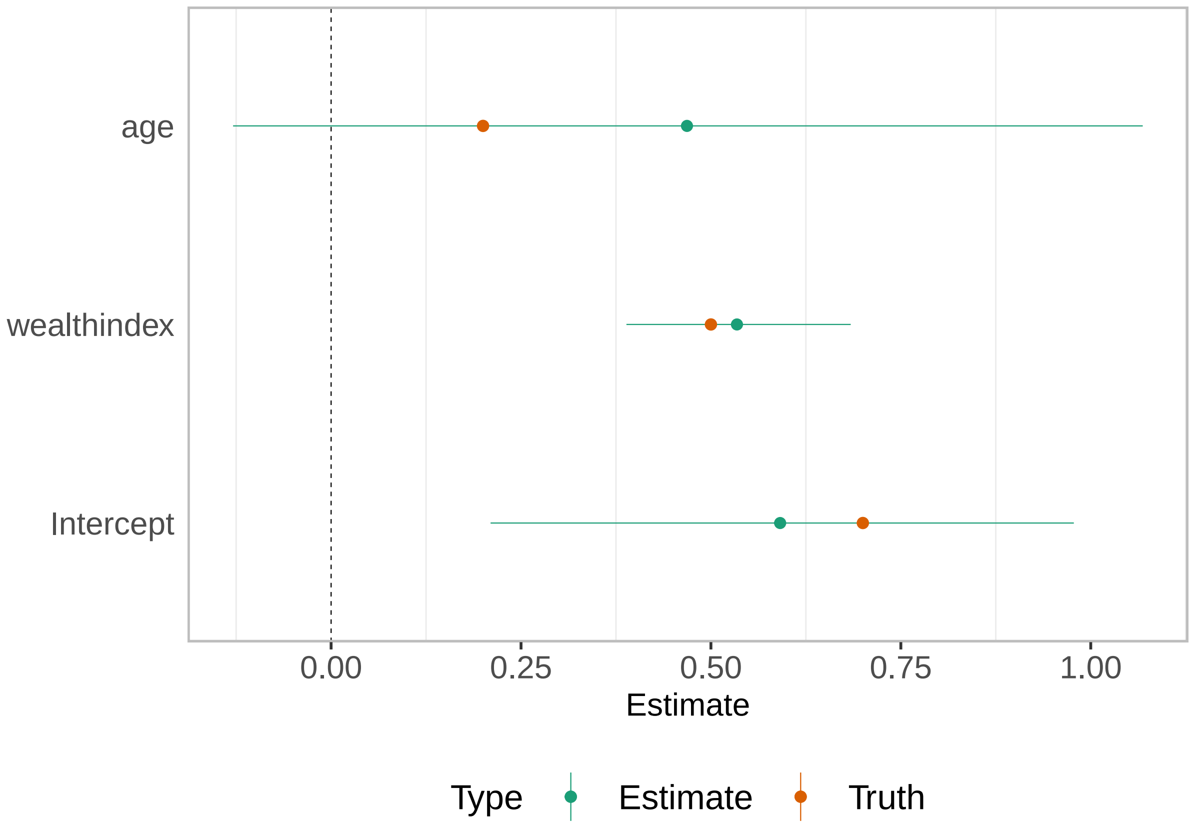
Coefficient plots

```r
## True beta
true_beta_df <- data.frame(term=c("Intercept", "age", "wealthindex")
    , estimate=c(beta0, betaA, betaW)
)

## Tidy coef estimates
coef_df <- (broom::tidy(simple_mod, conf.int=TRUE)
#   %>% dotwhisker::by_2sd(sim_df)
    %>% mutate(term = gsub("\\(|\\)", "", term))
)
print(coef_df)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 Intercept      0.591    0.196       3.02 2.53e- 3    0.210     0.978
## 2 age            0.468    0.305       1.53 1.25e- 1   -0.129     1.07
## 3 wealthindex    0.534    0.0752      7.10 1.24e-12    0.389     0.684
```

```r
simple_coef_plot <- (plotEsize(coef_df)
    + geom_point(data=true_beta_df, aes(x=term, y=estimate, colour="Truth"))
    + labs(colour="Type")
    + scale_colour_brewer(palette="Dark2")
)
print(simple_coef_plot)
```



Variable effect plots – with and without bias adjustment

```
# Age
## Not bias adjusted
simple_vareff_age <- varpred(simple_mod, "age", isolate=TRUE, modelname="Not adjusted")
print(sigma(simple_mod))
```

```
## [1] 1.084869
```

```
## Bias adjusted
simple_vareff_age_adjust <- varpred(simple_mod, "age", isolate=TRUE, bias.adjust=TRUE, modelname="Bias a

vareff_age <- simple_vareff_age
vareff_age$preds <- do.call("rbind", list(vareff_age$preds, simple_vareff_age_adjust$preds))
age_plot <- (plot(vareff_age)
    + labs(y="Prob. of improved \n service", colour="Model")
    + geom_hline(yintercept=true_prop, lty=2, colour="grey")
    + theme(legend.position="bottom")
)

# Wealth index
## Not bias adjusted
simple_vareff_wealthindex <- varpred(simple_mod, "wealthindex", isolate=TRUE, modelname="Not adjusted")

## Bias adjusted
simple_vareff_wealthindex_adjust <- varpred(simple_mod, "wealthindex", isolate=TRUE, bias.adjust=TRUE, m

vareff_wealthindex <- simple_vareff_wealthindex
vareff_wealthindex$preds <- do.call("rbind", list(vareff_wealthindex$preds, simple_vareff_wealthindex_ad
```

```r
wealthindex_plot <- (plot(vareff_wealthindex)
    + labs(y="", colour="Model")
    + geom_hline(yintercept=true_prop, lty=2, colour="grey")
    + theme(legend.position="bottom")
)
ggarrange(age_plot, wealthindex_plot, common.legend=TRUE)
```

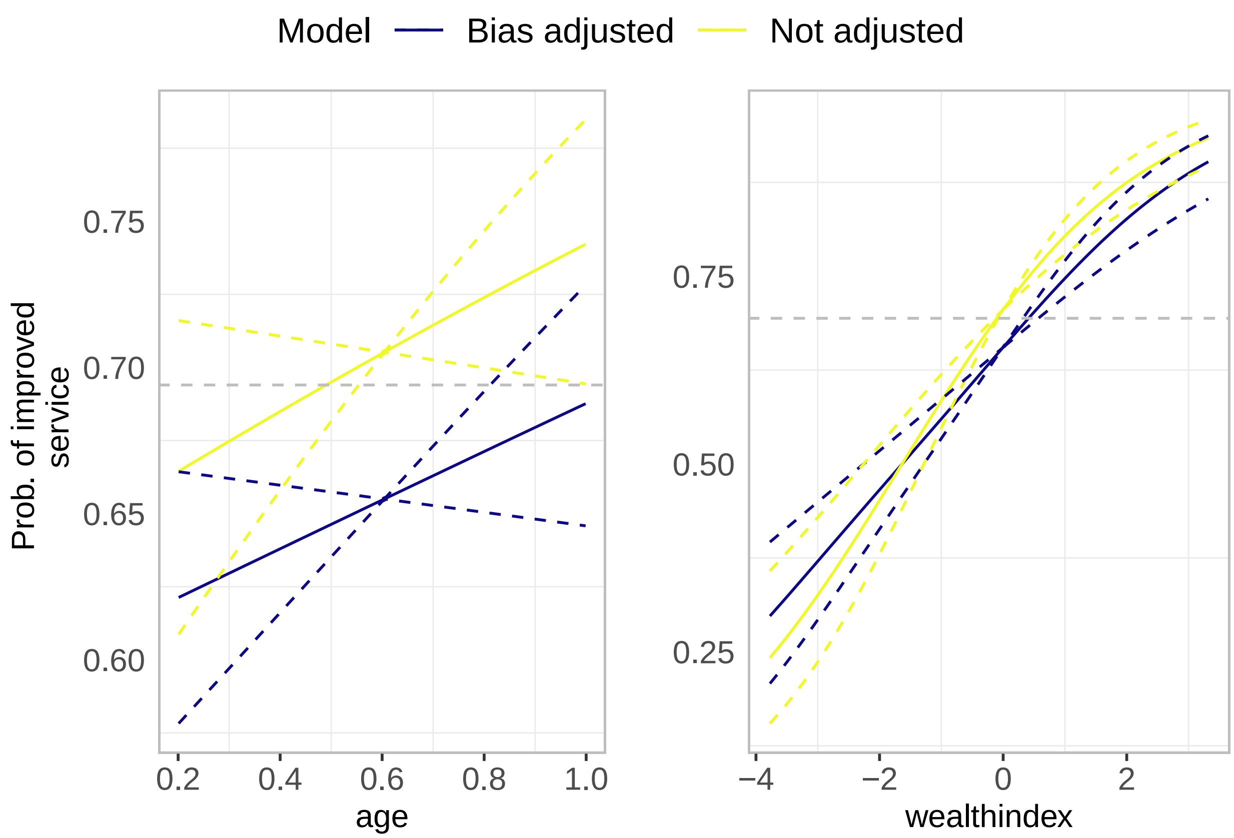


Figure 1: Variable prediction plots comparing bias unadjusted and adjusted predictions. The bias adjusted predictions are slighly lower but close to what we expect from effect size estimates (and their confidence intervals)?

Effect sizes on logit scale:

```r
coef_df_logit <- (coef_df
    %>% select(term, estimate, conf.low, conf.high)
    %>% group_by(term)
    %>% summarise_all(plogis)
)
print(coef_df_logit)
```

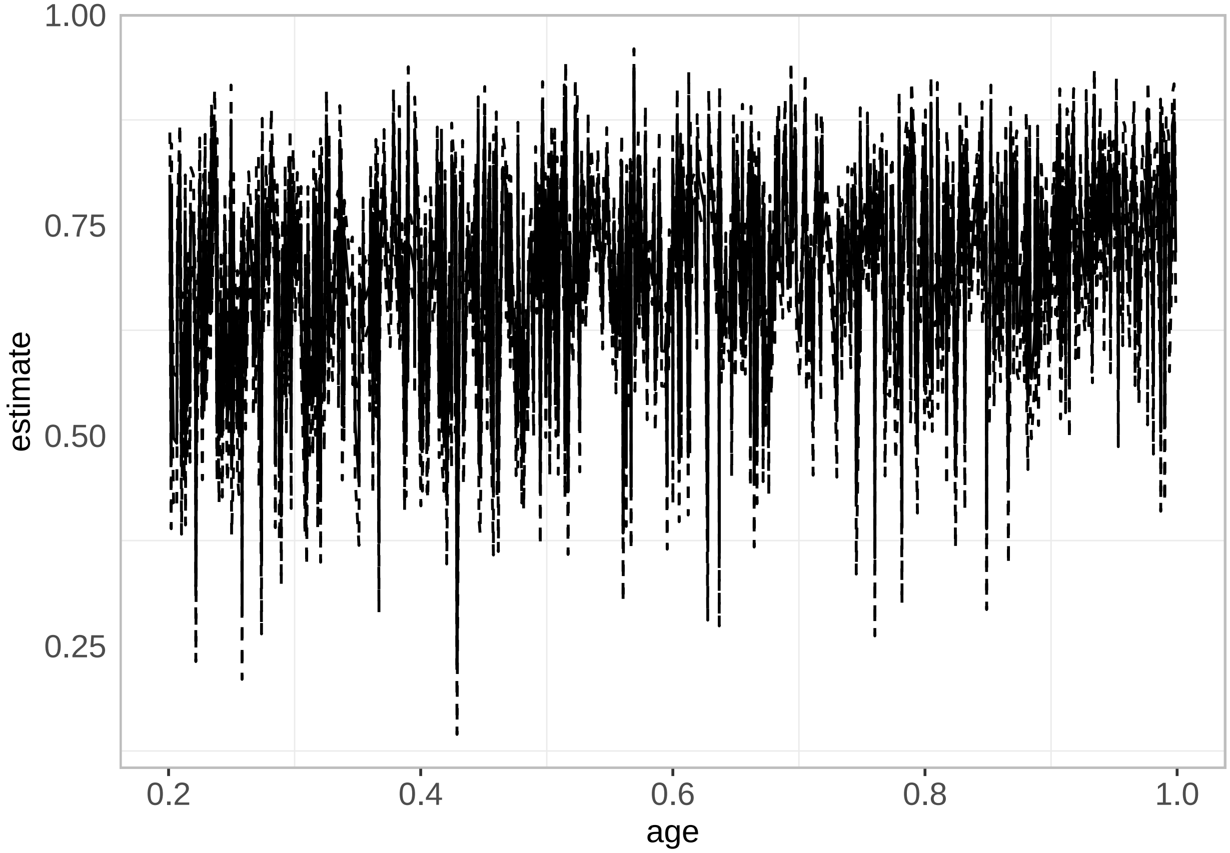
```
## # A tibble: 3 x 4
##   term        estimate conf.low conf.high
##   <chr>          <dbl>    <dbl>     <dbl>
## 1 age            0.615    0.468     0.744
## 2 Intercept      0.644    0.552     0.727
## 3 wealthindex    0.630    0.596     0.665
```

MRP approach

```
pred <- predict(simple_mod, type="link", se.fit=TRUE)
z.val <- qnorm(1 - (1 - 0.95)/2)
pred_df <- (data.frame(age=sim_df$age, wealthindex=sim_df$wealthindex, estimate=pred$fit, se=pred$se.fi
    %>% mutate(lwr = plogis(estimate-z.val*se)
        , upr = plogis(estimate+z.val*se)
        , estimate = plogis(estimate)
    )
)
head(pred_df)
```

```
##         age wealthindex  estimate        se       lwr       upr
## 1 0.8452918   1.5829806 0.8621003 0.1665990 0.8185148 0.8965414
## 2 0.2563395  -0.7920202 0.5715182 0.1319246 0.5073696 0.6333501
## 3 0.4192913   0.8592506 0.7767121 0.1196519 0.7334299 0.8147419
## 4 0.7882493   1.1605790 0.8292678 0.1351624 0.7884376 0.8635814
## 5 0.3893671   1.1220068 0.7978598 0.1378879 0.7507687 0.8379735
## 6 0.8806260   0.7425947 0.8022486 0.1304768 0.7585375 0.8397188
```

```
## Age plot
age_mrp <- (pred_df
    %>% select(age, estimate, lwr, upr)
    %>% group_by(age)
    %>% summarise_all(mean)
)
age_mrp_plot <- (ggplot(age_mrp, aes(x=age, y=estimate))
    + geom_line()
    + geom_line(aes(y=lwr), lty=2)
    + geom_line(aes(y=upr), lty=2)
#   + geom_smooth(aes(ymin=lwr, ymax=upr), stat="identity")
)
print(age_mrp_plot)
```


1.00
0.75
0.50
0.25
estimate
0.2
0.4
0.6
0.8
1.0
age