Bias correction in GLMs

Bicko, Jonathan & Ben

2021 Jun 21 (Mon)

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

—

Introduction

Generating variable 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, predictions from generalized linear (mixed) models (GLM(M)s) may be inaccurate due to a number of reasons:

1. choice of representative values of *focal* variable(s) and appropriate 'model center' for *non-focal* variables in the case of multiple predictors

2. Jensen's inequality and bias in the mean induced by nonlinear transformations of the response variable (e.g., link functions in GLM(M)s)

3. propagation of uncertainty ([how] can we incorporate uncertainty in nonlinear components of (G)LMMs? Should we exclude variation in non-focal parameters?)

Most common way of dealing with the first challenge is taking the taking unique levels of the *focal* variable(s) if they are discrete or taking appropriately sized quantiles (or bins) if they are continuous, and then conditioning these values on the mean of the non-focal predictors. We propose an alternative approach, the population-based approach. The two approaches result in predictions corresponding to each level of the focal predictor i.e., *variable prediction.* Second and third challenges are the focus of this article.

To get an intuition of how conditioning on the mean values of the non-focal predictors work, suppose we are interested in the variable predictions of a particular predictor (hence forth referred as *focal* predictor otherwise *non-focal*), x_f, from the set of predictors. To keep it simple, assume that the model has no interaction terms. Then the idea is to fix the values of *non-focal* predictor(s) at some typical values – typically determined by averaging (for now) in some meaningful way, for example, arithmetic mean and average over the levels of the factors of *non-focal* continuous and categorical predictors, respectively. An alternative to *averaging* is *anchoring* which involves picking a fixed value of the non-focal predictor. One way to achieve this is by averaging the columns of non-focal terms in model matrix, $\mathbf{X}$. For a simple linear model with identity link function, the estimated variable predictions corresponding to focal predictor, x_f, is $\eta(x_f, \bar{x}_{\{n\}}) = \beta_f x_f + \sum \beta_{\{n\}} \bar{x}_{\{n\}}$, where $\bar{x}_{\{n\}}$ are the appropriately averaged entries of non-focal predictors. Almost all existing **R** packages for constructing predictions employ this kind of averaging of the non-focal predictors across the levels of the focal predictors (Lenth and Lenth 2018; Leeper et al. 2017; Fox and Hong 2009). More complicated models and link functions are described in the subsequent sections.

Variable prediction and prediction plot

Variable predictions are individual predictions corresponding to the levels (or unique values) of focal predictor(s) conditioned on the "typical" values of the non-focal predictors. Prediction plot provides a way to describe variable predictions. In particular, 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.

Jensen's inequality and bias in the expected mean

Nonlinear relationship between the response and independent variables can either increase or decrease the predicted outcome, depending on the shape of the link function. A key challenge is understanding how the variability in the focal predictor(s) affect the (mean) predictions of the outcome. Classically, Jensen's inequality provides a way to examine the nature of the link function and deducing its effect on the prediction. Specifically, for a random variable x with mean of $\bar{x}$; a nonlinear function, $f(x)$, the mean of $f(x)$, $\overline{f(x)}$, does not equal the the nonlinear function applied to mean of x, $f(\bar{x})$. When $f(x)$ is accelerating ($f''(x) > 0$), $\overline{f(x)} > f(\bar{x})$; and when $f(x)$ is decreasing ($f''(x) < 0$), $\overline{f(x)} < f(\bar{x})$. In other words, the sign of the difference between $\overline{f(x)}$ and $f(\bar{x})$ depends on the nature of the link function.

For accelerating functions, the Jensen's inequality describes how the changes in the predictors elevate the predicted outcome and describes how these changes depress the predictions. **SC -> JD: graphical illustration?**

In many applications, it usually important to report the estimates that reflect the expected values of the untransformed respense. In such cases, bias-adjustment is needed when back-transforming the predictions to the original scales, due to the biases induced by the nonlinear transformation on the expected mean. More specifically, suppose the transformed response is η, the back-transformed response is $y = g^{-1}(\eta)$. Most common approach for bias-adjustment is second-order Taylor approximation (Lenth and Lenth 2018; Duursma and Robinson 2003). Here, we describe and implement a different approach, *population-based* approach for bias correction.

Population-based approach for bias correction

The most precise (although not necessarily accurate!) way to predict is to condition on a value F of the focal predictor and make predictions for all observations (members of the population) for which $x_f = F$ (or in a small range around $F \ldots$). A key point is that the nonlinear transformation involved in these computations is always *one-dimensional*; all of the multivariate computations required are at the stage of collapsing the multidimensional set of predictors for some subset of the population (e.g. all individuals with $x_f = F$) to a one-dimensional distribution of $\eta(x_f, x_{\{n\}})$.

Once we have got our vector of $\boldsymbol{\eta}$ (which is essentially a set of samples from the distribution over η for the conditional set), we want a mean and confidence intervals on the mean on the response (data) scale, i.e. after back-transforming. Most precisely the mean is $\int P(\eta')g^{-1}(\eta')d\,\eta$:

1. if we use the observations themselves then we just compute the individual values of $g^{-1}(\eta)$ and compute their mean
2. we could compute the quantiles of the distribution of η and use this to construct an approximate Riemann sum over the distribution

In principle, the first approach is more applicable, hence our focus, in our case since we are interested in individual (at each level of the focal predictor) predictions i.e., the *variable prediction*, not necessarily the mean. In addition, we can conveniently compute the mean of $g^{-1}(\eta)$.

There are two ways to compute η when we use the individuals observations, i.e., the first case. In particular for each level of the focal predictor, we can take

1. corresponding values of the non-focal linear (hence forth referred to as binned non-focal linear predictor // **BB**)
2. entire population of the non-focal linear predictor (hence forth referred to as population-based non-focal linear predictor // **JD**)

Binned non-focal linear predictor

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,

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

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 $\hat{y}_j = \text{mean}\ g^{-1}\left(\beta_f x_{j_f} + \eta_{j\,\{n\}}\right)$

Consider an example where we want to predict the probability of having clean water based on age and gender of household head, the prediction $\hat{y}_j$ would represent the expected probability of clean water for a 25-year-old, for example.

If we compute the individual back-transformed predictions for a poorly sampled/finely spaced set of focal values, we will get a noisy prediction line as the values of the non-focal predictors shift across the focal values. Simple example: suppose everyone below the median age has wealth index w_1, everyone above the median has w_2. Then the predicted value will have a discontinuity at the median age. We can deal with this by taking bigger bins (a form of smoothing), or by post-smoothing the results (by loess, for example). The principled form of this would be to assume/recognize that our uneven distribution of observed non-focal predictors actually represents a sample of a distribution that will vary *smoothly* as a function of the focal predictor.

Whole population non-focal linear predictor

Suppose we are now interested in *expected probability of having clean water for all x-year-old across the levels of gender??*, then at each level of the focal predictor, we want to add overall contribution of non-focal predictors. In particular:

- 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 $\hat{y}_j = \text{mean}\ g^{-1}\left(\beta_f x_{j_f} + \eta_{\{n\}}\right)$

Questions

- How to conceptualize **JD** approach. What does it represent in reality?
- Which approach is the most appropriate?

Propagation of uncertainty

What about the confidence intervals (CI)? The limits of the confidence intervals are points, not mean values. In principle, every observation/set of non-focal predictors has a different CI. The predictions are $\eta(x_f, x_{\{n\}})$ or $\eta(x_f, \bar{x}_{\{n\}})$, depending on how we treat non-focal predictors; the variances of the predictions are $\sigma^2 = \mathrm{Diag}(\mathbf{X}^\star \Sigma \mathbf{X}^{\star\top})$, where the entries in $\mathbf{X}^\star$ constitutes appropriately constructed (as per η) non-focal predictors (or corresponding terms) together with appropriate values of the non-focal predictors; and $\Sigma = V(\boldsymbol{\beta})$ is variance-covariance matrix of $\boldsymbol{\beta}$. If the non-focal predictors are averaged, the variances are directly computed from $\mathbf{X}^\star$. However, for the population-based approaches, we follow the same steps described above to compute the distributions of lower/upper CI, $\eta \pm q\sigma$, and compute the transformed mean values, i.e.,

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

where q is an appropriate quantile of the Normal or t distribution.

If we consider only the uncertainty of the focal predictor, so that the confidence intervals are $\eta \pm q\sigma_f$, we can construct *centered* CI by removing the uncertainty associated with the non-focal predictor(s) using either variance-covariance matrix, Σ or centered model matrix, $\mathbf{X}^\star$.

Variance-covariance

The computation of $\hat{\eta}$ remains the same as described above. However, to compute σ, Σ is modified by *zeroing-out* (the variance-covariance of all non-focal predictors are assigned zero) entries of non-focal terms$. This approach requires *centering* of the predictors in the model matrix, X. In other words, the fitted model should have centered predictors i.e., $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}}$.

—

Centered model matrix

Consider centered $\mathbf{X}_c^\star$. It follows that the *non-focal* terms in $\mathbf{X}_c^\star$ are all zero. Consequently the uncertainty due to non-focal predictors are all zeroed-out in the computation of σ. More generally, centered design matrix, $\mathbf{X}_c^\star$, impacts on the estimated value of the intercept and its associated variance but not the slopes. Thus since non-focal terms in $\mathbf{X}_c^\star$ are all zero, it does not matter what their corresponding values are in the variance-covariance matrix. Hence, we can compute variable predictions from non-centered predictors (in other words, fitted models with predictors in their natural scales).

Simulation examples

In this section, we illustrate bias in the context of linear and generalized (mixed) linear models. We compare predictions when non-focal predictors averaging and when population-based approaches are used. Later on, we illustrate and compare our approaches for describing uncertainty in the predictions to existing major **R** packages for prediction.

Simple linear model

Consider a simple simulation

$$\begin{aligned}
y &= \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \epsilon \\
\mathrm{x}_1 &\sim \mathrm{Normal}(0.2, 1) \\
\mathrm{x}_2 &\sim \mathrm{Normal}(0, 1) \\
\epsilon &\sim \mathrm{Normal}(0, 1) \\
\beta_0 &= 1.5 \\
\beta_{\mathrm{x}_1} &= 1.0 \\
\beta_{\mathrm{x}_2} &= 2
\end{aligned}$$

for 10000 observations.

We want to compare the model predicted mean with "true" (marginal) mean, i.e., 1.66. The first step is to fit the model and then construct variable predictions – non-focal predictor averaging (hence forth referred to as "average") and population-based, i.e., binned non-focal linear predictor (hence referred to as "binned-nlp") and whole population non-focal linear predictor (hence referred to as "whole-nlp").

Figure 1 compares variable predictions based on the three approaches highlighted above. All the three approaches give similar estimates (1.664) of expected mean, $\overline{f(x)}$, and very close to the observed mean, $f(\bar{x})$. In other words, for the simple linear model, $\overline{f(x)} \approx f(\bar{x}) = 1.66$. For perfect predictions, we expect the blue, grey and the black (or red) to intersect at the same (or very close) point. Additional smoothing step is needed for binned-nlp, Figure 1c, in order to generate smooth trend lines from noisy predictions.

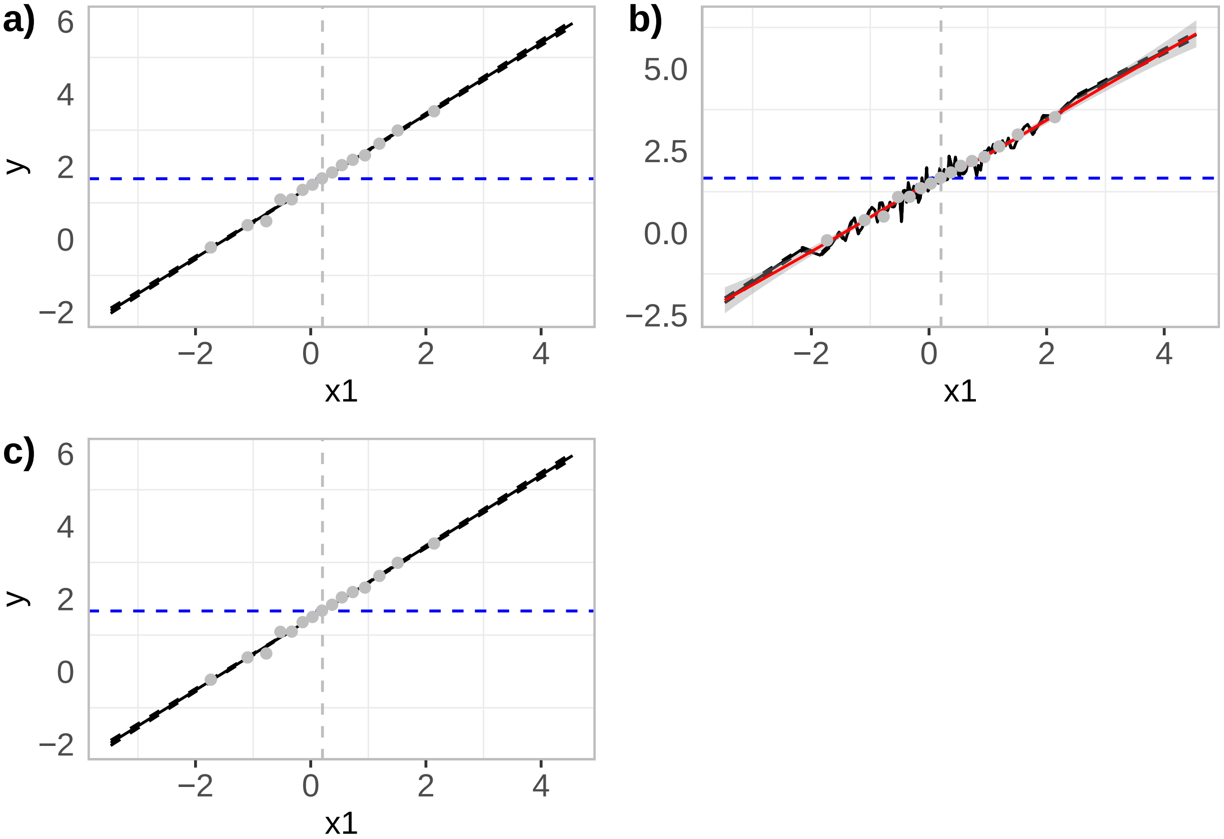

Figure 1: A comparison of variable predictions. The dotted blue and grey horizontal lines are the expected and observed/true means, respectively. The grey dots are the binned observations. The vertical grey line is the mean of the focal predictor (model center). In a) non-focal predictors averaged across the levels of focal predictor; b) non-focal linear predictor is binned across the levels of the focal predictor; and c) whole population non-focal linear predictor is used for each level of the focal predictor. In c, the noisy predictions, i.e., black lines, are smoothened resulting to the red trend-line with the corresponding CI represented by the grey shading.

Simple generalized linear model

Consider binary outcome simulation for improved (1) or unimproved (0) water services status, together with two socio-economic variables (age and wealth index), such that:

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

Figure 2 shows predicted probability of improved services for a particular year old household head. In comparison to binned-nlp ($\overline{f(x)} = 0.732$) and whole-nlp ($\overline{f(x)} = 0.732$) which closely estimate the expected marginal probability ($\overline{f(x)} \approx f(\bar{x}) = 0.73$), averaged non-focal predictors approach (Figure 2a) over-predicts the expected probabilities with $\overline{f(x)} = 0.82 > f(\bar{x})$. Because of the nonlinear link function in the model, our prediction are likely to be biased. However since population-based approaches (binned-nlp and whole-nlp) incorporate bias correction in their construction we see better estimates as opposed to averaged non-focal predictors approach.

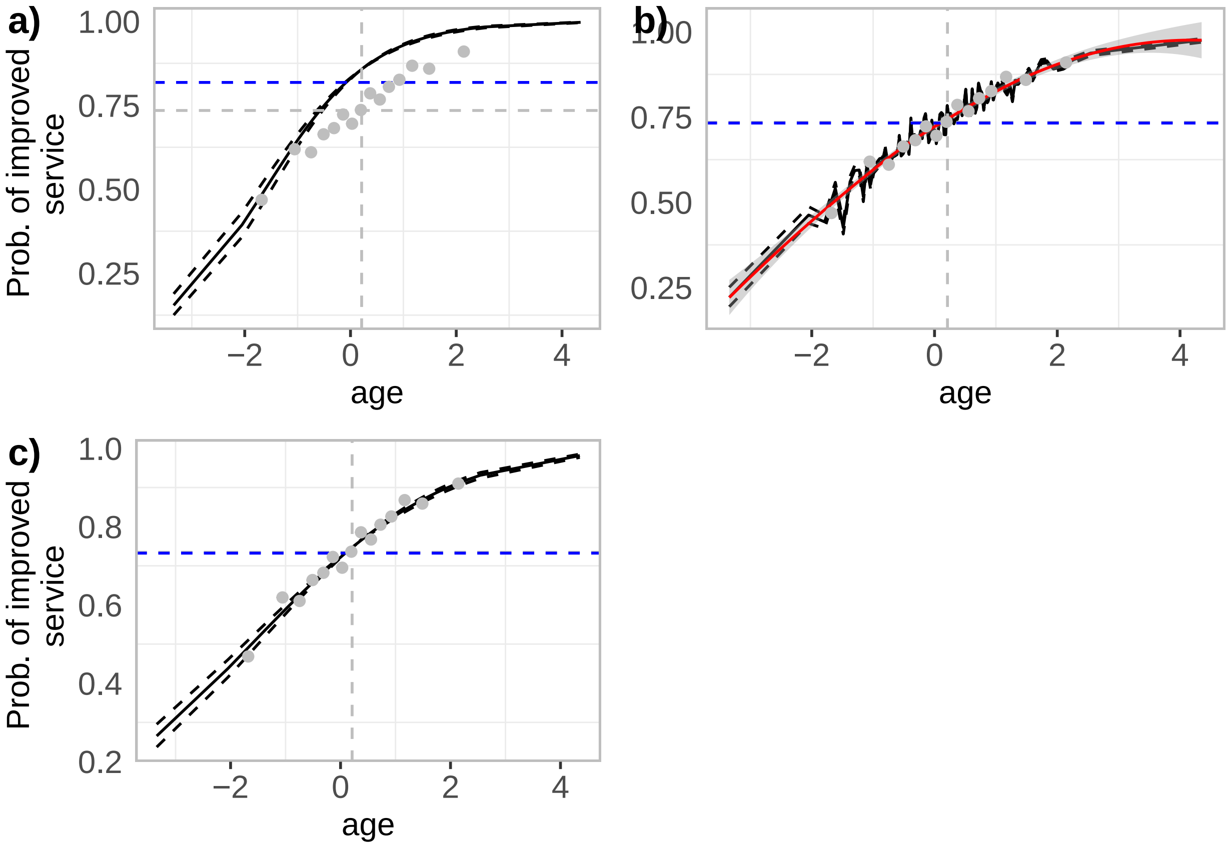

Figure 2: A comparison of variable predictions. The dotted blue and grey horizontal lines are the expected and observed/true means, respectively. The grey dots are the binned observations. The vertical grey line is the mean of the focal predictor (model center). In a) non-focal predictors averaged across the levels of focal predictor; b) non-focal linear predictor is binned across the levels of the focal predictor; and c) whole population non-focal linear predictor is used for each level of the focal predictor. In c, the noisy predictions, i.e., black lines, are smoothened resulting to the red trend-line with the corresponding CI represented by the grey shading.

Mixed models

We now consider more complex model involving fixed and random effects. We start with simple ones and then evolve to more complicated models.

One grouping factor, linear random effect model

Suppose in the first simulation (but now only a single predictor), the observations are recorded more than once from a number. In particular, let H be the number of households indexed by the grouping factor, and

$h[i]$ be household of the ith observation, so that:

$$
\begin{aligned}
y_i &= \beta_0 + \alpha_{h[i]} + \beta_{\mathrm{x}}\mathrm{x}_i + \epsilon_i \\
\alpha_h &\sim \mathrm{Normal}(0, 2), \quad \mathrm{h} = 1, \cdots, \mathrm{H} \\
\epsilon_i &\sim \mathrm{Normal}(0, 1) \\
\mathrm{x}_i &\sim \mathrm{Normal}(0.2, 1), \quad \mathrm{i} = 1, \cdots, \mathrm{n} \\
\beta_0 &= 1.5 \\
\beta_{\mathrm{x}} &= 1.0
\end{aligned}
$$

```
## tibble [10,000 x 3] (S3: tbl_df/tbl/data.frame)
##  $ hhid: int [1:10000] 1 1 1 1 1 1 1 1 1 1 ...
##  $ x   : num [1:10000] -2.291 -0.116 -0.159 -0.387 0.892 ...
##  $ y   : num [1:10000] -1.996 -0.5 -1.318 -1.244 -0.322 ...
```

```
##  Family: gaussian  ( identity )
## Formula:          y ~ x + (1 | hhid)
## Data: sim_lme_df
##
##      AIC      BIC   logLik deviance df.resid
##  28226.4  28255.2 -14109.2  28218.4     9996
##
## Random effects:
##
## Conditional model:
##  Groups   Name        Variance Std.Dev.
##  hhid     (Intercept) 3.1157   1.765
##  Residual             0.9762   0.988
## Number of obs: 10000, groups:  hhid, 10
##
## Dispersion estimate for gaussian family (sigma^2): 0.976
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.560193   0.558274    2.79   0.0052 **
## x           0.978748   0.009777  100.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In 3, we investigate the contribution of random effects to bias in predictions in the case of a simple linear mixed model with only one predictor and random intercept effect. Similar to the case of simple linear model, the expected mean is very close to the observed in all the three approaches.

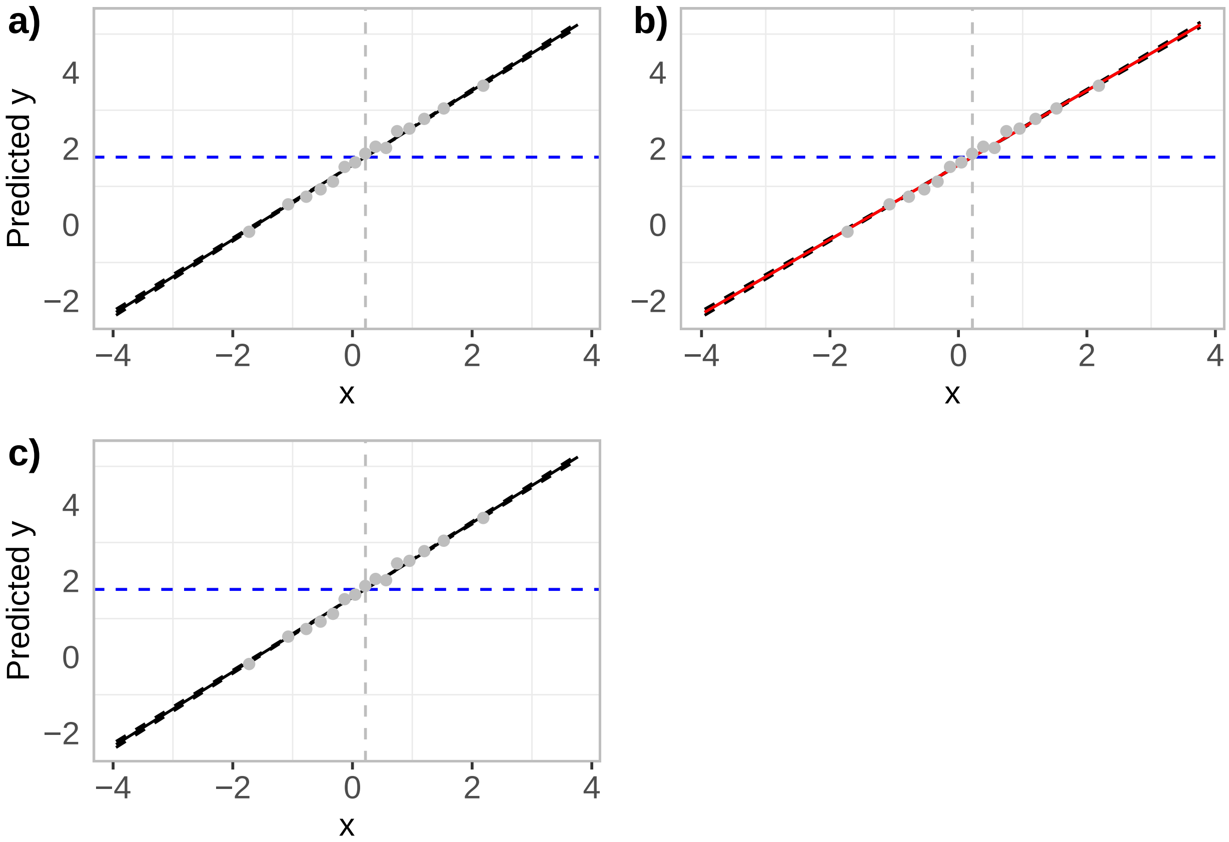

Figure 3: A comparison of variable predictions for simple linear mixed effect model. The dotted blue and grey horizontal lines are the expected and observed/true means, respectively. The grey dots are the binned observations. The vertical grey line is the mean of the focal predictor (model center). In a) non-focal predictors averaged across the levels of focal predictor; b) non-focal linear predictor is binned across the levels of the focal predictor; and c) whole population non-focal linear predictor is used for each level of the focal predictor. In c, the noisy predictions, i.e., black lines, are smoothened resulting to the red trend-line with the corresponding CI represented by the grey shading.

References

Duursma, RA, and AP Robinson. 2003. "Bias in the Mean Tree Model as a Consequence of Jensen's Inequality." *Forest Ecology and Management* 186 (1-3): 373–80.

Fox, John, and Jangman Hong. 2009. "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the Effects Package." *Journal of Statistical Software* 32 (1): 1–24.

Leeper, Thomas J, Jeffrey Arnold, Vincent Arel-Bundock, and Maintainer Thomas J Leeper. 2017. "Package 'Margins'." *Accessed December* 5: 2019.

Lenth, Russell, and Maintainer Russell Lenth. 2018. "Package 'Lsmeans'." *The American Statistician* 34 (4): 216–21.