Coefficients of Determination for Generalized Linear Mixed Models

Dabao Zhang
Department of Statistics, Purdue University

July 8, 2020

Abstract

For linear regression models, the coefficient of determination, a.k.a. R^2, is well-defined to measure the proportion of variation in the dependent variable explained by the predictors included in the model, following the law of total variance. While it is straightforward to extend such a measure for linear mixed models, the natural heteroscedasticity of generalized linear models challenges its extension. Following a variance-function-based measure recently proposed for generalized linear models, we define proper coefficients of determination for generalized linear mixed models, measuring the proportion of variation in the dependent variable modeled by either fixed effects or random effects or both. As the original measure defined for generalized linear models, the definition of our measures only need know the mean and variance functions, so applicable to more general quasi-models. It is consistent with the classical measure of uncertainty using variance, and reduces to the classical definition of the coefficient of determination when linear mixed regression models are considered.

Keywords: Exponential family distribution; Quasi-model; R^2; Variance function

1 Introduction

For a pair of random variables X and Y, the law of total variance states that $var(Y) = var(E[Y|X]) + E[var(Y|X)]$, decomposing the total variance of Y into two parts: the first part $var(E[Y|X])$ for the variation in Y explained by X, and the second part $E[var(Y|X)]$ for the variation in Y unexplained by X. With the ratio

$$\frac{var(E[Y|X])}{var(Y)} = 1 - \frac{E[var(Y|X)]}{var(Y)} \tag{1}$$

measuring the proportion of variation in Y explained by X, the law of total variance provides the theoretical basis for defining the coefficient of determination for linear models.

When a linear mixed model (McCulloch *et al.*, 2008) is considered for observed response variable Y_{ij} from the j-th individual inside the i-th cluster, we usually model it with both fixed and random effects, for $j = 1, \cdots, n_i$ within each $i = 1, \cdots, m$. For simplicity, we write the corresponding linear mixed model as,

$$Y_{ij} = \eta^F_{ij} + \eta^R_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{2}$$

where η^F_{ij} and η^R_{ij} respectively summarize all fixed and random effects on the response variable with $\eta^R_{ij}|\tau^2_{ij} \sim N(0, \tau^2_{ij})$. Following (1), the proportion of variation in Y_{ij} modeled by the fixed effects can be defined as

$$\rho^2_F = 1 - \frac{E[var(Y_{ij}|\eta^F_{ij}, \tau^2_{ij})]}{var(Y_{ij})} = 1 - \frac{E[(Y_{ij} - E[Y_{ij}|\eta^F_{ij}, \tau^2_{ij}])^2]}{E[(Y_{ij} - E[Y_{ij}])^2]}. \tag{3}$$

With η^F_{ij} estimated by $\hat{\eta}^F_{ij}$ and $E[Y_{ij}]$ estimated by $\bar{Y}_{\cdot\cdot}$, we have the following estimate of ρ^2_F,

$$R^2_F = 1 - \frac{\sum_{i,j}(Y_{ij} - \hat{\eta}^F_{ij})^2}{\sum_{i,j}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2}. \tag{4}$$

Since $E[(Y_{ij} - E[Y_{ij}|\eta^F_{ij}, \tau^2_{ij}])^2] = E[E[(Y_{ij} - \eta^F_{ij})^2|\tau^2_{ij}]] = E[\tau^2_{ij}] + \sigma^2$, we may also take estimated variance components to construct R^2_F to estimate ρ^2_F, see, e.g., Xu (2003); Nakagawa and Schielzeth (2013); Nakagawa *et al.* (2017); Jaeger *et al.* (2017). As our review of defining R^2 in linear mixed models is to shed light on their extensions to generalized linear mixed models (McCullagh and Nelder, 1989), we will not pursue this avenue as it cannot manage the heterogeneity in generalized linear mixed models.

The proportion of variation in Y_{ij} modeled by both fixed and random effects can be similarly defined as

$$\rho_M^2 = 1 - \frac{E[var(Y_{ij}|\eta_{ij}^F, \eta_{ij}^R)]}{var(Y_{ij})} = 1 - \frac{E[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])^2]}{E[(Y_{ij} - E[Y_{ij}])^2]}. \tag{5}$$

With $var(E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R]) = \sigma^2$, it is tempting to estimate ρ_M^2 on the basis of $\rho_M^2 = 1 - \sigma^2/var(Y_{ij})$ as in Xu (2003); Nakagawa and Schielzeth (2013); Nakagawa *et al.* (2017); Jaeger *et al.* (2017). However, the total variation in the response variable is described by $SST = \sum_{i,j}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$, and we thus would rather to calculate the total unexplained variation by emphasizing individual heterogeneity of τ_{ij} and the contribution of individual observation, which will help us extend to generalized linear models.

Note that,

$$E\left[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])^2\right] = E\left[E[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])^2|Y_{ij}, \eta_{ij}^F, \tau_{ij}^2, \sigma^2]\right],$$

which implies each observation contribute $E\left[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])^2|Y_{ij}, \eta_{ij}^F, \tau_{ij}^2, \sigma^2\right]$ with observed value Y_{ij} and estimable parameters in η_{ij}^F, τ_{ij}^2, and σ^2. That is, the expectation is on the random variable η_{ij}^R conditional on the observed values and these estimable parameters. With the conditional distribution

$$\eta_{ij}^R|Y_{ij}, \eta_{ij}^F, \tau_{ij}^2, \sigma^2 \sim N\left(\frac{\tau_{ij}^2}{\sigma^2 + \tau_{ij}^2}(Y_{ij} - \eta_{ij}^F), \frac{\sigma^2\tau_{ij}^2}{\sigma^2 + \tau_{ij}^2}\right),$$

we have

$$E[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])^2|Y_{ij}, \eta_{ij}^F, \tau_{ij}^2, \sigma^2] = \left(\frac{\sigma^2}{\sigma^2 + \tau_{ij}^2}\right)^2 \left(Y_{ij} - \eta_{ij}^F\right)^2 + \frac{\sigma^2\tau_{ij}^2}{\sigma^2 + \tau_{ij}^2}.$$

Therefore, ρ_M^2 will be estimated by

$$R_M^2 = 1 - \frac{\sum_{i,j} \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}_{ij}^2}\left[\hat{\tau}_{ij}^2 + \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}_{ij}^2}\left(Y_{ij} - \hat{\eta}_{ij}^F\right)^2\right]}{\sum_{i,j}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2}. \tag{6}$$

The proportion of variation in Y_{ij} modeled by random effects can be simply defined as

$$\rho_R^2 = \rho_M^2 - \rho_F^2 = \frac{E[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \tau_{ij}^2])^2] - E[(Y_{ij} - E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])^2]}{var(Y_{ij})}, \tag{7}$$

and can be estimated as

$$R_R^2 = R_M^2 - R_F^2. \tag{8}$$

While the law of total variance provides a clear path to extend R^2 for linear mixed models, the inherent heteroscedasticity makes it difficult to properly define R^2 for generalized linear mixed models (GLMMs). Indeed, such heteroscedasticity also challenges the proper definition of R^2 for generalized linear models (GLMs; McCullagh and Nelder, 1989). Therefore, many different measures have been proposed to define R^2 for GLMs from different aspects of view (Cameron and Windmeijer, 1997; Cox and Snell, 1989; Maddala, 1983; Magee, 1990; Nagelkerke, 1991; Zhang, 2017). However, it is difficult to extend these measures to account for random effects included in GLMMs.

A common strategy to define R^2 for GLMMs, adopted by Nakagawa and Schielzeth (2013); Nakagawa *et al.* (2017), is to recognize the linear function presented by the link function $g(\cdot)$, i.e.,

$$g(E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R]) = \eta_{ij}^F + \eta_{ij}^R, \tag{9}$$

and instead construct R^2 for the linear mixed regression model,

$$g(Y_{ij}) = \eta_{ij}^F + \eta_{ij}^R + \epsilon_{ij},$$

where $\epsilon_{ij} = g(Y_{ij}) - g(E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])$. However, such measures rely on the specified link function, and even the approximation method which is used to calculate the error variance $var(\epsilon_{ij})$ (Nakagawa *et al.*, 2017). On the other hand, the link function does not necessarily provide homoscedastic variance on the error term ϵ_{ij}, which is the primary challenge in extending classical R^2 from linear models to generalized linear models, although it describes the linear relationship of all effects on $g(E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])$, and presents additive variance components in $g(Y_{ij})$.

A critical concern of defining R^2 based on the approximate linear model for $g(Y_{ij})$ is that proportions of different variance components in $g(Y_{ij})$ may not represent the genuine proportions of different variance components in Y_{ij}. For example, it is well-known that the latent linear model of a probit model may present a much higher R^2, but the binomial response still hold a lot of uncertainty, which is well recognized in the study of genetic heritability, see Dempster and Lerner (1950).

Recently Zhang (2017) showed that quantifying the variation change along the variance function can measure explained variation of a heteroscedastic response variable and proposed to define a variable-function-based R^2. Unlike other likelihood-based measures, such

a measure neither overstate the proportion of explained variation, nor demand the specification of likelihood functions. While it only requires specification of the link function and variance function, it reduces to classical R^2 for general linear models so it is conceptually consistent with classical R^2.

Here we will follow the above extension of R^2 from linear models to linear mixed models, and propose R^2 for GLMMs by quantifying the variation change along the variance function. Although Giorgi (2018) proposed to use the Monte Carlo Markov Chain (MCMC) algorithm to implement such a R^2 to estimate the proportion of the explained total variation by a GLMM, inherent nonlinearity and heterogeneity in a GLMM still challenges the extension of R^2 to understand the proportions of the total variation explained by either or both of fixed- and random-effects of the model. In the next section, we introduce our definition of coefficient of determination, assuming only mean and variance functions are well-specified. We also propose an adjustment to account for the number of predictors in the model. Advantages and disadvantages of different definitions are investigated via simulation studies in Section 3, which also show the robustness of our proposed definition. We also apply and compare different coefficients of determination to a set of real data in Section 4, and conclude this paper with Section 5.

2 R^2 for Generalized Linear Mixed Models

For the generalized linear mixed model (9), the variance of Y_{ij}, given both fixed and random effects, can be specified via a dispersion parameter ϕ and a known variance function $V(\cdot)$, i.e.,

$$var(Y_{ij}|\eta_{ij}^F, \eta_{ij}^R) = \phi V(g^{-1}(\eta_{ij}^F + \eta_{ij}^R)).$$

In general, as long as the mean $g^{-1}(\eta_{ij}^F + \eta_{ij}^R)$ can be modeled well and linked appropriately to a set of predictors, a generalized linear model with known variance function $V(\cdot)$ can be investigated for the utility of the involved predictors.

The variance function describes the effect of the mean on the variation of the response variable besides the dispersion parameter. For a response variable with its mean moving from a to b, its variation changes accordingly along the variance function from $\phi V(a)$ to

$\phi V(b)$. Zhang (2017) therefore claimed that the variation change of the response variable should be measured using, instead of $(a-b)^2$, the squared length of the variance function $V(\cdot)$ between $V(a)$ to $V(b)$, that is,

$$d_V(a,b) = \left\{ \int_a^b \sqrt{1+[V'(t)]^2} dt \right\}^2 .$$

Our definition of R^2 for GLMMs will proceed by replacing the Euclidean distance by the above manifold distance along the variance function.

Replacing $(a-b)^2$ with $d_V(a,b)$ in (3), we can define the proportion of variation in Y_{ij} modeled by the fixed effects as

$$\rho_F^2 = 1 - \frac{E[d_V(Y_{ij}, E[Y_{ij}|\eta_{ij}^F, \tau_{ij}^2])]}{E[d_V(Y_{ij}, E[Y_{ij}])]}. \tag{10}$$

Note that

$$E[Y_{ij}|\eta_{ij}^F, \tau_{ij}^2] = E\left[E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R]|\eta_{ij}^F, \tau_{ij}^2\right] = E[g^{-1}(\eta_{ij}^F + \eta_{ij}^R)|\eta_{ij}^F, \tau_{ij}^2],$$

which follows the model (9). We can rewrite

$$\rho_F^2 = 1 - \frac{E\left[d_V(Y_{ij}, E[g^{-1}(\eta_{ij}^F + \eta_{ij}^R)|\eta_{ij}^F, \tau_{ij}^2])\right]}{E[d_V(Y_{ij}, E[Y_{ij}])]}. \tag{11}$$

With η_{ij}^F estimated by $\hat{\eta}_{ij}^F$ and τ_{ij}^2 estimated by $\hat{\tau}_{ij}^2$, ρ_F^2 can be estimated by

$$R_F^2 = 1 - \frac{\sum_{i,j} d_V(Y_{ij}, E[g^{-1}(\hat{\eta}_{ij}^F + \eta_{ij}^R)|\hat{\eta}_{ij}^F, \hat{\tau}_{ij}^2])}{\sum_{i,j} d_V(Y_{ij}, \bar{Y}_{\cdot\cdot})}. \tag{12}$$

The involved expectation is calculated over the random effect η_{ij}^R conditional on estimates of η_{ij}^F and τ_{ij}^2. As the random effects are usually assumed to be normally distributed, such a expectation can be easily evaluated by numerical methods available for one-dimensional integration, e.g., Piessens *et al.* (1983).

With (5), we can similarly define the proportion of variation in Y_{ij} modeled by both fixed and random effects as

$$\rho_M^2 = 1 - \frac{E[d_V(Y_{ij}, E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])]}{E[d_V(Y_{ij}, E[Y_{ij}])]} = 1 - \frac{E[d_V(Y_{ij}, g^{-1}(\eta_{ij}^F + \eta_{ij}^R))]}{E[d_V(Y_{ij}, E[Y_{ij}])]}. \tag{13}$$

It is tempting to estimate η_{ij}^R, as well as η_{ij}^F, to estimate ρ_M^2. However, estimation of the parameter τ_{ij}^2 should be preferred to that of random effects as the latter is relatively

unstable. We here estimate both η_{ij}^F and τ_{ij}^2 for each observation, and together with Y_{ij} from each observation, we have

$$E[d_V(Y_{ij}, g^{-1}(\eta_{ij}^F + \eta_{ij}^R))] = E\left[E[d_V(Y_{ij}, g^{-1}(\eta_{ij}^F + \eta_{ij}^R))|Y_{ij}, \eta_{ij}^F, \tau_{ij}^2]\right].$$

Therefore, with estimates $\hat{\eta}_{ij}^F$ and $\hat{\tau}_{ij}^2$, we can estimate ρ_M^2 by

$$R_M^2 = 1 - \frac{\sum_{i,j} E[d_V(Y_{ij}, g^{-1}(\eta_{ij}^F + \eta_{ij}^R))|Y_{ij}, \hat{\eta}_{ij}^F, \hat{\tau}_{ij}^2]}{\sum_{i,j} d_V(Y_{ij}, \bar{Y}_{..})}. \tag{14}$$

The expectation in the above definition of R_M^2 is calculated on the random effect η_{ij}^R conditional on Y_{ij}, and estimates of η_{ij}^F and τ_{ij}^2.

Denote $f(\cdot|g^{-1}(\eta_{ij}^F + \eta_{ij}^R))$ the density function of Y_{ij} with mean value $g^{-1}(\eta_{ij}^F + \eta_{ij}^R)$. Then the expectation involved in (14) can be rewritten as

$$\begin{aligned} & E[d_V(Y_{ij}, g^{-1}(\eta_{ij}^F + \eta_{ij}^R))|Y_{ij}, \eta_{ij}^F, \tau_{ij}^2] \\ = \; & \frac{E_{\eta_{ij}^R|\tau_{ij}^2}[d_V(Y_{ij}, g^{-1}(\eta_{ij}^F + \eta_{ij}^R)) \times f(Y_{ij}|g^{-1}(\eta_{ij}^F + \eta_{ij}^R))]}{E_{\eta_{ij}^R|\tau_{ij}^2}[f(Y_{ij}|g^{-1}(\eta_{ij}^F + \eta_{ij}^R))]}, \end{aligned}$$

where $E_{\eta_{ij}^R|\tau_{ij}^2}[\cdot]$ calculates the mean of the underlying term over the random effect $\eta_{ij}^R \sim N(0, \tau_{ij}^2)$, and can be easily evaluated via numerical integration (Piessens *et al.*, 1983). For quasi-models, the likelihood function $f(\cdot)$ can be replaced by the underlying quasi-likelihood function (McCullagh, 1983).

The proportion of variation in Y_{ij} modeled by random effects can be simply defined as

$$\rho_R^2 = \rho_M^2 - \rho_F^2 = \frac{E[d_V(Y_{ij}, E[Y_{ij}|\eta_{ij}^F, \tau_{ij}^2])] - E[d_V(Y_{ij}, E[Y_{ij}|\eta_{ij}^F, \eta_{ij}^R])]}{var(Y_{ij})}, \tag{15}$$

and can be estimated by R_R^2 in (8) using R_M^2 in (14) and R_F^2 in (12). Note that we do not parallel the definition of ρ_R^2 and R_R^2 to that of ρ_F^2 and R_F^2 for two reasons: firstly, both random and fixed effects may share the same predictors and are correlated, so we would rather evaluate the contribution due to random effects by removing those attributable to the fixed effects; secondly, defining ρ_R^2 using $E[Y_{ij}|\eta_{ij}^R]$ demands more computation, which is also difficult. Note that ρ_R^2 defined in (15) may be more appropriate to be named as $\rho_{R|F}^2$, accordingly R_R^2 as $R_{R|F}^2$ which should serve our routine purpose of evaluating the variation due to random effects.

Remarks. (i) The above coefficients of determination are well-defined as long as the mean and variance functions are specified, like the quasi-models. Therefore, $\hat{\eta}_{ij}^F$, $\hat{\tau}_{ij}^2$, and $\hat{\sigma}^2$ may be derived from quasi-likelihood estimators, other than MLE; (ii) Because $V'(\cdot)$ is constant for normal and Poisson distributions, each R^2 defined here is consistent with those defined for linear mixed models following the law of total variance; (iii) The coefficient of partial determination can also be defined to measure the proportion of variation in the response variable not explained by a set of predictors that can be explained by an additional set of predictors, following Zhang (2017); (iv) each of above R^2 suffers to increasing numbers of predictors as the classical R^2, and may increase even if irrelevant predictors are added to the underlying model. As shown in Zhang (2017), we can accordingly define an adjusted version of each above R^2 by accounting for number of predictors involved in calculate the fixed effect η_{ij}^F.

3 Simulation Studies

The R package `performance` has implemented the R^2 proposed by Xu (2003) for linear mixed models, and the one proposed by Nakagawa *et al.* (2017) for mixed-effects models. We will compare the performance of our proposed R^2 to them, by simulating a total of 100 data sets for each model under investigation.

3.1 Linear Mixed Models

For linear mixed models, we simulate each data set with a total of 200 random samples, evenly clustered inside m groups with $m = 5$ and 50, respectively. A binary covariate X_1 is generated for each observation, with half observations within the same group taking 1 and the other half taking -1. A second variable X_2 is generated from the standard normal distribution, independent of X_1 and the response variable. The j-th response value inside the i-th cluster, i.e., y_{ij}, is generated by

$$y_{ij} = \mu_i + x_{1ij}\beta + \epsilon_{ij},$$

where x_{1ij} is the corresponding value of the binary covariate X_1, the random effect $\mu_i \overset{iid}{\sim} N(0,1)$, and $\epsilon_{ij} \overset{iid}{\sim} N(0,1)$. For each data set, we fit the different models with both

maximum likelihood (ML) and restricted maximum likelihood (REML) methods, and then estimate both ρ^2_M and ρ^2_F with different approaches as shown in Figure 1 and Figure 2, respectively.

The difference between REML and ML methods lies only in the estimated variances of the random effects, i.e., τ^2_{ij}, and such difference increases when m decreases. So it is not surprising to observe much wider difference in the case of $m = 5$ than $m = 50$ between estimated ρ^2_M by Nakagawa *et al.* (2017) based on the REML and ML methods respectively, and the ML based estimation of ρ^2_M is usually larger than the one based on the REML method. But the method by Xu (2003) is rarely affected by the choice of REML or ML method. Such difference between REML and ML methods narrows when estimating ρ^2_F by Nakagawa *et al.* (2017), however, there is no difference when R^2_F is used. It is interesting to observe that our proposed methods and the method by Nakagawa *et al.* (2017) coincide when the ML method is used.

When m is large, the method by Xu (2003) may significantly overestimate ρ^2_M as shown in Figure 1. When m is small, the method by Nakagawa *et al.* (2017) may slightly underestimate ρ^2_F when the REML method is used.

3.2 Logistic Mixed Models

For logistic models, we simulate each data set with a total of 400 random samples, evenly clustered inside m groups with $m = 10$ and 50, respectively. The binary X_1 and continuous X_2 are similarly generated as in the previous section. The j-th response value inside the i-th cluster, i.e., y_{ij}, is generated by

$$E[y_{ij}|\mu_i, x_{1ij}] = \frac{1}{1 + \exp\{-\mu_i - x_{1ij}\beta\}},$$

where x_{1ij} is the corresponding value of the binary covariate X_1, and the random effect $\mu_i \overset{iid}{\sim} N(0, 1)$. For each data set, we fit the different models with the maximum likelihood (ML) method, and then estimate both ρ^2_M and ρ^2_F with our approach and the method by Nakagawa *et al.* (2017) as shown in Figure 3 and Figure 4, respectively.

Overall, either the estimated ρ^2_M or ρ^2_F by our methods and the one by Nakagawa *et al.* (2017) demonstrate similar patterns in each regression model. Specifically, a regression

model including the true predictor X_1, i.e., when regressing Y vs. X_1 or regressing Y vs. (X_1, X_2), both estimated ρ_M^2 or ρ_F^2 increase when β increases, and estimated ρ_F^2 nears zero when $\beta = 0$. However, estimated ρ_M^2 implies the explained variation due to the random effects when $\beta = 0$.

For small β, we observe higher estimated values of both ρ_M^2 or ρ_F^2 by Nakagawa *et al.* (2017) than our approach. However, for large β, the method by Nakagawa *et al.* (2017) tends to underestimate ρ_F^2 in either case of $m = 10$ and $m = 50$, and underestimate ρ_M^2 in the case of $m = 10$. However, the estimated ρ_M^2 by Nakagawa *et al.* (2017) and our approach eventually converge for very large β when $m = 50$ or the true predictor X_1 is not included in the model (see Figures 3.e and 3.f). In summary, R_M^2 and R_F^2 capture well the increasing proportion of explained variation in a binary response variable, and perform better than the method by Nakagawa *et al.* (2017).

When regressing Y vs. X_2, ρ_M^2 measures only the variation in Y explained by the random effects and ρ_F^2 should be zero as X_2 is not related to Y. The close to zero values in Figures 4.e and 4.f imply good measurement by our method and the method by Nakagawa *et al.* (2017), although the method by Nakagawa *et al.* (2017) slightly overestimates in comparison to our method. The estimated ρ_M^2 shows decreasing patterns in both Figures 3.e and 3.f because of constant τ^2 in comparison to the increasing variation in Y caused by increasing β. We also observe that the method by Nakagawa *et al.* (2017) may claim much higher proportion of explained variation due to the random effects, in comparison to our approach.

4 Real Data Analysis

4.1 Analysis of the Sleep Study Data via Linear Mixed Models

Eighteen subjects have been followed on their reaction times for nine days in a sleep deprivation study (Belenky *et al.*, 2003). As shown in Figure 5, individual trajectory demonstrated the linear trend in increasing reaction times, but the heterogeneity between trajectories provides strong evidence in favor of random effects of both intercepts and slopes. Here we will investigate explained variation of the reaction time under different variance structures.

There are two sets of variance structures, i.e., the correlation between the random intercept and slope within the same participant, and the correlation between the error terms within the same participant. For either set, we will model both independence and dependence, and the dependence between error terms will be modeled by a first-order autoregression, i.e., AR(1).

As shown in Table 1, R^2_F stably estimated ρ^2_F, the proportion of variation explained by factors with fixed-effects, across different variance structures whether REML or ML method was used. However, the method by Nakagawa *et al.* (2017) provided the estimates ranging from .2771 to .3018, relying on not only the variance structures but also whether REML or ML method was used. In general, When the ML method was used, the method by Nakagawa *et al.* (2017) provided larger estimates.

Table 1: Analysis of the Sleep Study Data

		Independent Errors				AR(1) Errors			
		Ind. RE		Dep. RE		Ind. RE		Dep. RE	
		REML	ML	REML	ML	REML	ML	REML	ML
ρ^2_F	R^2_F	.2865	.2865	.2865	.2865	.2863	.2863	.2862	.2863
	Nakagawa	.2830	.2927	.2786	.2876	.2928	.3018	.2771	.2861
ρ^2_M	R^2_M	.7998	.7973	.8004	.7981	.7316	.7303	.7194	.7206
	Nakagawa	.7965	.7897	.7992	.7928	.7185	.7137	.7134	.7103

Ind. RE – Independent Random Effects; Dep. RE – Dependent Random Effects; Nakagawa – Nakagawa *et al.* (2017).

As for ρ^2_M, i.e., the proportion of variation explained by the model in total, estimates by both R^2_M and the method by Nakagawa *et al.* (2017) vary across different variance structures, especially between the models with independent errors and the ones with AR(1) errors. R^2_M provides slightly larger estimates than the method by Nakagawa *et al.* (2017) in all models, and it also performs slightly more stable across REML and ML.

4.2 Analysis of the Balance Study Data via Logistic Mixed Models

Steele (1998) conducted an experiment to study the effects of surface and vision on balance. Each of a total of twenty males and twenty females was tested twice in each combination of two different surfaces and three vision conditions. We consider logistic models to investigate how the factors, such as sex, height, surface, and vision, affect a subject's balance. We will consider different models to include subject as a factor with random, fixed, and no effects, respectively, see Table 2. For GLMs, R^2_F reduces to R^2 proposed by Zhang (2017), and we also calculated R^2 prposed by Cameron and Windmeijer (1997) and Nagelkerke (1991) respectively.

Table 2: Analysis of the Balance Study Data

effects of	ρ^2_M		ρ^2_F		others	
subject	R^2_M	Nakagawa	R^2_F	Nakagawa	Cameron	Nagelkerke
random	.7278	.9128	.4527	.6693	–	–
fixed	–	–	.7832	–	.7694	.8557
none	–	–	.4542	–	.4270	.5614

Cameron – Cameron and Windmeijer (1997); Nagelkerke – Nagelkerke (1991).

For the model with random effects of subject, the method by Nakagawa *et al.* (2017) provided much larger estimates, i.e., .9128 and .6693 respectively, of both ρ^2_M and ρ^2_F than our approach. On the other hand, by including the fixed effects of subject, the proportion of variation explained by all factors is only .7832 as measured by Zhang (2017), and .7694 as measured by Cameron and Windmeijer (1997), which are close to .7278 estimated by R^2_M in the mixed model. Furthermore, the model excluding subject provides a measure of the portion of variation explained on the factors with fixed effects, and has R^2 at .4542 by Zhang (2017), which is very close to .4527 estimated by R^2_F in the mixed model. As noted by Zhang (2017), the method by Nagelkerke (1991) tends to overestimate R^2 in GLMs, which is evidenced by the value of .8557 for the model including subject with fixed effects, and the value of .5614 for the model excluding subject. Nonetheless, the estimated ρ^2_M by

Nakagawa *et al.* (2017) is even higher than the R^2 estimated by Nagelkerke (1991) for the model including subject with fixed effects, and the estimated ρ_F^2 by Nakagawa *et al.* (2017) is much higher than the R^2 estimated by Nagelkerke (1991) for the model excluding subject. We conclude the more appropriateness of our proposed R_M^2 to account for the proportion of variation explained by the model in total, and R_F^2 to account for the proportion of variation explained by the fixed effects.

5 Conclusion

Unlike p-values which signal the variable significance but rely on the sample size, the coefficient of determination, a.k.a. R^2, measures the proportion of the variation in the response variable explained by a set of predictors. R^2 is a key statistic, and plays an important role in molecular biology to measure the heritability of different traits (Visscher *et al.*, 2008). The popularly used mixed-effects models in genomic studies demand appropriate extension of R^2 which is well defined in linear regression models. We have followed the law of total variance to extend R^2 to LMM, i.e., R_M^2, to properly account for the proportion of the variation explained by all predictors in the model, including those with fixed effects and random effects. We also define R_F^2 in LMM to account for the proportion of the variation explained by predictors only with fixed effects, and R_R^2 in LMM to account for the proportion of the variation explained by predictors only with random effects. These extensions are undertaken in a way which allows to similarly define R_M^2, R_F^2, and R_R^2 for GLMMs by following Zhang (2017).

Nakagawa *et al.* (2017) proposed R^2 for GLMMs by assuming an underlying model with additive variance components due to fixed-effects predictors, random-effects predictors, and the model itself. Such an additive model is defined by approximating an transformed response variable with the transformation attached to the link function. However, besides the approximation error, the link function does not necessarily provide homoscedastic variance on the assumed error term, which is the primary challenge in extending R^2 to GLM. Indeed, well demonstrated in the study of genetic heritability, R^2 defined for transformed response variable may not represent well the genuine proportions of different variance components in the original response variable (Dempster and Lerner, 1950). Our simulation study and real

data analysis both evidence such concern. Instead, our properly defined R^2_M and R^2_F for GLMMs match well with those defined in GLMs. Furthermore, they are also well defined for any, even quasi, GLMMs, as long as they have well defined mean and variance functions.

Acknowledgement

This work is partially supported by NCI R03CA235363.

References

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., and Balkin, T. J. (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, **12**, 112.

Cameron, A. C. and Windmeijer, A. G. (1997) An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, **77**, 329-342.

Cox, D. R. and Snell, E. J. (1989) *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

Dempster, E. R. and Lerner, I. M. (1950) Heritability of threshold characters. *Genetics*, **35**, 212-236.

Giorgi, E. (2018) On the goodness-of-fit of generalized linear geostatistical models. *Spatial Statistics*, **28**, 79-83.

Jaeger, B. C., Edwards, L. J., Das, K., and Sen, P. K. (2017) An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, **44**, 1086-1105.

Maddala, G. S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University.

Magee, L. (1990) R^2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, **44**, 250-253.

McCullagh, P. (1983) Quasi-likelihood functions. *The Annals of Statistics*, **11**, 59-67.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. Chapman and Hall/CRC.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008) *Generalized, Linear, and Mixed Models*, 2nd Edition. New York: Wiley.

Nagelkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691-692.

Nakagawa, S., Johnson, P. C. D., and Schielzeth, H. (2017) The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, **14**, 20170213.

Nakagawa, S. and Schielzeth, H. (2013) A general and simple method for obtaining R^2 from generalized linear mixed models. *Methods in Ecology and Evolution*, **4**, 133-142.

Piessens, R., DonckerKapenga, E. de, Uberhuber, C. W., and Kahaner, D. K. (1983) Quadpack: a Subroutine Package for Automatic Integration. Springer Verlag.

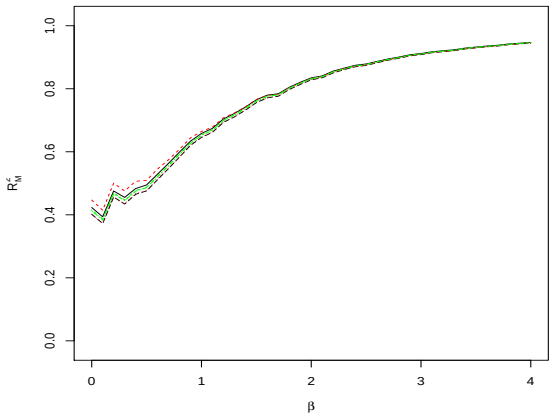
Steele, R. (1998) *Effect of Surface and Vision on Balance.* Ph. D. thesis, Department of Physiotherapy, University of Queensland.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008) Heritability in the genomics eraconcepts and misconceptions. *Nature Reviews Genetics*, **9**, 255-266.

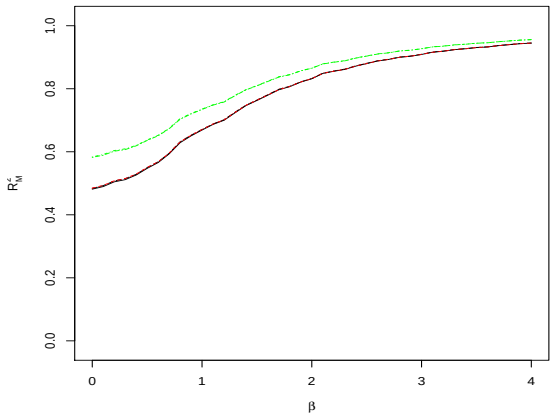
Xu, X. (2003) Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, **22**, 3527-3541.

Zhang, D. (2017) A coefficient of determination for generalized linear models. *The American Statistician*, **71**, 310-316.

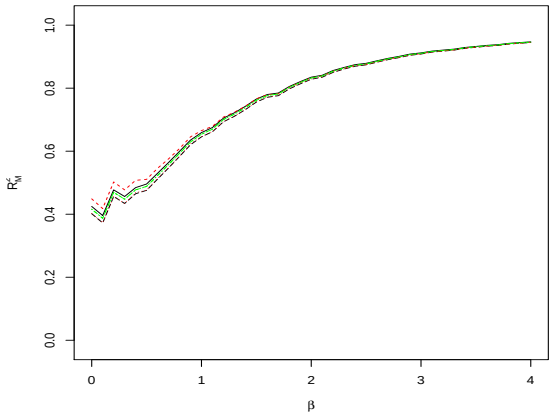
a. Regressing Y *vs.* X_1 when $m = 5$



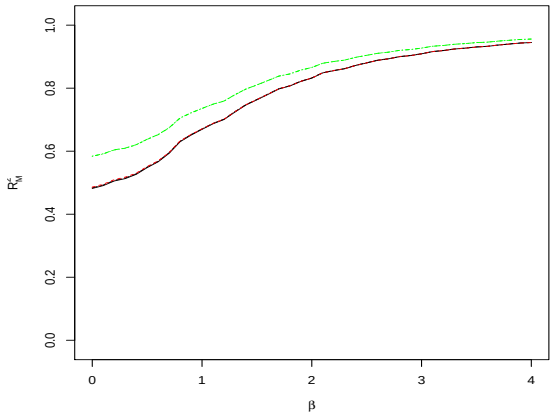
b. Regressing Y *vs.* X_1 when $m = 50$



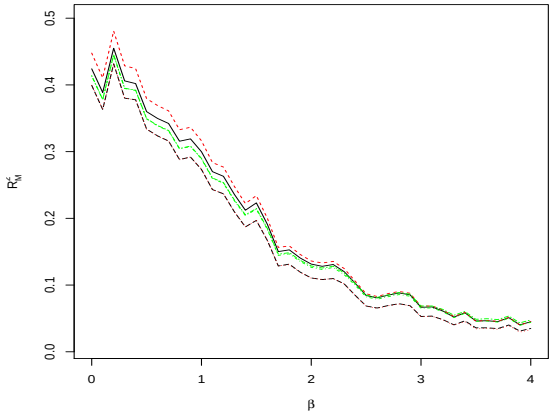
c. Regressing Y *vs.* (X_1, X_2) when $m = 5$



d. Regressing Y *vs.* (X_1, X_2) when $m = 50$



e. Regressing Y *vs.* X_2 when $m = 5$



f. Regressing Y *vs.* X_2 when $m = 50$

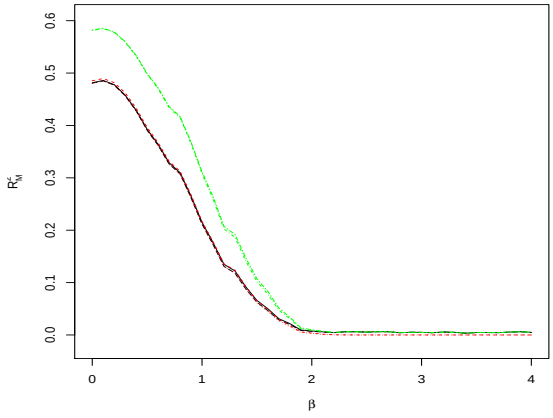
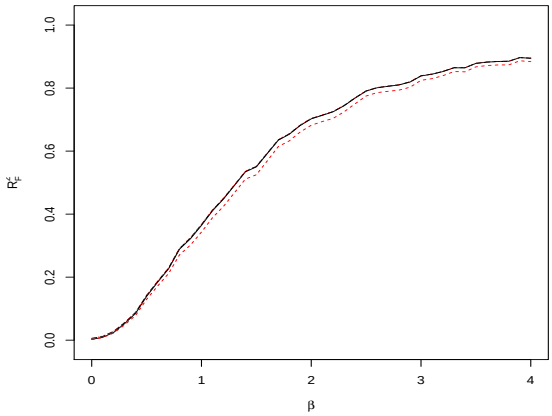
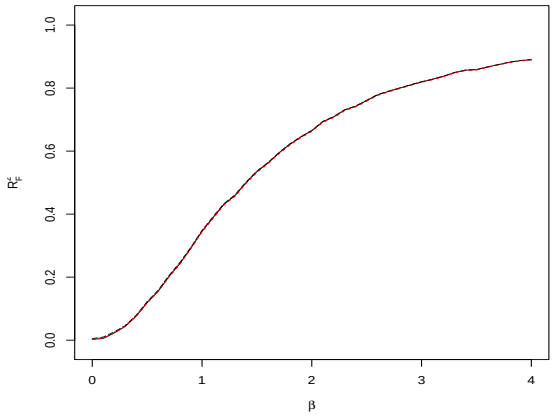


Figure 1: Estimation of ρ^2_M in linear mixed models. Shown in the plot are R^2_M based on REML (black solid) and ML (black long-dashed), the method by Nakagawa *et al.* (2017) based on REML (red dashed) and ML (red dotted), and the method by Xu (2003) based on REML (green two-dashed) and ML (green dotted-dashed).

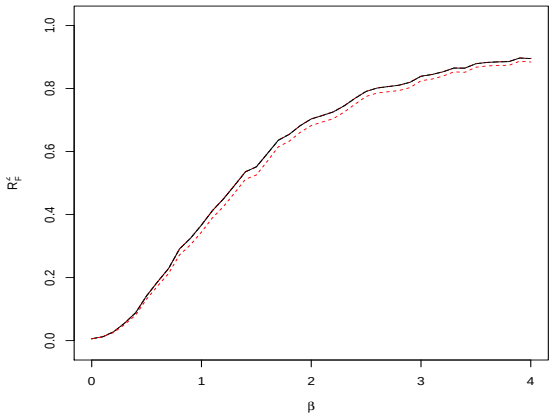
a. Regressing Y *vs.* X_1 when $m = 5$



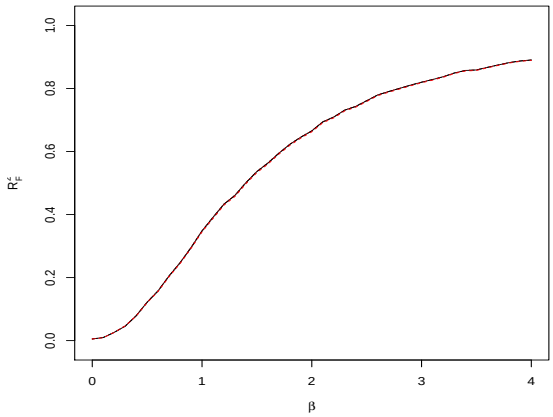
b. Regressing Y *vs.* X_1 when $m = 50$



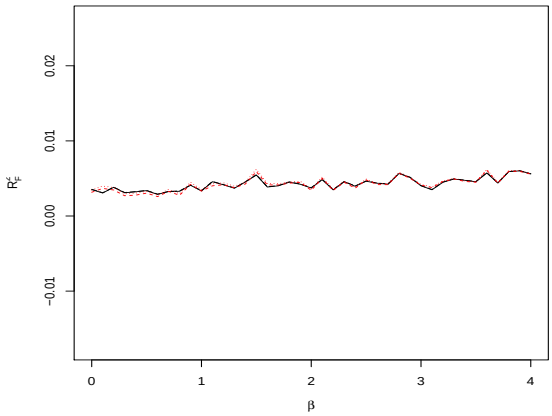
c. Regressing Y *vs.* (X_1, X_2) when $m = 5$



d. Regressing Y *vs.* (X_1, X_2) when $m = 50$



e. Regressing Y *vs.* X_2 when $m = 5$



f. Regressing Y *vs.* X_2 when $m = 50$

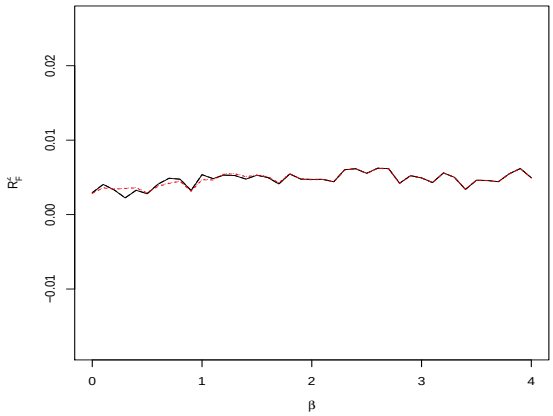
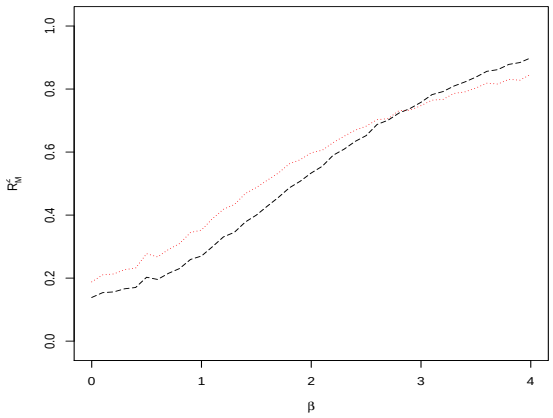
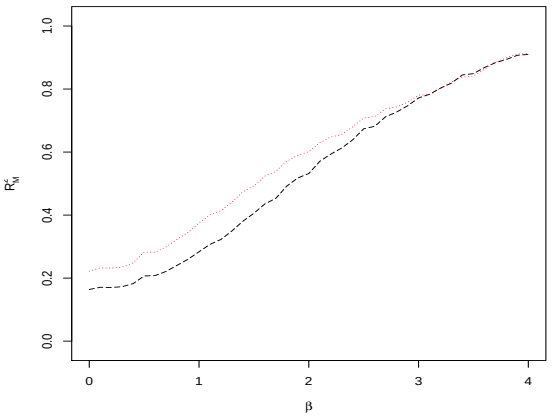


Figure 2: Estimation of ρ^2_F in linear mixed models. Shown in the plot are R^2_F based on REML (black solid) and ML (black long-dashed), and the method by Nakagawa *et al.* (2017) based on REML (red dashed) and ML (red dotted).

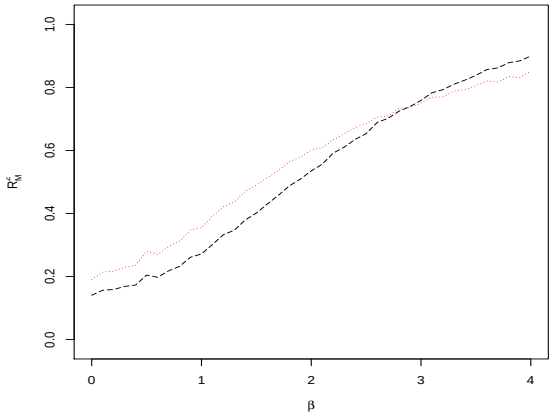
a. Regressing Y *vs.* X_1 when $m = 10$



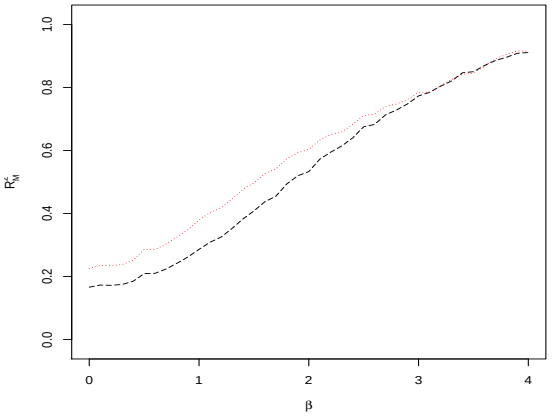
b. Regressing Y *vs.* X_1 when $m = 50$



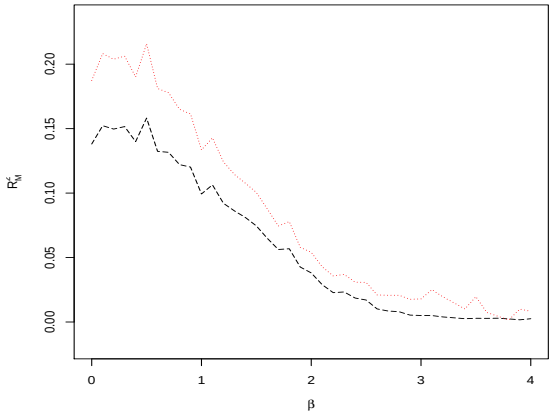
c. Regressing Y *vs.* (X_1, X_2) when $m = 10$



d. Regressing Y *vs.* (X_1, X_2) when $m = 50$



e. Regressing Y *vs.* X_2 when $m = 10$



f. Regressing Y *vs.* X_2 when $m = 50$

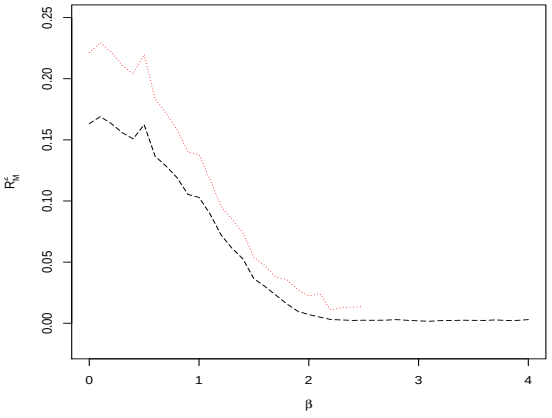
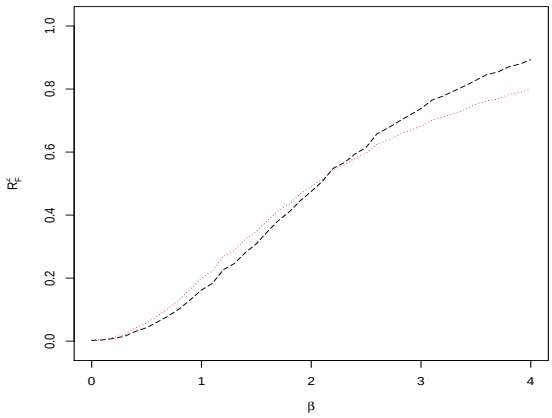
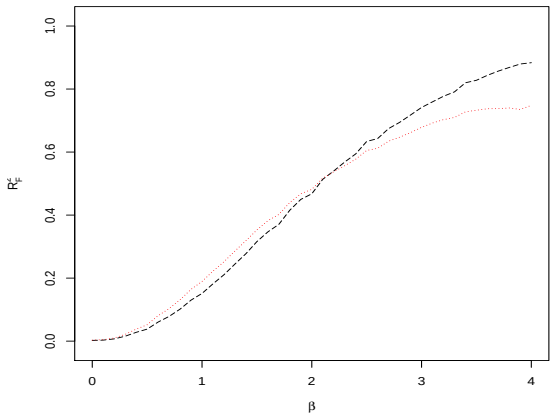


Figure 3: Estimation of ρ^2_M in logistic mixed models. Shown in the plot are R^2_M (black long-dashed), and the method by Nakagawa *et al.* (2017) (red dotted).

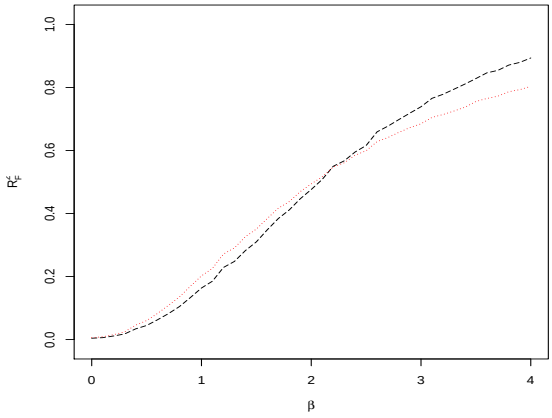
a. Regressing Y *vs.* X_1 when $m = 10$



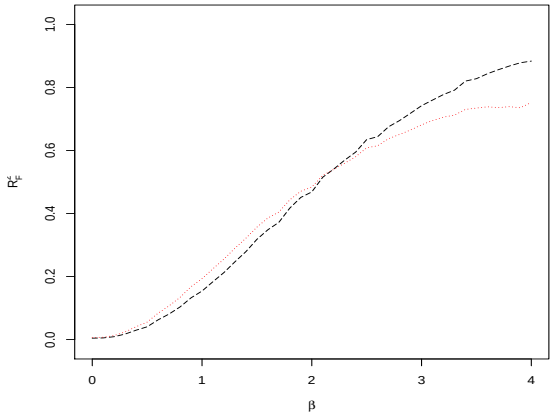
b. Regressing Y *vs.* X_1 when $m = 50$



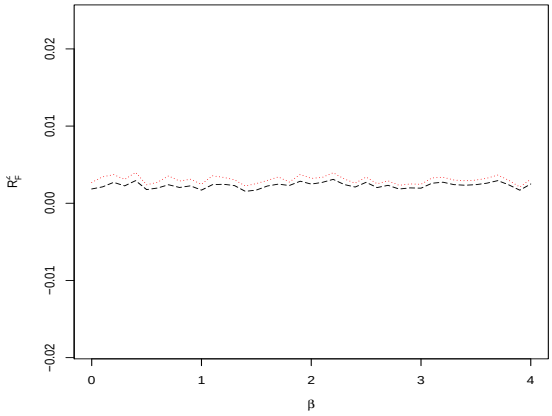
c. Regressing Y *vs.* (X_1, X_2) when $m = 10$



d. Regressing Y *vs.* (X_1, X_2) when $m = 50$



e. Regressing Y *vs.* X_2 when $m = 10$



f. Regressing Y *vs.* X_2 when $m = 50$

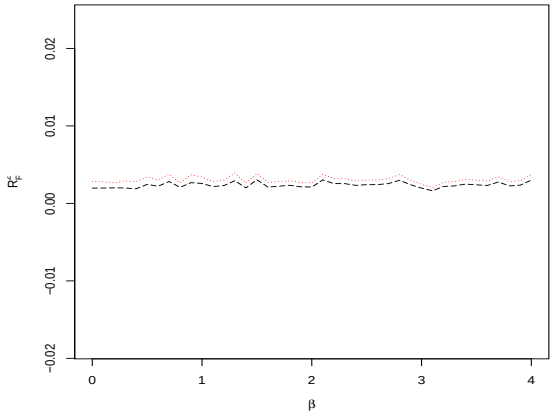


Figure 4: Estimation of ρ^2_F in logistic mixed models. Shown in the plot are R^2_F (black long-dashed), and the method by Nakagawa *et al.* (2017) (red dotted).

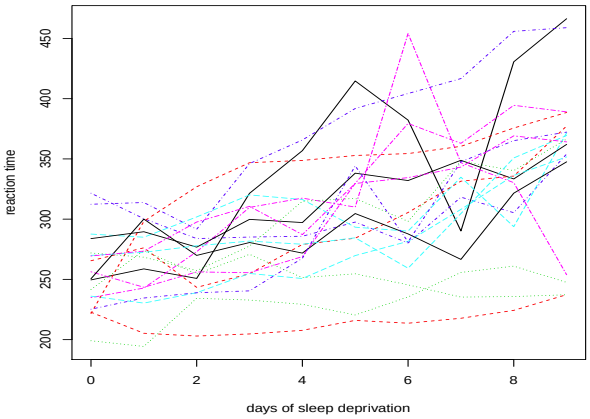


Figure 5: Reaction time (in milliseconds) trajectories of eighteen participants involved in the sleep deprivation study.