Safe models for risky decisions

Steingröver, H.M.

Appendix H

Appendix to Chapter 10: "Using Bayesian Regression to Incorporate Covariates into Hierarchical Cognitive Models"

H.1 Complete Results of the Simulation Study

In this appendix we provide the results of our simulation study for all four parameters of the PVL-Delta model. Figure H.1 gives the log Bayes factors for all PVL-Delta parameters from our simulations with uncorrelated covariates. Dark grey dots show the Bayes factors obtained in the regression analysis, light grey dots show the Bayes factors obtained in the median-split analysis. Recall that our simulated data were generated so that the first covariate would be positively correlated with the A parameter and the second covariate would be negatively correlated with the w parameter. The correlations between A and the second covariate, and between w and the first covariate were set to 0, and the relationships between the remaining model parameters and the covariates were set to the values estimated from Steingroever et al.'s (submitted) data, and were negligible. As described in the main text, the Bayes factors from the regression analysis showed strong evidence for an effect of the first covariate on the A parameter (dark grey dots, left column in the top row), whereas the median-split analysis provided much weaker evidence for such an effect (light grey dots, left column in the top row). Similarly, the regression analysis strongly supported an effect of the second covariate on the w parameter (dark grey dots, second column in the bottom row), whereas the median-split analysis provided weaker evidence for such an effect (light grey dots, second column in the bottom row). For the null-effects of the first covariate on the w parameter (second column, top row) and of the second covariate on the A parameter (left column, bottom panel), both analyses performed similarly without any appreciable differences in Bayes factors. Finally, both analyses provided only weak support if any for an effect the covariates on the a and c parameters and there were no sizable differences in Bayes factors.

Figure H.2 gives the log Bayes factors for all PVL-Delta parameters from our simulations with correlated covariates. As described in the main text, the Bayes factors obtained from the regression analysis again showed stronger evidence for an effect of the first covariate on the A parameter (dark

$\rho(X_1, X_2) = 0$

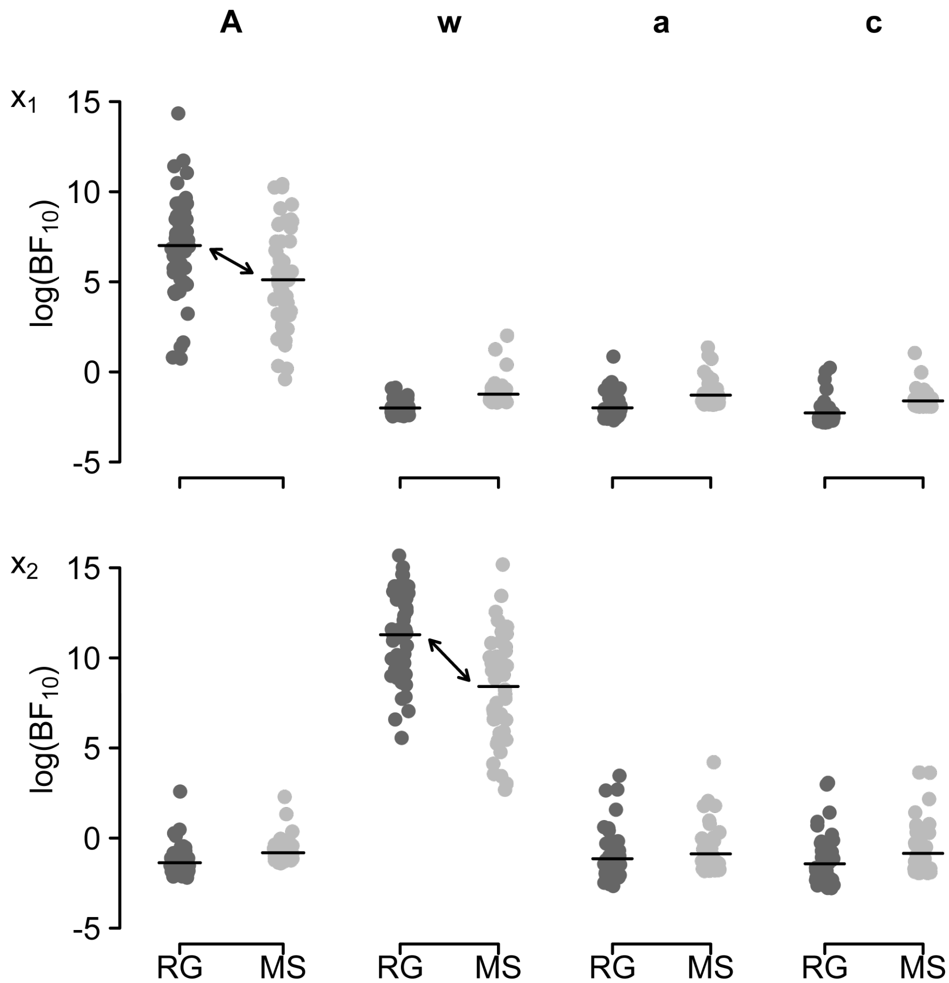

Figure H.1: Bayes factors from 50 simulated data sets for the regression and median-split analysis with uncorrelated covariates. Data points show the log Bayes factors for the alternative hypothesis ($\log(\mathrm{BF}_{10})$) obtained in the regression (RG, dark grey dots) and median-split (MS, light grey dots) analysis for the PVL-Delta model's A and w parameters (columns) and two covariates (rows). Lines indicate the mean log BF. Data points are jittered along the x-axis for improved visibility.

grey dots, left column in the top row) than the median-split analysis (light grey dots, left column in the top row). Similarly, the regression analysis provided stronger support for an effect of the second covariate on the w parameter (dark grey dots, second column in the bottom row) than the median-split analysis (light grey dots, second column in the bottom row). However, unlike in the case of uncorrelated covariates, in the case of correlated covariates the median-split analysis now suggested spurious effects of the first covariate on the w parameter (second column, top row) and of the second covariate on the A parameter (left column, bottom row). Finally, the regression as well as the median-split analysis did not provide strong evidence for any effects of the covariates on the a and c parameters, and there were no clear differences in Bayes factors visible between the two analyses.

Taken together, these results illustrate that, in the case of uncorrelated covariates, a median-split analysis tends to understate the evidence for existent effects. In the case of correlated covariates, a median-split analysis also understates the evidence for existent effects but in addition suggests spurious effects of covariates on model parameters that are in fact unrelated. Furthermore, our results show that in cases where model parameters do not have any appreciable relationships with any of the covariates, as was the case for the a and c parameters, regression and median-split analyses perform similarly and there are no appreciable biases associated with a dichotomization-based analysis.

Figure H.3 shows the posterior means of the standardized effect sizes estimated in the regression analysis (RG, left panels in each subplot) and the posterior means of the standardized effect sizes estimated in the median-split analysis (MS, right panels in each subplot). The left subplot shows results for the case of uncorrelated covariates, the right subplot shows the results for the case of correlated covariates. The top row shows the results for the first covariate, the bottom row for the second covariate. The results are complementary to the results for the Bayes factors. In the case of uncorrelated covariates (left subplot), the estimated effects in both models are largest for effects that we created to be non-zero (i.e., the effect of the first covariate on A and the effect of the second covariate on w, leftmost column of the panels in the top row and second-from-left column of the panels in the bottom row, respectively). Moreover, both models correctly estimate the direction of the effect of the first covariate on the A parameter to be positive (leftmost column of the panels in the top row), and the direction of the effect of the second covariate on w to be negative (second-from-left column of the panels in the bottom row). Both models also correctly estimate the effects of the first and second covariate on a and c to be close to 0 (second-from-right and rightmost columns of each panel).

In the case of correlated covariates (right subplot), both analyses again correctly estimate the size and direction of the strong effects of the first covariate on the A parameter (leftmost column of the panels in the top row) and of the second covariate on the w parameter (second-from-left column of the panels in the bottom row). However, whilst the regression analysis correctly estimates the relationships between the first covariate and the w parameter (left panel, second-from-left column in the top row) and between second covariate and the A parameter (left panel, leftmost column in the bottom row) to be approximately 0, the median-split analysis suggests a weakly negative association between the first covariate and w (right panel, second-from-left column in the top row) and between the second covariate and A (right panel, leftmost column in the bottom row). Finally, both models correctly estimate the effects of the covariates on the a and c parameters to be close to 0.

These results align well with the results for the Bayes factors. In the case of uncorrelated covariates, the regression analysis as well as the median-split analysis correctly indicate the direction and size of the relationships between covariates and model parameters. However, in the case of correlated covariates, the median-split analysis tends to suggest spurious relationships

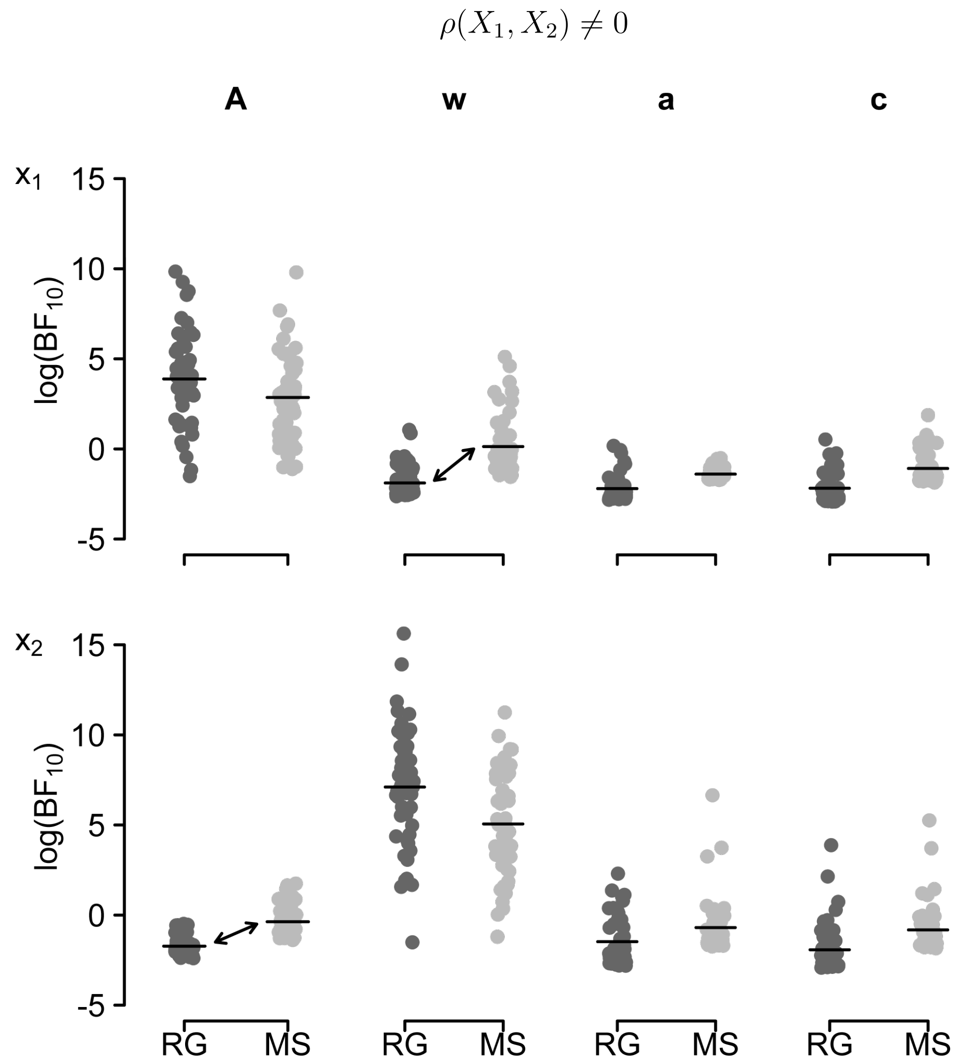

Figure H.2: Bayes factors from 50 simulated data sets for the regression and median-split analysis with correlated covariates. Data points show the log Bayes factors for the alternative hypothesis ($\log(\text{BF}_{10})$) obtained in the regression (RG, dark grey dots) and median-split (MS, light grey dots) analysis for the PVL-Delta model's A and w parameters (columns) and two covariates (rows). Lines indicate the mean log BF. Data points are jittered along the x-axis for improved visibility.

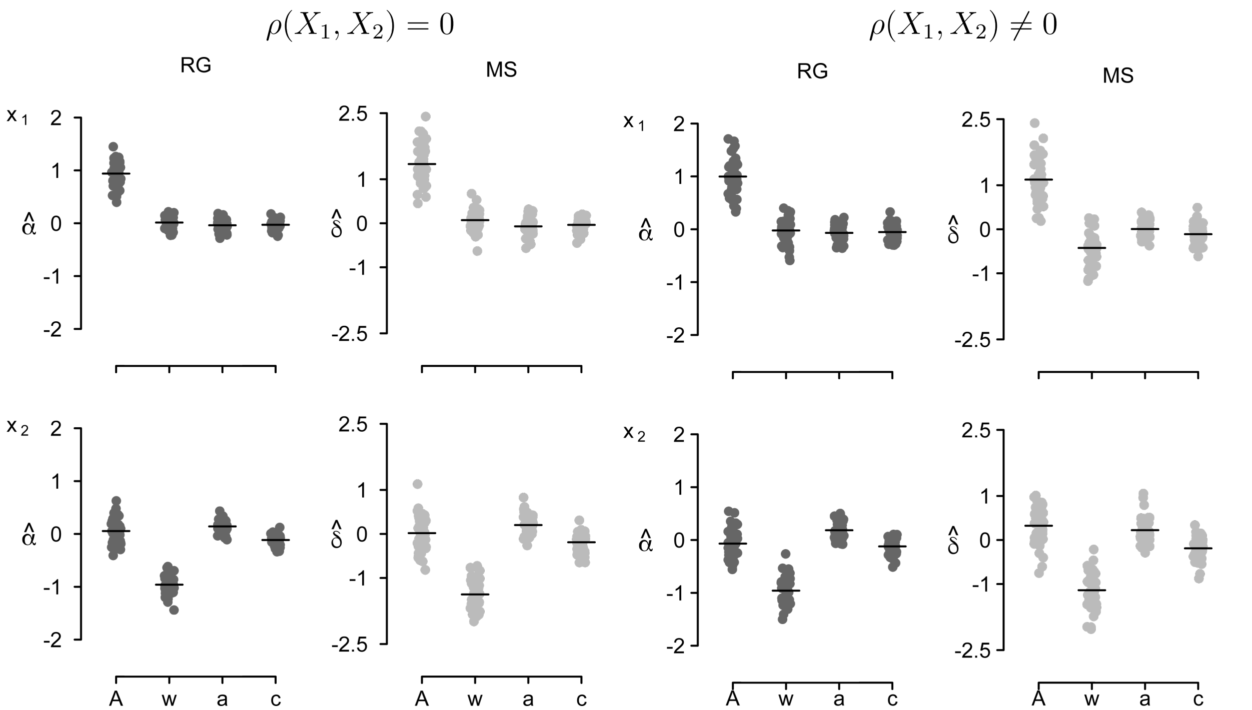

Figure H.3: Mean posterior estimates from 50 simulated data sets of effects for the regression and median-split analysis for uncorrelated (left subplot) and correlated (right subplot) covariates. Data points show estimated standard effect sizes ($\hat{\alpha}$) for the PVL-Delta+R (RG; left panels in each subplot, dark grey dots) and the estimated standardized effect size ($\hat{\delta}$) from the median-split analysis (MS; right panels in each subplot, light grey dots) for the four PVL-Delta parameters. The top row shows the results for the first covariate, the bottom row for the second covariate. Black lines indicate the mean across simulations. Data points are jittered along the x-axis for improved visibility.

between covariates and model parameters. The direction of these spurious effects is equal to the direction of the true effects. This suggests a spill-over from one covariate to the other that arises from the fact that the median-split analysis ignores the correlation between covariates, whereas the regression analysis partials out correlations between covariates.

H.2 Priors for the Regression Coefficients

In this appendix we provide detailed information about the priors for the regression coefficients in our model. The overarching goal in regression analysis as used in the present work is to select the set of predictors that best accounts for the observed data. This selection process can be understood as a model comparison exercise. Consider, for example, a situation where a researcher has a criterion variable y and a single predictor variable x and wants to know whether x has any predictive value for y. To answer this question, the researcher constructs two models, a null model $\mathcal{M}_0$ that only includes an intercept term, and an alternative model $\mathcal{M}_1$ that includes an intercept term and the predictor variable x, and compares the adequacy of the two models. Following the Bayesian approach, such model comparisons can be carried out by computing the

Bayes factor, that is, the ratio of the marginal likelihood of the observed data under the two models, $\mathrm{BF}_{10} = p(y \mid \mathcal{M}_1)/p(y \mid \mathcal{M}_0)$.

Bayes factors need to fulfill a number of theoretical desiderata (Bayarri, Berger, Forte, & García-Donato, 2012; Rouder & Morey, 2012). First, Bayes factors should be location and scale invariant. In the case of regression models, this means that the scale on which the criterion and predictor variables are measured (e.g., kilograms, grams, milligrams) and the location of the zero-point of the scale (e.g., temperature in Celsius vs. in Kelvin) should not influence the Bayes factor. Second, Bayes factors should be consistent, which means that as sample size approaches infinity, the Bayes factor should converge to the correct bound (i.e., $\mathrm{BF}_{10} \to 0$ if $\mathcal{M}_0$ is correct and $\mathrm{BF}_{10} \to \infty$ if $\mathcal{M}_1$ is correct). Third, Bayes factors should be consistent in information, which means that the Bayes factor should not approach a finite value as the information provided by the data in favor of the alternative model approaches infinity. In the case of regression models this means that, as the coefficient of determination, R^2, in $\mathcal{M}_1$ approaches 1, the Bayes factor should go to infinity.

One factor that critically influences the behavior of the Bayes factors is the choice of the priors for the model parameters. Assigning improper priors to model-specific parameters, for instance, leads to indeterminate Bayes factors (Jeffreys, 1961). In our example with a single predictor x, the corresponding regression weight β_x is included in $\mathcal{M}_1$ but not in $\mathcal{M}_0$. If β_x is assigned an improper prior that is only determined up to a multiplicative constant, this constant will appear in the numerator of the Bayes factor but not in the denominator, which means that it will not cancel and the Bayes factor will depend on the multiplicative constant. Consequently, researchers need to choose the prior distribution for the model parameters in such a way that model comparisons yield Bayes factors with the desired theoretical properties.

An additional criterion in selecting priors for the model parameters is the degree to which priors are noninformative. In many situations, researchers have little information about the range in which the model parameters, that is, the regression weights, should fall. Therefore, the weights should be assigned a prior that puts little constraint on the possible values. One prior that has regularly been used in regression problems is Zellner's g-prior (Zellner, 1986). In the case of P predictor variables and N observations for each variable, this prior takes the form:

$$\boldsymbol{\beta} \mid g \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}),$$

where $\boldsymbol{\beta}$ is the $P \times 1$ vector of regression weights, g is a scaling factor, σ^2 is the residual variance of the criterion variable, $\mathbf{0}$ is $P \times 1$ vector of zeros, $\boldsymbol{X}$ is the $N \times P$ design matrix containing the predictor variables, and $\boldsymbol{\mathcal{N}}$ denotes the multivariate normal distribution. The degree to which this prior is informative is controlled by its variance-covariance matrix, which in turn depends on g, σ^2, and $\boldsymbol{X}$. In Zellner's (1986) conceptualization of this prior, the design matrix should be treated as a constant; the prior can then be interpreted as the prior on the regression weights arising from a repetition of the experiment with the same design matrix. The reciprocal of the residual variance, that is, the precision ϕ should be assigned Jeffreys' prior (Jeffreys, 1961) $p(\phi) \propto 1/\phi$. Finally, the scaling factor g controls the weight given to the prior relative to the data. For example, $g = 10$ means that the data are given 10 times as much weight as the prior. The scaling factor thus controls how peaked or how informative the prior is.

Another way to understand the effect of the scaling parameter is to consider the shrinkage factor $g/(1+g)$ (Liang et al., 2008; Wetzels, Grasman, & Wagenmakers, 2012). Using this shrinkage factor, the posterior mean for $\boldsymbol{\beta}$ can be estimated as the product of the shrinkage factor and the least-squares estimate of the regression weights, $\boldsymbol{\beta}_{OLS}$. Consequently, if g is set to a small value, the posterior estimate of $\boldsymbol{\beta}$ will be pulled towards 0 whereas a high value of g leads to a posterior mean that is similar to the least-squares estimate.

The question that remains is how g should be set. One popular choice is to set $g = N$, which yields a unit information prior (Kass & Raftery, 1995). Specifically, the term $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ in the expression for the variance-covariance matrix of the prior equals the variance of the maximum-likelihood estimator of the regression weights, $\text{var}(\boldsymbol{\beta}_{OLS})$. This estimate is based on the design matrix with N rows, which conveys the information of N observations. Therefore, by setting g to N, the influence of the design matrix on the prior can be made equivalent to the information contained in a single observation[1].

However, as shown by Liang et al. (2008), the Zellner prior in its general form suffers from two shortcomings. First, if g is set to a fixed value, the resulting Bayes factors will suffer from the "information paradox". This means that when a model $\mathcal{M}_1$ is compared to the null model $\mathcal{M}_0$, and the coefficient of determination R^2 under $\mathcal{M}_1$ goes to 1 (i.e., there is infinite support for $\mathcal{M}_1$), the Bayes factor will tend to a finite value, thus violating the theoretical desideratum of consistency in information. Second, if g is set to a very large value to make the prior noninformative, Bayes factors will suffer from the Jeffreys-Lindley-Bartlett paradox. This means that $\mathcal{M}_0$ will unduly be favored. In the limiting case when $g \to \infty$, the Bayes factor will go to zero, irrespective of the information provided by the data, thus violating the theoretical property of consistency.

The problems of the Zellner prior are resolved by Jeffreys-Zellner-Siow prior (JZS prior; Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015). In the approach suggested by Zellner and Siow (1980), the regression coefficients are assigned a multivariate Cauchy prior, which satisfies the consistency requirements on the Bayes factors (Liang et al., 2008; Wetzels et al., 2012). One slight drawback, however, of the multivariate Cauchy prior is that the marginal likelihood of the data under a model cannot be computed in closed form and numerical approximations require the computation of the P-dimensional integrals, which become unstable for models with large numbers of predictors. As pointed out by Liang et al. (2008), a remedy to this problem is to express the multivariate Cauchy distribution as a mixture of g-priors. In this approach, the scaling factor g in the Zellner prior is treated as a random variable:

$$\boldsymbol{\beta} \mid g \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, g\sigma^2(\boldsymbol{X}^T\boldsymbol{X}/N)^{-1}),$$

and g is assigned an inverse-gamma prior:

$$g \sim \mathcal{IG}(1/2, s^2/2)$$

with shape parameter $1/2$ and scale parameter $s^2/2$. Note that the scale parameter s of the inverse gamma distribution is equal to the scale parameter of the multivariate Cauchy distribution. This form of the JZS prior combines the favorable theoretical properties of the multivariate Cauchy prior with the computational advantages of the mixture representation. Specifically, using the mixture representation of the JZS prior reduces the computation of a Bayes factor to a one-dimensional integral that can be computed numerically with great precision (Rouder & Morey, 2012).

The above discussion shows that using the JZS prior yields Bayes factors that are consistent and consistent in information, and thus satisfy two of the three desiderata. The third desideratum, location and scale invariance, can be achieved by assigning the JZS prior to the vector of standardized effect sizes $\boldsymbol{\alpha}$, rather than the vector of regression weights $\boldsymbol{\beta}$. The elements of the vector of standardized effect sizes are given by:

$$\alpha_i = \beta_i \frac{s_{x_i}}{\sigma},$$

[1]Note that because the design matrix appears in the expression for the prior in the inverse of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, multiplying $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ by N is equivalent to diving $(\boldsymbol{X}^T\boldsymbol{X})$ by N.

where β_i is the regression weight for the ith predictor variable, s_{x_i} is the standard deviation of the ith predictor variable and σ is the residual standard deviation of the criterion variable. Returning to our initial example with a single predictor variable x and criterion variable y, the standardized effect size is given by $\alpha_x = \beta_x \frac{s_x}{\sigma}$. The researcher could now assign α_x a JZS prior and compute the marginal likelihood of the data under model $\mathcal{M}_1$, which includes an intercept term and x as a predictor, and under model $\mathcal{M}_0$, which only includes an intercept term. The resulting Bayes factor will then satisfy all three desiderata, being invariant to linear transformations of x and y, favoring the correct model as the sample size goes to infinity, and not approaching a finite asymptote as evidence for $\mathcal{M}_1$ approaches infinity.