Multivariate quasi-beta regression models for continuous bounded data

Supplementary Material

Web Appendix A

In this Appendix we present additional details of the percentage of body fat data set. The participants of the study signed the informed consent form, approved by the Research Ethics Committee of the HC-UFPR.

Healthy men and women between the ages of 18 and 90 years whitout any hormonal treatment or medications that could interfere with body composition, either for replacement or supplementation, with a body mass index (BMI) of between 18.5 and 29.9 kg/m^2, without any physical incapacity and walking without the aid of orthoses or prostheses. Individuals with chronic diseases and licit or illicit drugs or drugs known to affect body composition, such as insulin-dependent diabetes, corticosteroids, thyroid hormone in suppressive doses, and those with low weight, consistent with BMI less than 18.5 kg/m^2 or obese, with BMI of at least 30 kg/m^2 were excluded of the research.

All participants performed the same day anthropometric measures (weight and height) and answered the questionnaire on sociodemographic data, followed by the total body densitometry (Lunar Prodigy Advance PA + 302284) for the analysis of the body's fat, lean and bone masses total. The test was evaluated according to the recommendation of the International Society for Clinical Densitometry (Petak et al., 2013, Kendler et al., 2013).

All participants responded to the IPAQ (International Physical Activity Questionnaire), validated in Portuguese (Matsudo et al., 2001), which is an instrument used to estimate the level of physical activity practiced routinely. The IPAQ-short, composed of eight questions about the performance, frequency and duration of moderate, vigorous or walking physical activities was used. The IPAQ was answered in the form of self-administration for the majority of volunteers or as an individual interview, applied by the investigator or trained evaluator, in cases in which there was difficulty of understanding. The volunteers were then divided into three groups, according to the level of physical activity performed (Nahas, 2001, Sonati, 2012): sedentary are those who do not perform any physical activity for at least 10 continuous minutes during the week; insufficiently active, perform at least 10 continuous minutes of physical activity, at least 5 days a week or 150 minutes a week, but insufficiently to be classified as active. Assets are individuals who perform at least 20 minutes of vigorous physical activity per session, at least 3 times a week or moderate activities, or 30 minutes walk per session, at least 5 times a week, or any activity added for 5 days week or more, with a total duration of 150 minutes per week (Silva et al., 2007).

Figure 1 shows dispersion diagrams with smoothing curves estimated by the loess method (Cleveland, 1979), in addition to showing the correlations between fat percentage in the arms, legs, trunk, android and gynoid regions.

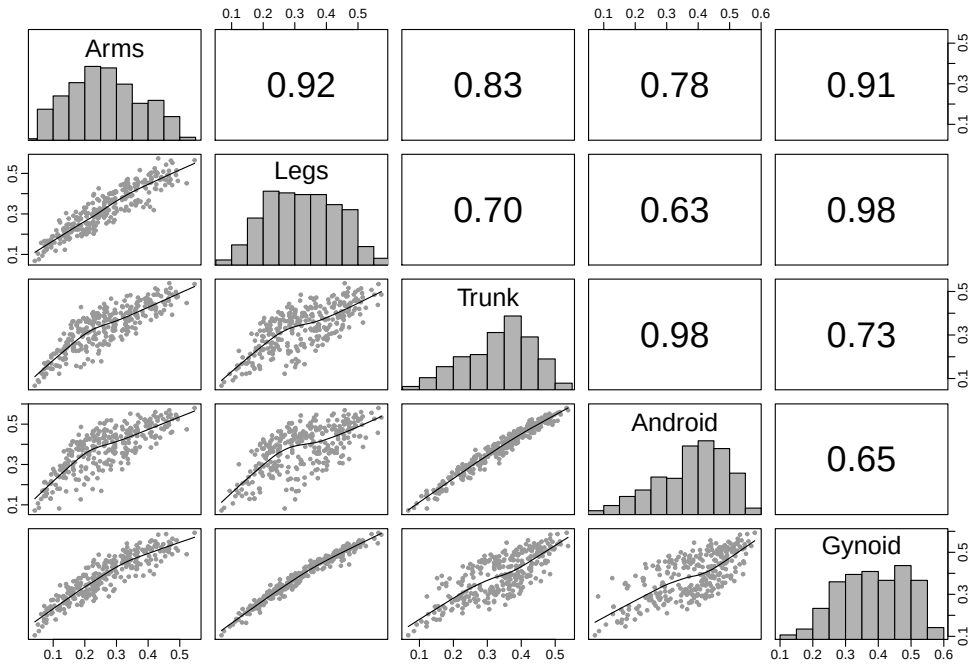

Figure 1: Dispersion diagrams and correlations between the body fat percentages in the regions of the arms, legs, body, android and gynecoid.

According to the results presented in Figure 1, all correlations are positive, with stronger correlations between the fat percentages in the arms and legs regions ($\hat{\rho} = 0.92$), trunk and android ($\hat{\rho} = 0.98$), arms and gynecoid ($\hat{\rho} = 0.91$) and between legs and gynecoid ($\hat{\rho} = 0.98$). On the other hand, moderate correlations can be observed between the body fat percentages in the legs and android regions ($\hat{\rho} = 0.63$) and between android and gynecoid ($\hat{\rho} = 0.65$). It is important to note that these correlations were estimated by the Spearman correlation coefficient ($\hat{\rho}$) and do not take into account the effect of the covariates available in the study.

Finally, Figure 1 shows the empirical distribution of each response variable by means of a histogram, indicating symmetric distributions for

most of them. However, asymmetric distributions on the left can be seen for body fat percentages in the trunk and android regions.

Web Appendix B

NORTA algorithm

The NORTA algorithm (Cario and Nelson, 1997) is one of the most popular methods for simulating non-Gaussian correlated random vectors. The method works as a two-step process. First, a multivariate normal random vector $\mathbf{Z}$ is generated. Then, this vector is transformed into a multivariate uniform vector $\mathbf{U}$, which is again transformed into vector $\mathbf{Y}$ which has distribution NORTA (*NORmal To Anything*), where each element of the vector has a desired arbitrary marginal distribution. Therefore, its representation is given by:

$$\mathbf{Y} = \left[F_{Y_1}^{-1}(\Phi[Z_1]), F_{Y_2}^{-1}(\Phi[Z_2]), \ldots, F_{Y_p}^{-1}(\Phi[Z_p])\right]^\top \text{for } l = 1, \ldots, p, \quad (1)$$

where $\Phi[\cdot]$ is the cumulative distribution function (cdf) of the standard Gaussian distribution applied to each element of the vector $\mathbf{Z}$ and $F_{Y_l}^{-1}(u) \equiv \inf\{y : F_{Y_l}(y) \geq u\}$ denotes the inverse cdf.

The correlation matrix of $\mathbf{Z}$ directly determines the correlation matrix of $\mathbf{Y}$, provided that

$$\rho_Y(l, l') = \text{Corr}(Y_l, Y_{l'}) = \text{Corr}\big(F_{Y_l}^{-1}(\Phi[Z_l]), F_{Y_{l'}}^{-1}(\Phi[Z_{l'}])\big),$$

for all $l \neq l'$. The correlation is defined by:

$$\text{Corr}(Y_l, Y_{l'}) = \frac{\text{E}(Y_l, Y_{l'}) - \text{E}(Y_l)\text{E}(Y_{l'})}{\sqrt{\text{Var}(Y_l)\text{Var}(Y_{l'})}}, \quad (2)$$

where marginal quantities $\text{E}(Y_l), \text{E}(Y_{l'}), \text{Var}(Y_l)$ and $\text{Var}(Y_{l'})$ are defined by F_{Y_l} and $F_{Y_{l'}}$. It is worth mentioning that $(Z_l, Z_{l'})$ has stardard bivariate Gaussian distribution with correlation $\text{Corr}(z_l, z_{l'}) = \rho_Z(l, l')$ where the quantity $\text{E}(Y_l, Y_{l'})$ in (2) is calculated by:

$$\begin{aligned} \text{E}(Y_l, Y_{l'}) &= \text{E}\Big(F_{Y_l}^{-1}(\Phi[Z_l]) F_{Y_{l'}}^{-1}(\Phi[Z_{l'}])\Big) \quad (3) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{Y_l}^{-1}(\Phi[Z_l]) F_{Y_{l'}}^{-1}(\Phi[Z_{l'}]) \varphi_{\rho_Z(l,l')}(z_l, z_{l'}) dz_l dz_{l'}, \end{aligned}$$

where $\varphi_{\rho_Z(l,l')}$ denotes the probability density function of a stardard bivariate Gaussian distribution with correlation given by $\rho_Z(l, l')$.

It should be noted that the integral (3) will have a solution, due to the mean/variance relationship of the beta distribution. Such a constraint directly impacts the parametric space of the correlation (2) which also depends on the specification of the marginal means.

Evaluating the behavior of the NORTA algorithm to simulate correlated beta random variables

The main goal of this simulation study is to investigate the behavior of the NORTA algorithm to simulate bivariate beta random variables. We used the `R` statistical software (R Core Team, 2019) and the `NORTARA` package (Su, 2014), which provides the computational implementation of the NORTA algorithm.

For the case of the multivariate beta distribution the main challenge is to identify the minimum and maximum values allowed for the correlation between responses given the marginal expectations and dispersion parameters. Thus, we considered a bivariate case, where we denote the response variables by Y_1 and Y_2 and set five different values for the dispersion parameters $\sigma^2 = (0.99, 0.60, 0.20, 0.10, 0.04)$. These values ranging from a very challenging case, i.e $\sigma^2 = 0.99$ where the generated data are approximately only 0's and 1's to a simple situation where we have symmetric data. For each marginal distribution, we fixed the marginal expectation as a sequence with 100 values between $(0,1)$. Then, we constructed a grid of values with 10,000 points (100×100) to evaluate the correlation matrix between the marginal beta distributions. Our main interest is to find the minimum and maximum correlation allowed for the bivariate distribution given the marginal expectation and dispersion parameters. In order to obtain such values, we used the function `valid_input_cormat()` of the NORTARA package. This function returns the minimum (ρ_L) and maximum (ρ_U) values that the correlation matrix can assume given the marginal distributions specified.

For example, when $\mu_1 = 0.495$ and $\mu_2 = 0.851$ with σ^2 fixed at 0.99, the lower and upper limits are $\rho_L = -0.421$ and $\rho_U = 0.413$, respectively. However, for the same values of the marginal expectation, but σ^2 fixed at 0.10, both lower and upper limits are $\rho_L = -0.956$ and $\rho_U = 0.954$. These results shows how the covariance structure depending on the marginal expectation of the beta distribution and, consequently, the limits of the correlation matrix.

To make this idea more general, we constructed Figure 2. In Figure 2, the upper part shows the minimum limits, while the lower part shows the

maximum limits that the correlation between the two beta random variables assume as a function of their marginal expectations for each value of the dispersion parameter σ^2. According to the results presented in Figure 2, when we have high values for σ^2 the obtained minimum correlation becomes restricted, especially when $\sigma^2 = (0.99 \text{ and } 0.60)$. As the value of σ^2 decreases, stronger correlations are allowed in the darker regions of the graph. The same is observed for the maximum values of the correlation (Figure 2). Thus, it was observed that low values of the dispersion parameter combined with high/low values of the marginal expectations allow stronger correlations.

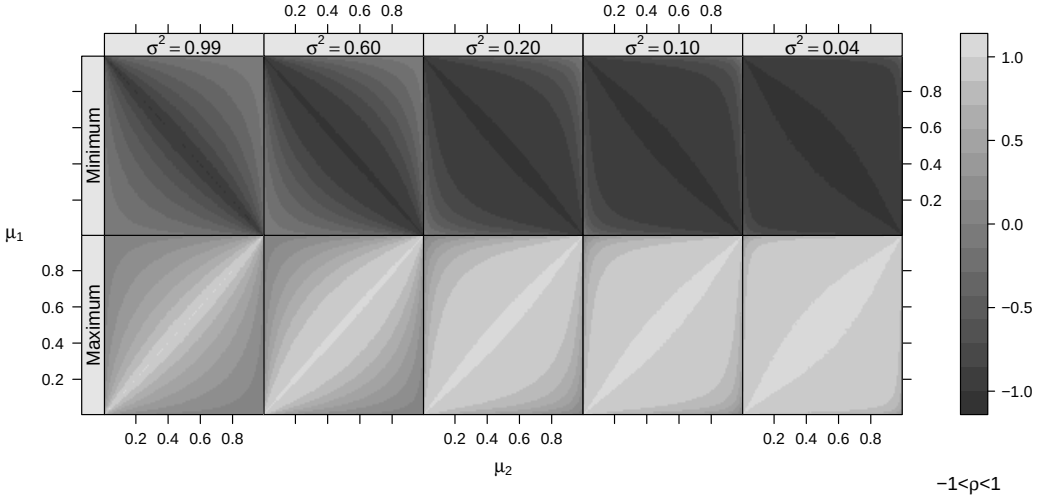

Figure 2: Minimum and maximum values for the correlation between two beta random variables as a function of the marginal expectations and different values of the parameter σ^2.

The results of the simulation study showed that the parametric space of the correlation parameter was reduced when high values were obtained for the dispersion parameters associated with high/low values of the marginal means. Considering the results obtained, we have an idea of the behavior of the NORTA algorithm that will be used in the next simulation study conducted to evaluate the performance of the estimating functions estimator for the parameters of the model proposed in Section 4 (main document).

Web Appendix C

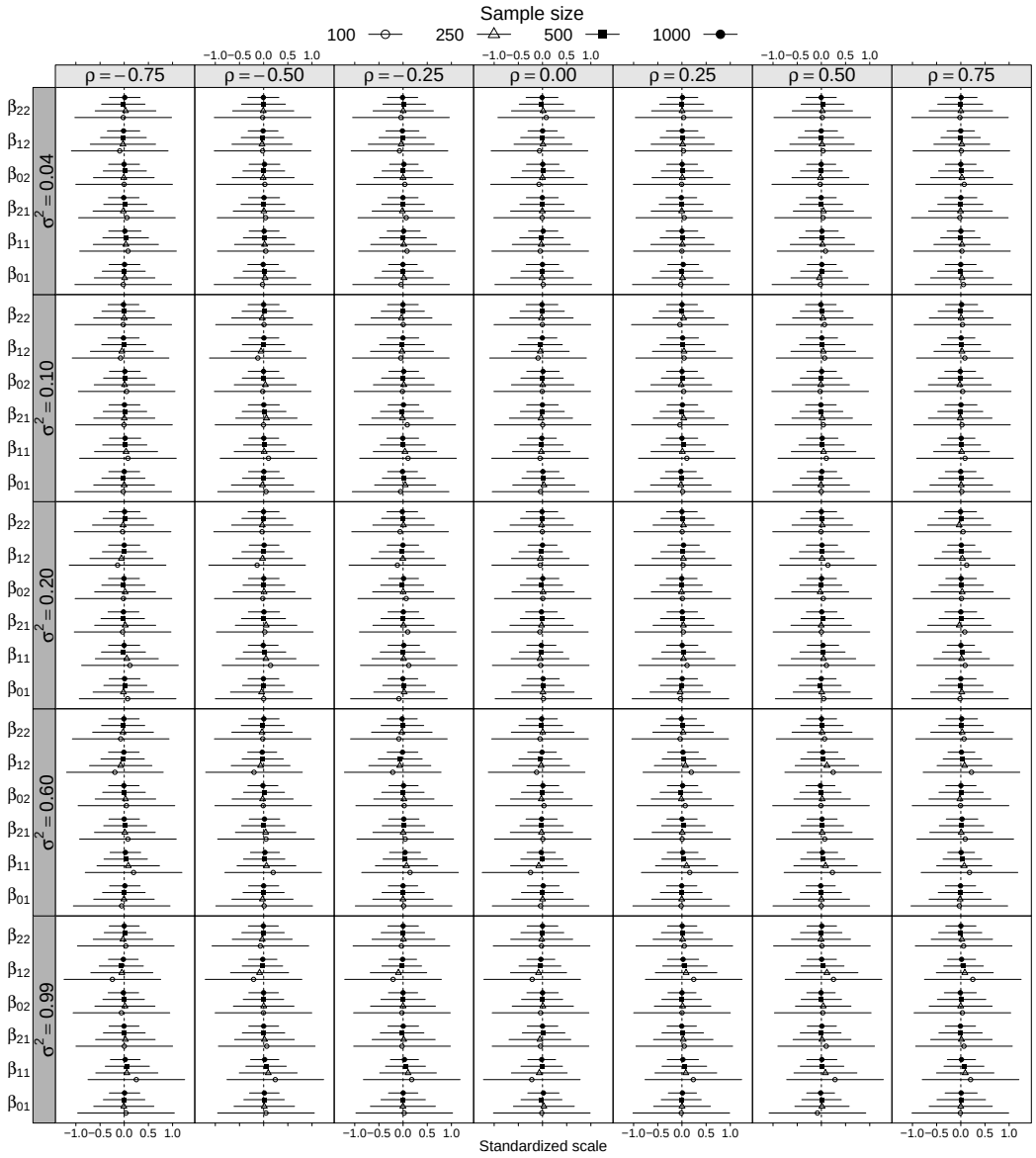

Figure 3: Average bias and confidence intervals on a standardized scale for the regression coefficients $(\beta_{01}, \beta_{11}, \beta_{21}, \beta_{02}, \beta_{12}, \beta_{22})$ by sample size and simulation scenarios.

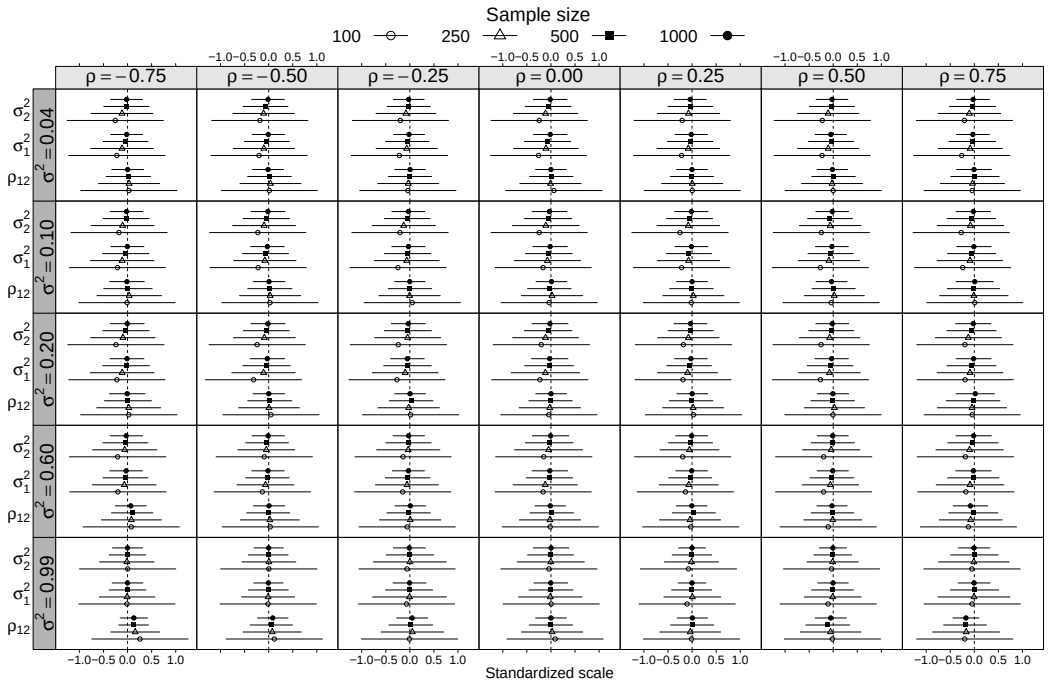

Figure 4: Average bias and confidence intervals on a standardized scale for each parameter $(\rho_{12}, \sigma_1^2, \sigma_2^2)$ by sample size and simulation scenarios.

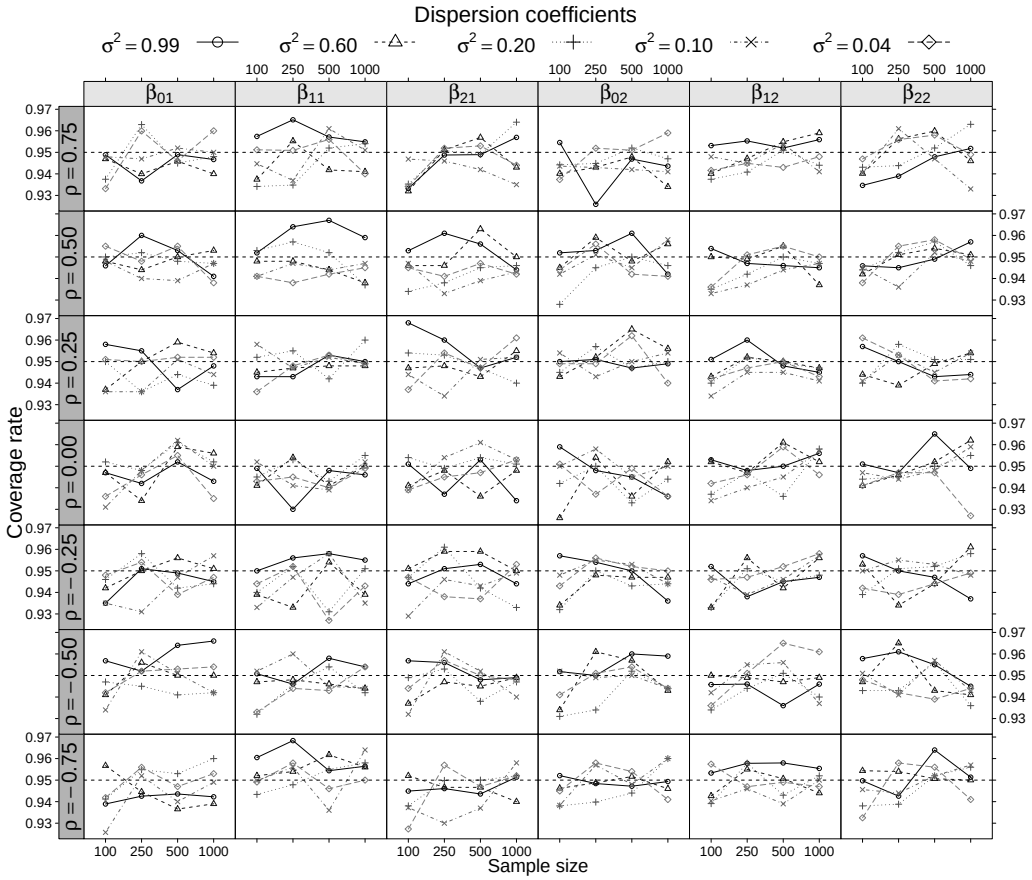

Figure 5: Coverage rate for each parameter $(\beta_{01}, \beta_{11}, \beta_{21}, \beta_{02}, \beta_{12}, \beta_{22})$ by sample size and simulation scenarios.

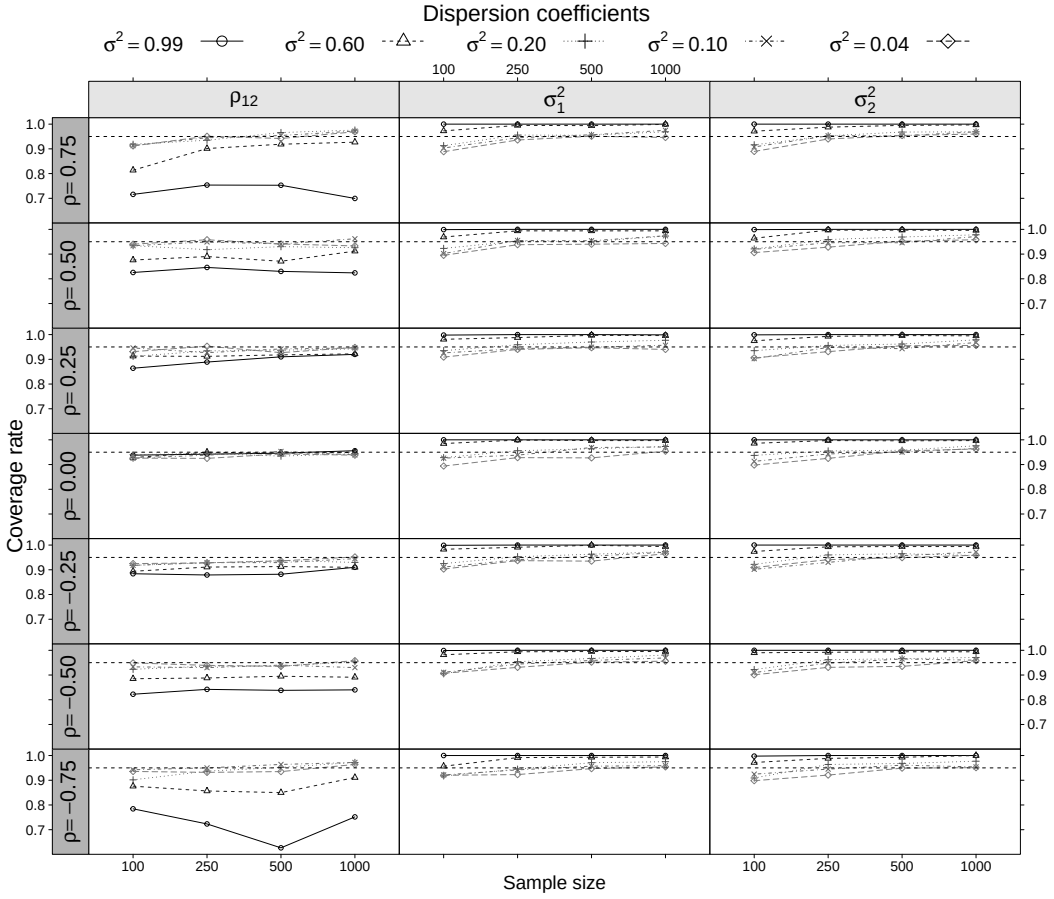

Figure 6: Coverage rate for each parameter $(\rho_{12}, \sigma_1^2, \sigma_2^2)$ by sample size and simulation scenarios.

Web Appendix D

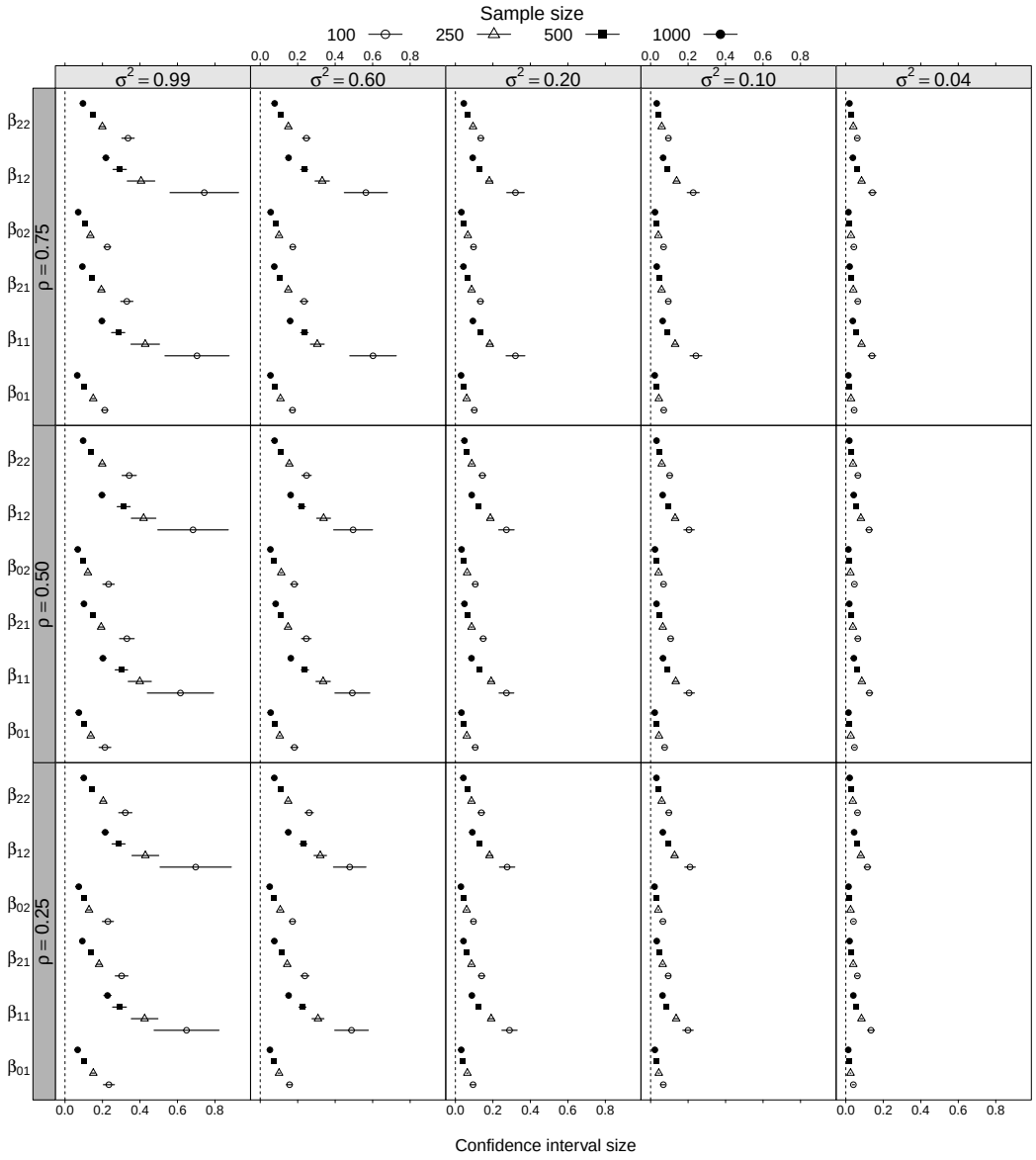

Figure 7: Confidence interval size for the regression coefficients $(\beta_{01}, \beta_{11}, \beta_{21}, \beta_{02}, \beta_{12}, \beta_{22})$ by sample size and some simulation scenarios.

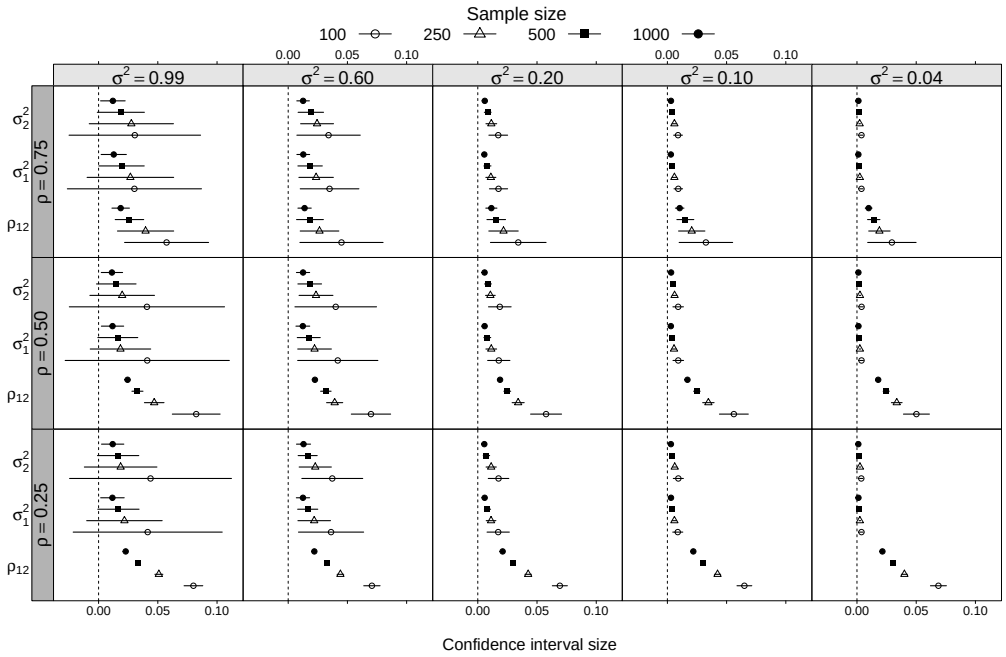

Figure 8: Confidence interval size for each parameter $(\rho_{12}, \sigma_1^2, \sigma_2^2)$ by sample size and some simulation scenarios.

Web Appendix E

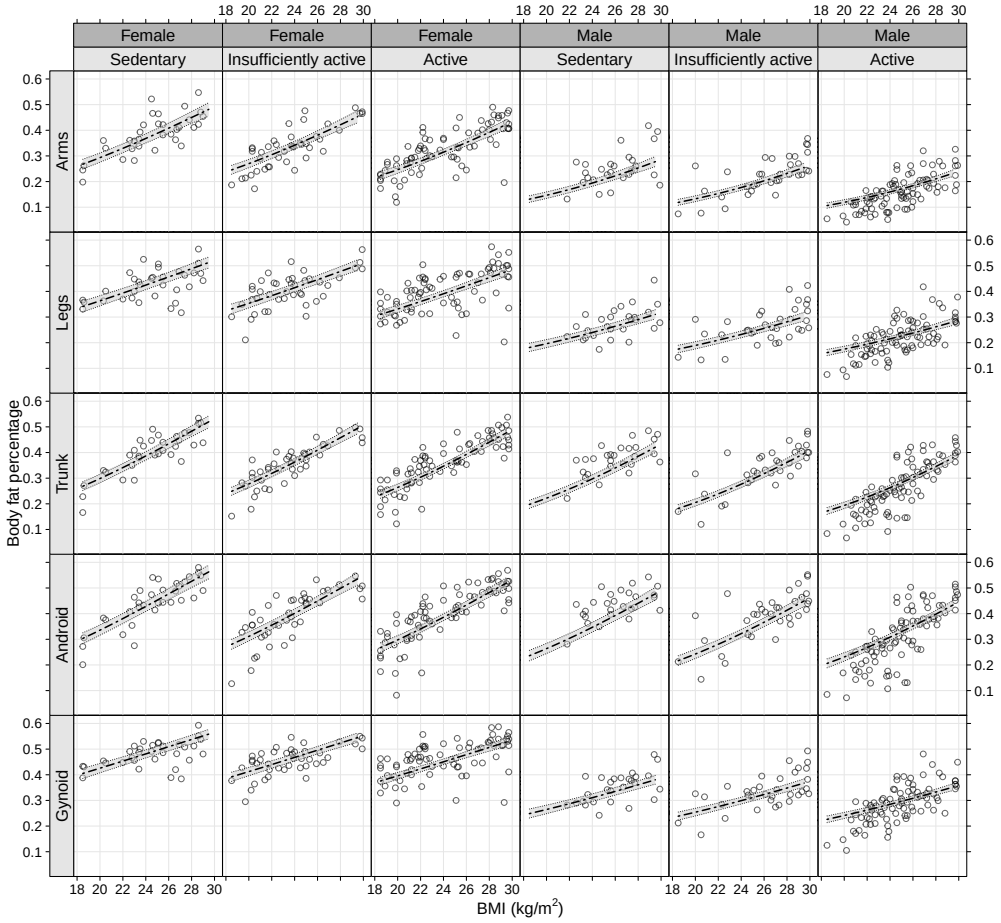

Figure 9: Curves of fitted values with 95% confidence intervals by gender, IPAQ and BMI for the quasi-beta regression models by response variables.

References

Cario, M. C. and B. L. Nelson (1997): "Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix," Technical report, Citeseer.

Cleveland, W. S. (1979): "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, 74, 829–836.

Kendler, D. L., J. L. Borges, R. A. Fielding, A. Itabashi, D. Krueger, K. Mulligan, B. M. Camargos, B. Sabowitz, C.-H. Wu, W. Y. Elaine, et al. (2013): "The official positions of the international society for clinical densitometry: indications of use and reporting of dxa for body composition," *Journal of Clinical Densitometry*, 16, 496–507.

Matsudo, S., T. Araújo, V. Matsudo, D. Andrade, E. Andrade, L. C. Oliveira, and G. Braggion (2001): "Questionário internacional de atividade física (ipaq): Estudo de validade e reprodutibilidade no brasil," *Revista Brasileira de Atividade Física & Saúde*, 6, 5–18.

Nahas, M. V. (2001): *Atividade física, saúde e qualidade de vida: conceitos e sugestões para um estilo de vida ativo*, Midiograf.

Petak, S., C. G. Barbu, W. Y. Elaine, R. Fielding, K. Mulligan, B. Sabowitz, C.-H. Wu, and J. A. Shepherd (2013): "The official positions of the international society for clinical densitometry: body composition analysis reporting," *Journal of Clinical Densitometry*, 16, 508–519.

R Core Team (2019): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Silva, G. d. S. F. d., R. Bergamaschine, M. Rosa, C. Melo, R. Miranda, and M. Bara Filho (2007): "Avaliação do nível de atividade física de estudantes de graduação das áreas saúde/biológica," *Revista Brasileira de Medicina do Esporte*, 13, 39–42.

Sonati, J. (2012): *Qualidade de Vida e Composição Corporal: Características do Envelhecimento Bem Sucedido. 84 p.*, Ph.D. thesis, Universidade Estadual de Campinas.

Su, P. (2014): *NORTARA: Generation of Multivariate Data with Arbitrary Marginals*, R package version 1.0.0.

Supplementary material

http://www.leg.ufpr.br/doku.php/publications:papercompanions:multquasibeta