Impact of collinearity on classical and Bayesian model choice criteria

---

Master Thesis in Biostatistics (STA495)

by

Georgios Kazantzidis
20-742-169

supervised by

Malgorzata Roos, PD Dr.

Zurich, February 2022

Acknowledgments

The present master thesis is part of the Biostatistics Master Programme of the University of Zurich. It was completed in the academic semester of Spring 2022.
Before I start the analysis of my work, it is important to express my acknowledgments to everyone who contributed in achieving this work. Firstly, I would like to deeply thank my supervisor, PD Dr Malgorzata Roos, who guided me through this master thesis, orienting my investigation while at the same time allowing me to be flexible and independent. Along with her, I would like to thank the members of her team. Lucas Kook, who provided me with meaningful comments throughout the semester, Samuel Pawel, who inspired me with his feedback, and Sona Hunanyan. Moreover, I would like to thank all my professors who taught me the principles and applications of statistics. Special thanks to Professor Leonhard Held, who provided consultation and motivated me through many courses of the programme. Professor Torsten Hothorn, whose lectures helped me to improve my coding skills in the R programming language. Dr Eva Furrer, who always supported me thoroughly throughout the duration of the Master's programme. Professor Reinhard Furrer, Professor Beate Sick, Dr Zofia Baranczuk-Turska and Professor Michael Höhle for their insightful lectures. My former supervisors, Professor John M. Halley and Professor Jason Matthiopoulos for introducing me in the best way to the world of classical and Bayesian statistics.
Last but not least, I would like to especially thank: Corina, Annina, Lina and Carlos for supporting me with their hospitality. Their help was extremely important for my studies in Zurich. My friends who supported me through this demanding period and my family, Savvas, Chrysoula and Aglaia-Io, who are supporting me at all times and circumstances.

Georgios

Version February 28, 2022

Contents

Preface

Multiple linear regression models are one of the most common methods of statistical analysis. However, these models can be affected by collinearity, which leads to misleading estimates of regression coefficients. Therefore, predictors should be diagnosed for collinearity. To clarify the impact of collinearity, Belsley proposed a principled approach based on singular value decomposition of equilibrated design matrices. The choice of the optimal model is also a relevant question for multiple linear regression models. However, collinearity can potentially affect classical and Bayesian model choice criteria and the choice of the best model. Therefore, it is necessary to investigate the sensitivity of model choice criteria to collinearity. In this project, we investigate the impact of collinearity on both classical and Bayesian model choice criteria through both a simulation and a case study. Moreover, we address the instability of the regression estimates caused by collinearity. Finally, we provide on GitHub a new and openly accessible R package for collinearity diagnostics based on the approach suggested by Belsley. This package can support other researchers who want to clarify collinearity issues in their multiple linear regression models.

Georgios Kazantzidis
February 2022

Chapter 1

Introduction

One of the most frequently used methods of statistical analysis in science is the multiple linear regression. Often, the analysed data set contains more than one predictor variables. In such cases, a model selection procedure is performed. The goal of the model selection procedure is to obtain the optimal regression model out of all considered. For the selection of the best available model, a number of criteria for model comparisons are available. The model choice criteria is a frequently used index which identifies the best available model. Each model choice criterion, uses a different approach to identify the best model.

Multiple linear regression models are applied both in classical (frequentist) and Bayesian framework. The frequentist approach uses the least squares method to estimate the regression coefficients. The main difference in the Bayesian framework is the involvement of prior distributions and the distributional formulation of the estimated parameters. Usually, normal priors with a large variance are assumed for all the predictors in order to obtain the posterior distribution of the coefficients.

Numerous different model choice criteria are available for evaluation of classical and Bayesian models. AIC and BIC are the mostly used frequentist model choice criteria and DIC, WAIC, LCPO and LML are Bayesian model choice criteria [Akaike, 1974, Schwarz, 1978, Spiegelhalter et al., 2002, Watanabe and Opper, 2010, Pettit, 1990, Gkisser, 2017, Roos and Held, 2011, Chib, 1995, Chib and Jeliazkov, 2001, Gómez-Rubio and Rue, 2018, Gómez-Rubio et al., 2021]. Although the main goal of the model choice criteria is to indicate the best model, as we explain below, each model choice criterion follows its own principles.

Collinearity is the phenomenon where at least one of the predictors in a regression can be described as a linear combination of other predictors. The statistical and inferential problems of collinearity in multiple regression are well established in the statistical literature [Cohen et al., 1983, Hocking, 2013, Neter et al., 1996, Tabachnick and Fidell, 1996, Draper and Smith, 1998, Chatterjee et al., 2000, Graham, 2003]. The level of collinearity in a multiple linear regression model depends only on the predictor variables and it can be measured based on the design matrix X. Montgomery et al. [2021] introduces the condition number as a preferable measure of collinearity diagnostics. The condition number is based on the eigensystem analysis of the X^TX matrix of the regression predictors. In this setting, small (close to zero) eigenvalues of X^TX matrix identify near linear relationships between the predictors. Moreover, their eigenvectors indicate the predictors involved in collinearity.

Belsley [1991] suggests the singular value decomposition of the design matrix X along with the equilibration of this matrix as a more principled approach towards collinearity diagnostics. Although both Montgomery's and Belsley's methods aim at identifying collinear relations between predictors, the singular value decomposition of matrix X is far more numerically stable than the eigensystem analysis of X^TX. In addition Belsley [1991] suggests the use of condition indexes which compare the maximal singular value to all other singular values. Furthermore, for comparability issues he suggests that the matrix X needs to be equilibrated first. That is, the Euclidean lengths of the columns

of the design matrix X should be equal. Moreover, Belsley [1991] suggests variance decomposition proportions, which weight the contribution of each predictor to collinearity. In practice, singular value decomposition approach combined with equilibration facilitates the comparison of collinearity levels among different models by taking account of the different patterns within different predictor combinations.

The negative effects of collinearity on model estimates are well established [Montgomery et al., 2021]. When collinearity is present, the regression estimates are biased and can be misleading. Thus, the model fit can be dramatically deteriorated by collinearity. In a model selection procedure, the consideration of candidate models with many predictors, increases the chance of harmful collinearity. Identification of such models is of key importance. Although model choice criteria are designed to detect the best model, their sensitivity to collinearity is yet unknown.

Collinearity between the predictors of a regression model does not necessarily imply that the predictors are correlated. Correlation is a simple form of collinearity which involves only two predictors. Thus, correlation diagnosis is a sub-optimal and inadequate alternative to full collinearity diagnosis. The distinction between correlation and collinearity is long known to the scientific community [Belsley, 1991]. Yet, many researchers still consider only correlation coefficients when performing collinearity diagnostics. Therefore, a numerical example is needed to show that collinearity and correlation is not the same thing.

Alongside with the existence of collinearity itself, the sample size has an important impact on regression estimates. Morrissey and Ruxton [2018], Becker et al. [2015], Efendi and Effrihan [2017], Salmerón Gómez et al. [2016] all argue that a sufficient sample size can decrease, or even diminish [Midi et al., 2010] the effects of collinearity. Yet, the role of sample size in the relationship between model choice criteria and collinearity needs to be clarified.

To motivate the applications of the proposed collinearity diagnostics and to research the relationship between model choice criteria and collinearity, we use the case study of Bodyfat data. The Bodyfat dataset has been thoroughly used to exemplify model selection procedures in literature. For example, Heinze et al. [2018] defines a fixed framework model and uses backward AIC-based selection method to improve the model selection procedure in Bodyfat data. Moreover, Pavone et al. [2020] uses the Bodyfat data to demonstrate the benefits from the projected prediction method. Both studies mention that the Bodyfat data can be affected by collinearity.

The R programming language is used as the main tool for statistical analysis. Several packages for collinearity diagnostics already exist, such as for example: `collin` Basagaña and Barrera-Gómez [2021], `mctest` Imdad and Aslam [2020], Imdadullah et al. [2016], Imdad et al. [2019], `lrmest`, `mcvis` Lin et al. [2020] and `multiColl` García et al. [2019]. Moreover these packages provide various tools and methods for diagnosing and illustrating collinearity. However, none of the above packages has implemented the singular value decomposition and equilibrated design matrix methods as suggested by [Belsley, 1991].

In this project, we investigate the impact of collinearity on model choice criteria. We use the approach from Belsley [1991] to quantify collinearity and we create a new R package implementing these methods. We provide simulation results for theoretical scenarios of different levels of collinearity in simple regression models. Finally, we extend the analysis to the Bodyfat case study illustrating the effect of collinearity on regression estimates and model choice criteria in real data.

Chapter 2

Methods

2.1 Multiple linear regression

This project considers classical and Bayesian multiple linear regression models to describe relationships between predictors and the response. Linear regression is a method used in statistical analysis aiming to describe linear patterns between a set of variables. The linear relationships are quantified with the parameter estimates. We refer to the total number of estimated parameters including the intercept, slope and σ^2 of a model with d, whereas the number of predictors is p.

Frequentist method

For the estimation of the regression coefficients in the frequentist case, the least squares estimation was used. In R, this can easily be achieved with the `lm` function provided by the basic R package.

Bayesian method

For the application of the linear models withing a Bayesian framework, the `INLA` program was used. `INLA`, allows the user to conveniently perform approximate Bayesian inference in latent Gaussian models. The R package `INLA` serves as an interface to the `INLA` program and its usage is similar to the `lm` function in R. Its standard output encompasses marginal posterior densities of all parameters in the model together with summary characteristics. Moreover, several model choice criteria can be extracted from the resulting output.

2.2 Model choice criteria in regression

2.2.1 Akaike information criterion

The Akaike information criterion [Akaike, 1974] is one of the most frequently used model choice criterion for a classical simple linear regression model. The formula is:

$$\text{AIC} = -2\log p(y|\hat{\theta}_{mle}) + 2d \tag{2.1}$$

where $\hat{\theta}$ is the maximum likelihood estimator. Lower values indicate a better model.

2.2.2 Schwarz information criterion

The Schwarz, or Bayesian information criterion is another efficient way for the selection of the best model. The formula:

$$\text{BIC} = -2\log p(y|\hat{\theta}_{mle}) + d\ln(n) \tag{2.2}$$

where $\hat{\theta}$ is the maximum likelihood estimator and n is the number of observations [Schwarz, 1978]. This criterion gives a greater penalty on adding predictors as the sample size increases. Lower values indicate a better model.

2.2.3 Deviance information criterion

For Bayesian models, DIC is a frequently used model choice criterion.

$$\text{DIC} = -2\log p(y|\hat{\theta}_{Bayes}) + 2p_{DIC} \tag{2.3}$$

$$p_{DIC} = 2(\log p(y|\hat{\theta}_{Bayes}) - E_{post}(\log p(y|\hat{\theta}))) \tag{2.4}$$

where $\hat{\theta}_{Bayes} = E(\theta|y)$ and p_{DIC} is the effective number of parameters of the model [Spiegelhalter et al., 2002]. Lower values indicate a better model.

The effective number of parameters can be calculated using simulations θ^s, $s = 1, \ldots, S$.

$$\text{computed } p_{DIC} = 2(\log p(y|\hat{\theta}_{Bayes}) - \frac{1}{S}\sum_{s=1}^{S} \log p(y|\theta^s)) \tag{2.5}$$

2.2.4 Watanabe-Akaike information criterion

$$\text{WAIC} = -2\log p(y|\hat{\theta}_{Bayes}) + 2p_{WAIC} \tag{2.6}$$

where the p_{WAIC} can be measured as the variance of individual terms in the log predictive density summed over the the n data points [Watanabe and Opper, 2010, Watanabe, 2013, Gelman et al., 2014]. Lower values indicate a better model.

$$p_{WAIC} = \sum_{i=1}^{n} \text{Var}_{post}(\log p(y_i|\theta^s)) \tag{2.7}$$

Equation (2.7) can be computed from the posterior variance of the log predictive density for each data point y_i, that is, $V_{s=1}^{S} log\, p(y_i|\theta^s)$, where $V_{s=1}^{S}$ is the sample variance over the S posterior draws of θ^s.

$$\text{computed } p_{WAIC} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i|\theta^s)) \tag{2.8}$$

2.2.5 Logarithmic score of conditional predictive ordinate

Conditional predictive ordinates [Pettit, 1990] are a cross - validatory criterion for model assessment that is computed for each observation

$$\text{CPO}_i = \pi(y_i|y_{-i}) \tag{2.9}$$

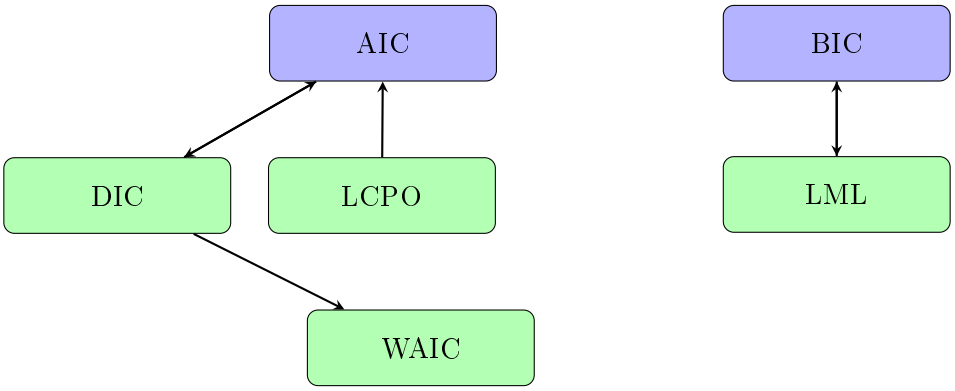

Figure 2.1: Connections between model choice criteria. Blue color indicates classical models and green represents Bayesian framework. The arrows mark the direct connections between the criteria.

That represents the probability of observing an observation y_i when the model is fit using all the observations but y_i. We can calculate this measure from the simulations as

$$\text{LCPO} = -\sum_{i=1}^{n} \log(\text{CPO}_i) \tag{2.10}$$

For this study, we use the negative of LCPO criterion as in equation (2.10) so that smaller values pointing to a better model fit [Gkisser, 2017, Roos and Held, 2011].

2.2.6 Log-marginal likelihood

The marginal likelihood of a model is the probability of the observed data under a given model. For a set of different models M, we can indicate the marginal likelihood as $\{M_m\}_{m=1}^{M}$. The approximation of the marginal likelihood provided by `INLA` is computed as

$$\tilde{\pi}(y) = -\int \frac{\pi(\theta, x, y)}{\tilde{\pi}_G(x|\theta, y)}\bigg|_{x=x*(\theta)} d\theta \tag{2.11}$$

[Chib, 1995, Chib and Jeliazkov, 2001, Gómez-Rubio and Rue, 2018, Gómez-Rubio et al., 2021].

We use the negative value of the LML in order to have the same patern of evaluation with the rest of the used model choice criteria. That is, smaller values of the negative LML correspond to a better model. LML is preferred to compare models with identical priors on the same parameters.

2.2.7 Connections between model choice criteria

Figure 2.1 illustrates the connections between the six model choice criteria considered. Whereas criteria used in the frequentist framework are coloured blue, criteria for comparisons of Bayesian models are green. Arrows between two criteria mark the existing connections. In general, AIC, DIC and WAIC [Gelman et al., 2014], which all are in the left side of the figure, can be seen as approximations to different versions of cross-validation [Stone, 1977]. Firstly, DIC can be seen as a Bayesian version of AIC. Two alterations from the latter are the replacement of the maximum likelihood estimate with the posterior mean and the replacement of the number of parameters used with a data-based bias correction.

Moving on to the WAIC, it is a Bayesian method of estimation of the out-of-sample expectation. Thus, it is connected with DIC and AIC. First, it computes the log pointwise posterior predictive density. Then, the correction for the effective number of parameters is added to adjust for overfitting.

The logarithmic conditional predictive ordinate (LCPO) which is a cross validated predictive density at an observation measures the predictive quality of the model. It is asymptotic equivalent with AIC [Stone, 1977] in the Bayesian framework.

Because BIC aims at the estimation of the predictive fit by approximating the marginal probability of the data under the model, it differs from the previously mentioned model choice criteria. At the right side of the plot, we notice the connection between the BIC and the log marginal likelihood. BIC is asymptotically equivalent to the marginal log likelihood for Bayesian models. For the log marginal likelihood, the effect of the prior becomes less important for increasing sample sizes. At the same way, BIC criterion improves for large sample sizes.

2.3 Collinearity

The presence of collinearity has a number of potentially serious effects on the least squares estimates of the regression coefficients. Some of these effects can be easily demonstrated. Suppose that there are only two regressor variables x_1 and x_2. The model, assuming that x_1, x_2 and y are scaled to unit length, is

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{2.12}$$

and the least squares normal equations are

$$(X^T X)\hat{\beta} = X^T y \tag{2.13}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the simple correlation between x_1 and x_2 and r_{jy} is the simple correlation between x_j and y, (j = 1,2). Now the inverse of $(X^T X)$ is

$$C = (X^T X)^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \tag{2.14}$$

and the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}, \tag{2.15}$$

$$\hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2} \tag{2.16}$$

If there is strong collinearity between x_1 and x_2, then the correlation coefficient r_{12} will be large (close to one). From equation (2.14) we see that as $|r_{12}| \to 1$, $\mathrm{Var}(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$ and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \mp\infty$ depending on whether $r_{12} \to +1$ or $r_{12} \to -1$. Therefore, strong collinearity between x_1 and x_2 results in large variance and covariances for the least-squares estimators of the regression coefficients. This implies that different samples taken at the same x levels could lead to unstable estimates of the models parameters.

Figures 2.2 – 2.5 illustrate the instability of the least squares plane when collinearity is present. The data used for the four figures use the exact same dataset. The fist row contains Figures 2.2 and 2.3. The sample size for this row is $n = 20$. We use the same simulated data for the two plots and we

scale the errors between the two predictors with a scale factor f of 0.1 for the right plot and 50 for the left plot. Whereas the left plot presents the case of low collinearity, the right plot compiles the case of severe collinearity. As it can be observed, the regression plane created for the two cases is more stable for the left, blue plot. On the other hand, the right, red plot has an unstable regression plane. To further illustrate the instability of the least squares estimates we draw a bootstrap sample of same size ($n = 20$) from the used data and we estimate the regression plane with a different colour. In the right figures, the two regression planes are coloured with red and green while in the left figures the regression planes are, nearly identical.

We recreate the same comparison between low an high collinearity with increased number of observations ($n = 300$) in Figures 2.4 and 2.5. We use the same scale factors f. Greater sample sizes are beneficial regarding the stability of the regression plane when collinearity is present. However, the problem still exists with smaller probability of occurrence.

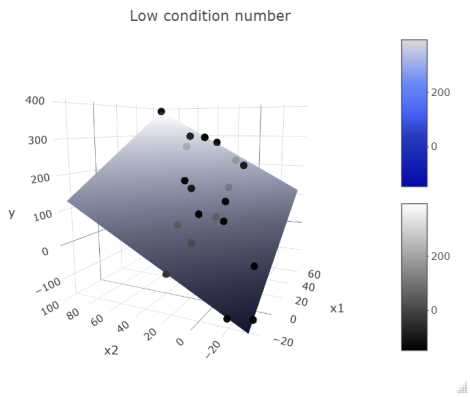

Figure 2.2: Stable estimation. (n=20)

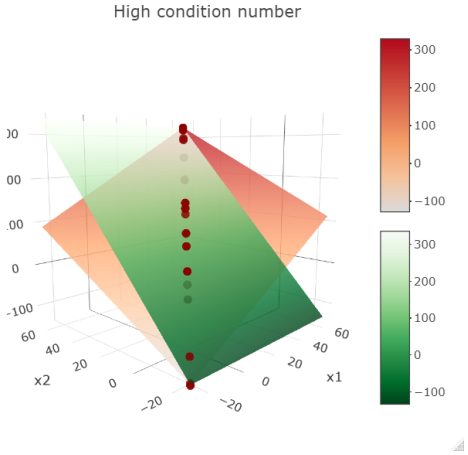

Figure 2.3: Unstable estimation. (n=20)

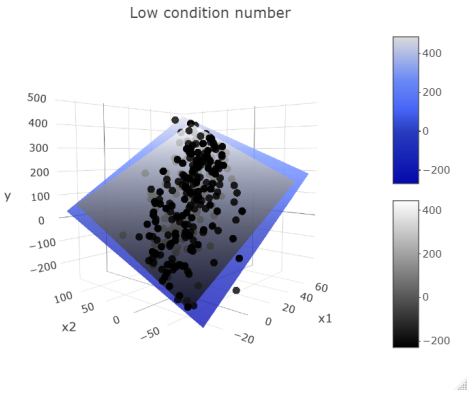

Figure 2.4: Stable estimation. (n=300)

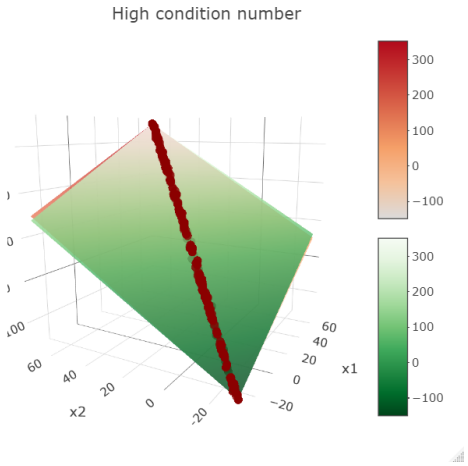

Figure 2.5: Unstable estimation. (n=300)

2.4 Collinearity and correlation

In the literature, one often finds the term correlation as a synonym for collinearity. Indeed this connection is partially true as when correlation is present, collinearity is present too. However, the existence of collinearity does not necessary imply correlation. Collinearity can exist when one of the predictors can be linearly described from the rest of the predictors. Thus, more than two predictors can be involved to the collinearity phenomenon. On the other hand, correlation can only provide inference for pairs of predictors. To motivate the difference, we consider the dataset in Table 2.1 with nine predictors. The first eight predictors (*pred 1* to *pred 8*), are numbers drawn from different uniform distributions. The ninth predictor (*pred 9*), is the sum of all eight previous predictors. A small error term ($\epsilon_{x_9} \sim N(0, 0.001)$) was added to the ninth predictor. Since the ninth predictor is a linear combination of the previous eight predictors, the predictor matrix is ill conditioned with severe collinearity. The response y is the summation of all nine predictors as in equation (2.17).

$$y = 2x_1 + 3x_2 + x_3 + 7x_4 + x_5/2 + 2x_6 + 2x_7 + 2x_8 + 2x_9 + \epsilon_y \tag{2.17}$$

Table 2.1: Uncorrelated data example.

y	pred 1	pred 2	pred 3	pred 4	pred 5	pred 6	pred 7	pred 8	pred 9
81.86	-7.03	64.43	-26.72	4.85	-86.33	-5.80	24.48	0.49	-31.61
198.43	38.87	79.73	-23.40	1.12	-96.20	-18.47	1.42	3.95	-12.98
113.34	-28.30	97.58	-40.96	3.35	-85.81	-12.92	17.35	8.26	-41.45
-142.16	-21.72	37.06	-39.33	1.81	-80.41	-10.53	3.58	1.85	-107.70
-66.32	-16.28	39.04	-47.79	3.91	-81.11	15.11	1.95	5.36	-79.79
280.19	25.94	79.03	-70.90	-0.08	-93.87	6.31	15.80	33.77	-4.01
256.16	36.97	88.91	-69.80	3.24	-95.84	-10.64	20.35	21.01	-5.80
175.95	-2.30	57.46	-18.65	0.77	-93.03	6.03	14.83	24.09	-10.79
251.75	27.83	89.97	-35.61	3.82	-86.56	7.62	2.82	0.32	10.21
185.83	30.47	63.45	-48.29	3.19	-88.04	-6.55	23.95	11.86	-9.96
212.61	-21.63	99.50	-40.14	4.63	-87.83	19.03	14.94	5.74	-5.76
136.23	30.49	35.81	-17.19	1.18	-81.86	-13.48	2.52	32.51	-10.04
100.23	-17.59	67.04	-44.19	-0.49	-90.46	-17.12	22.94	29.61	-50.25
223.01	-3.19	94.64	-14.15	2.42	-94.92	9.32	2.48	1.24	-2.16
27.00	5.94	48.97	-76.23	2.90	-86.14	4.20	3.63	29.58	-67.16
17.55	-18.36	31.48	-24.95	3.58	-93.96	12.34	32.60	8.09	-49.18
251.54	26.32	57.13	-13.70	1.81	-86.28	1.57	9.49	26.34	22.69
67.88	-6.67	73.64	-79.93	-0.85	-95.96	-11.79	33.73	8.81	-79.01
85.51	-3.46	37.05	-34.62	2.37	-99.91	5.81	28.99	25.51	-38.25
204.89	-13.97	94.06	-65.21	-0.52	-81.60	4.16	28.62	7.33	-27.13

Figure 2.6 illustrates all the pairwise correlation coefficients between the nine predictors. The largest correlation value is 0.62 between *pred 9* and *pred 2*. 0.62 is a relatively small correlation coefficient. Thus the exclusion of one of the two involved predictors because of correlation is a rather radical decision. Having no indication of correlation, one could ignorantly fit a classical linear model with all nine predictors included. The results of the linear regression model are summarised in Table 2.2. As it can be seen, both the least square estimates and their standard errors are extremely large. The unexpectedly large regression estimates are a direct consequence of collinearity. Moreover, none of the predictors p-values is significant.

We now take a look at a second regression model where the ninth, problematic predictor, *pred 9*, is excluded. Table 2.3 summarizes the results of this regression model. The estimated coefficients are very good estimations of the initially selected factors and the confidence intervals's width is reasonably close to the estimated value. It is clear that the model better describes the data. Tables 2.4–2.6 illustrate Belsley's suggested collinearity diagnostics approach which is described in section 2.5.3.

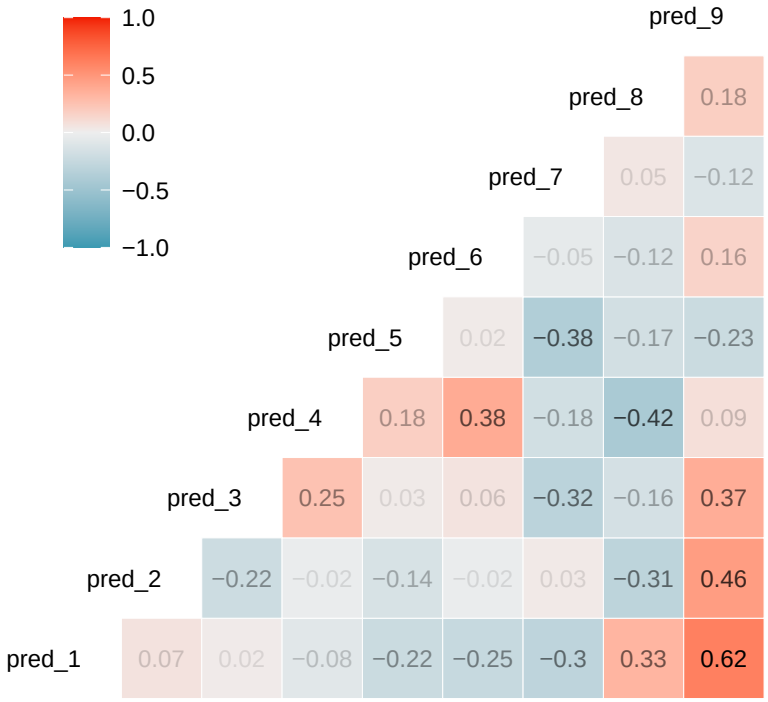

Figure 2.6: Pairwise Pearson correlation plot for the nine predictors.

Table 2.2: Regression output with nine predictors, high collinearity and no correlation between the predictors.

	Coefficient	95%-confidence interval	p-value
intercept	14.70	from -50.60 to 80.00	0.63
Pred 1	15.70	from -49.60 to 80.99	0.60
Pred 2	13.69	from -51.61 to 78.99	0.65
Pred 3	5.76	from -59.52 to 71.03	0.85
Pred 4	13.22	from -52.11 to 78.54	0.66
Pred 5	14.69	from -50.61 to 79.98	0.63
Pred 6	14.69	from -50.62 to 80.00	0.63
Pred 7	14.70	from -50.59 to 79.99	0.63
Pred 8	-11.69	from -76.99 to 53.60	0.70
Pred 9	1.50	from -4.16 to 7.16	0.57

Table 2.3: Regression putput with eight predictors, low collinearity and no correlation between the predictors.

	Coefficient	95%-confidence interval	p-value
intercept	3.00	from 2.99 to 3.02	< 0.0001
Pred 1	4.00	from 3.99 to 4.01	< 0.0001
Pred 2	2.00	from 1.98 to 2.01	< 0.0001
Pred 3	-5.93	from -6.11 to -5.76	< 0.0001
Pred 4	1.52	from 1.47 to 1.57	< 0.0001
Pred 5	2.99	from 2.97 to 3.02	< 0.0001
Pred 6	3.00	from 2.97 to 3.02	< 0.0001
Pred 7	3.01	from 2.98 to 3.04	< 0.0001
Pred 8	0.91	from -3.48 to 5.31	0.66

Table 2.4: Singular values μ_i and condition indexes η_i for the two regression models (Tables 2.2 and 2.3). The upper part of the table includes the values for the model with the nine predictors x and a constant. The second half of the table includes the values for the model with the eight predictors x and a constant.

Singular Values		Condition Indexes	
$\mu_1 =$	562.06	$\eta_1 =$	1
$\mu_2 =$	171.65	$\eta_2 =$	3.27
$\mu_3 =$	95.55	$\eta_3 =$	5.88
$\mu_4 =$	82.01	$\eta_4 =$	6.85
$\mu_5 =$	61.25	$\eta_5 =$	9.18
$\mu_6 =$	47.88	$\eta_6 =$	11.74
$\mu_7 =$	40.03	$\eta_7 =$	14.04
$\mu_8 =$	6.75	$\eta_8 =$	83.25
$\mu_9 =$	0.24	$\eta_9 =$	2376.46
$\mu_{10} =$	0.01	$\eta_{10} =$	100542.75
$\mu_1 =$	546.23	$\eta_1 =$	1
$\mu_2 =$	102.71	$\eta_2 =$	5.32
$\mu_3 =$	89.3	$\eta_3 =$	6.12
$\mu_4 =$	81.74	$\eta_4 =$	6.68
$\mu_5 =$	48.56	$\eta_5 =$	11.25
$\mu_6 =$	40.05	$\eta_6 =$	13.64
$\mu_7 =$	39.74	$\eta_7 =$	13.74
$\mu_8 =$	6.31	$\eta_8 =$	86.56
$\mu_9 =$	0.24	$\eta_9 =$	2309.91

Table 2.5: Condition indexes and variance decompositon proportions for the equilibrated design matrix X of the regression model with nine predictors x and a constant (Table 2.2).

Eq. cond. Index	V(xc)	V(x1)	V(x2)	V(x3)	V(x4)	V(x5)	V(x6)	V(x7)	V(x8)	V(x9)
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2.41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
58.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66
90188.73	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33

Table 2.6: Condition indexes and variance decomposition proportions for the equilibrated design matrix X of the regression model with eight predictors x and a constant (Table 2.3).

Eq. cond. Index	V(xc)	V(x1)	V(x2)	V(x3)	V(x4)	V(x5)	V(x6)	V(x7)	V(x8)
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2.07	0.19	0.00	0.00	0.01	0.00	0.26	0.00	0.01	0.00
2.62	0.32	0.00	0.00	0.02	0.00	0.34	0.02	0.00	0.00
3.43	0.00	0.00	0.01	0.19	0.00	0.21	0.02	0.19	0.00
4.88	0.26	0.01	0.04	0.11	0.00	0.10	0.23	0.34	0.00
5.47	0.05	0.07	0.14	0.17	0.00	0.01	0.47	0.02	0.00
7.05	0.00	0.13	0.74	0.21	0.00	0.03	0.01	0.00	0.00
12.13	0.04	0.77	0.01	0.24	0.02	0.02	0.03	0.44	0.02
64.11	0.14	0.01	0.06	0.04	0.98	0.03	0.23	0.00	0.98

2.5 Methods to quantify collinearity

2.5.1 Design matrix X

Collinearity is a phenomenon created only by the predictors of a model. Therefore, all the procedures aiming to identify collinearity are based on the design matrix X. The design matrix X is a matrix with as many columns as the number of predictors and as many rows as the number of the observations. Note, that if a constant is included in a model formula, one column of the design matrix X must represent the constant. Whereas some methods to diagnose collinearity use the raw design matrix X, others use a transformed design matrix or a product of the design matrix such as the X^TX.

2.5.2 Collinearity diagnostics based on X^TX

We first consider Montgomery et al. [2021] approach on collinearity diagnostics using X^TX matrix and its eigenvalues.

Condition number

The condition number of matrix (X^TX) is defined as

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \tag{2.18}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the roots of the the equation $(X^TX - \lambda I)v = 0$ with eigenvectors $v_1, v_2, \ldots, v_p$. If a linear association exists among predictors, it causes at least one of the eigenvalues (λ_{min}) to be very small (close to zero). Thus, the condition number κ will be high. Therefore, the condition number κ provides a relative measure for the distance of X from the exact collinearity.

The procedure of condition number calculation is

1. Define the design matrix X of the predictors for the given model
2. Create the matrix (X^TX)
3. Calculate the eigenvalues and the corresponding eigenvectors of the matrix (X^TX)
4. Calculate κ as in (2.18)

The condition number of X is related to the sensitivity of the diagonal elements of $(X^TX)^{-1}$ (and therefore the least square estimates). It meaningfully measures the difficulty of inverting a matrix.

Apart from the existence of collinearity and the level of severity of it, eigenvectors can be used to identify which variables are involved in the collinear relations. The eigenvectors define the magnitude each dimension has in an eigenvalue (the dimensions are defined by the number of predictors). By looking at the eigenvectors of the smallest eigenvalue, we can identify the impact of each predictor to collinearity. Moreover, one can understand high dimensional (≥ 3 predictors) patterns of collinear predictors.

2.5.3 Collinearity diagnostics based on X

Singular value decomposition

An alternative method proposed by Belsley [1991] uses the singular value decomposition of the design matrix X.

$$X = UDV^T \tag{2.19}$$

$$U^TU = V^TV = I \tag{2.20}$$

where D is a diagonal with non-negative diagonal elements and $\mu_1, \ldots, \mu_p$ are the singular values of X. Each column of the design matrix X has a unique singular value μ. U is a column orthogonal matrix with dimensions equal to the design matrix X. V is row and column orthogonal matrix with both dimensions equal to the number of predictors. Finally, D is non-negative and diagonal matrix with both dimensions equal to the number of predictors. Note the relation between the eigenvalues based on X^TX and singular values decomposition the condition indexes of design matrix X is equal to $\lambda_i = \mu_i^2$.

Moreover, Belsley [1991] defines condition indexes

$$\eta_i = \frac{\mu_{max}}{\mu_i},\ i = 1, \ldots, p. \tag{2.21}$$

Therefore, the maximal condition index is

$$\eta_{max} = \frac{\mu_{max}}{\mu_{min}} = \sqrt{\kappa} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{2.22}$$

Variance decomposition proportions

Belsley [1991] suggests variance decomposition proportions. In the case of usual assumptions, variance-covariance matrix $V(b)$ of the least squares estimator $b = (X^TX)^{-1}X^Ty$ is $\sigma^2(X^TX)^{-1}$ where σ^2 is the common variance of the components of ϵ in the linear model $y = X\beta + \epsilon$. Using the singular value decomposition, $X = UDV^T$, the variance covariance matrix is

$$V(b) = \sigma^2(X^TX)^{-1} = \sigma^2 VD^{-2}V^T \tag{2.23}$$

The variance of the kth regression coefficient, b_k, is

$$\text{var}(b_k) = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2} \tag{2.24}$$

where μ_j's are the singular values of X and $V \equiv (v_{ij})$

We define the (k, j)th variance-decomposition proportion as the proportion of the variance of the kth regression coefficient associated with the jth component of its decomposition in equation (2.24). Those proportions are calculated as follows:

$$\phi_{kj} \equiv \frac{v_{kj}^2}{\mu_j^2} \text{ and } \phi_k \equiv \sum_{j=1}^{p} \phi_{kj},\ k = 1, \ldots, p. \tag{2.25}$$

Then, the variance-decomposition proportions are

$$\pi_{jk} \equiv \frac{\phi_{kj}}{\phi_k},\ k, j = 1, \ldots, p. \tag{2.26}$$

Each row corresponds to a singular value μ_j and the associated condition index $\eta_j \equiv \mu_{max}/\mu_j$.

Equilibration

Belsley [1991] proposes to transform the design matrix X by equilibration. Equilibration allows for condition numbers from different design matrixes to be comparable. The suggested transformation is to scale each column to have equal length - column equilibration. That is $X = [X_1 \ldots X_p], s_i \equiv (X_i^TX_i)^{-1/2}$, and $S \equiv \text{diag}(s_1, \ldots, s_p)$ then the equilibrated design matrix is (XS). The equilibrated design matrix (XS) has columns of equal Euclidean length equal to one.

Condition indexes for equilibrated design matrix

We use the maximal equilibrated condition index which is the square root of the condition number of the equilibrated design matrix. Similarly with the equilibrated condition index, the equilibrated condition indexes of an equilibrated design matrix X are defined as

$$\tilde{\eta}_i = \eta_i(XS),\ i = 1, \ldots, p \tag{2.27}$$

Belsley [1991] suggests a threshold of 30 for condition indexes based on equilibrated design matrix. If the value is larger than 30 it indicates collinearity that we should investigate in greater detail.

2.5.4 Scaling and centering

Scaling the design matrix X ensures equal magnitude changes among the predictors. However, such transformations do affect the numerical properties of the data matrix and lead to different variance decomposition proportions and condition indexes. Nevertheless, the presence of exact linear dependencies among the columns of X is not affected by scaling. That is because for any non-singular matrix B there exist a nonzero c such that $Xc = 0$ if and only if $[XB][B^{-1}c] \equiv \bar{X}\bar{c} = 0$ where $\bar{X} = XB$ and $\bar{c} = B^{-1}c$ [Belsley, 1991, p.171].

Centering produces collinearity diagnostics that are often distorted and misleading. Near linear relations among the predictors due to the intercept are not to be ignored. Collinearity diagnostics of mean centered data ignore the effects of ill conditioning because of the missing intercept and provide diagnostic information relevant to a different problem. Although the constant term can not be correlated with any of the predictors, it can be involved in high dimensional collinearity patterns.

Table 2.7: Models considered and their notation

Notation	Model
(β_0)	$y \sim \beta_0$
(β_0,β_1)	$y \sim \beta_0 + \beta_1 x_1$
(β_0,β_2)	$y \sim \beta_0 + \beta_2 x_2$
$(\beta_0,\beta_1,\beta_2)$	$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2$
$(\beta_0,\beta_1,\beta_2,\beta_3)$	$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Table 2.8: Prior parameters for INLA.

Coefficient	Mean	Precision
beta0	0	0.01
beta1	0	0.01
beta2	0	0.01
beta3	0	0.01

2.6 Simulation

Simulation protocol

The aim of the simulation is to identify and illustrate the impact of collinearity in the least square estimations and the model choice criteria. We primarily focus on severe collinearity scenarios where the least squares estimates collapse. We use different methods to quantify collinearity. The existing levels of collinearity in a design matrix were described with the scale factor f and the condition number from Montgomery et al. [2021] described in section 2.5.3. Different combinations of a set of parameters such as the sample size and the used predictors were investigated. We consider two sample sizes of $n = 20$ and $n = 300$.

Models considered

We considered five different regression models as in Table 2.7. The first model (β_0) is described by a constant term only and it is the simplest model that can be fit to the data. The second and the third model, (β_0, β_1) and (β_0, β_2) respectively, have two coefficients each. The former is the intercept β_0 and the second one is a coefficient for a predictor (x_1 or x_2). The fourth model, $(\beta_0, \beta_1, \beta_2)$, has an intercept and two predictor variables. Finally, the fifth model, $(\beta_0, \beta_1, \beta_2, \beta_3)$, has an intercept and three predictors. The third predictor, which corresponds to coefficient β_3 is the product of the first two predictors. The prior parameters used for the Bayesian regression are presented in Table 2.8.

Random number generator

We use only one set of generated data set for each sample size ($n = 20$, $n = 300$). The predictors x, the response y and the error terms ϵ remain the same for all the simulations. We generate one predictor x_2 from the other x_1 using a uniform distribution. The errors of the predictors were generated from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.5$.

Method of data generation

For the generation of different cases of collinearity we use the equation (2.28). We create a response variable y from an intercept β_0, two predictors x_1 and x_2 and a noise factor ϵ_y.

Table 2.9: Estimates obtained from the simulations.

	Classical	Bayesian
Estimates	$(\hat{\beta}_0)$ $(\hat{\beta}_0,\hat{\beta}_1)$ $(\hat{\beta}_0,\hat{\beta}_2)$ $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2)$ $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$ SE σ^2	
Model choice criteria	AIC BIC	DIC WAIC LCPO LML

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_y \tag{2.28}$$
$$x_2 = \gamma_0 + \gamma_1 x_1 + \epsilon_x \mathrm{f} \tag{2.29}$$
$$\epsilon_y \sim N(0, \sigma_\epsilon^2) \tag{2.30}$$
$$\epsilon_x \sim N(0, \sigma_x^2) \tag{2.31}$$
$$\tag{2.32}$$

We construct the two regressors so that they are collinear. Here, we use $\gamma_0 = 0$ and $\gamma_1 = 1$ as the simplest scenario. The scale factor f is a scaling factor which is used to create different levels of collinearity. The closer the scale factor f is to zero, the greater the level of collinearity. Contrariwise, if the scale factor f is greater than one, the amount of collinearity is decreased.

Regressor x_1 is randomly sampled from a uniform distribution $x_1 \sim \mathcal{U}_{-30,70}$. The noise of the regressors is also randomly generated $\epsilon_x \sim \mathcal{N}_{0,0.5}$. The second regressor, x_2 is created by adding noise to x_1. We repeat the procedure after scaling the noise values with the scale factor f. For the scaling, a sequence of positive factors was used (Table 3.1).

Estimates obtained

Five different linear regression models fits were considered with the predictors being x_1 and x_2. Both a classical and a Bayesian linear regression were applied. For the later, we used the `INLA` package. For models fits in the classical framework, we obtained the coefficient estimates, the standard errors, the AIC and the BIC values (Table 2.9). Similarly, for models fits in Bayesian framework, we obtained the median of the posterior distributions for the parameters, the DIC, the WAIC, the LCPO and the LML.

Number of simulations

For the simulation for the single scenario model choice criteria, 60 replications were required for each different value of the scale factor f. This procedure was done both in frequentist and Bayesian framework thus the total number of simulations was increase to 120.

Regarding the instability estimates, an increase number of simulation was required. 100 bootstrap draws were taken from the initial generated values. Each draw was scales with the 60 values of the

scale factor f. Finally, the procedure was repeated in frequentist and Bayesian framework. The total number of iterations for the instability estimates of the five considered models was 1200.

2.6.1 Bootstrap replications

Bootstrap is a well-known technique to check stability of results. [Heinze et al., 2018, Pavone et al., 2020] In order to investigate the instability of the regression estimates and the model choice criteria for small and large sample size, we repeat the simulation using 100 bootstrap replications. For the bootstrap draws, we use the initially generated data set and we sample with replacement. For each bootstrap draw, we scale the errors between the two predictors x_1 and x_2 with the scale factor f. For each scaling scenario, we fit a classical linear regression model and we extract the estimated coefficients and the model choice criteria AIC and BIC. Moreover, we fit a Bayesian linear regression model using `INLA` and we extract the posterior median and the model choice criteria DIC, WAIC, LCPO and LML. We repeat the procedure for each of the five model fits considered (Table 2.7). When the simulation is finished, we have 100 estimates for each scaling scenario.

To illustrate the instability of regression estimates for collinear data we illustrate the change of the estimates for increasing scale factors f for all 100 bootstrap replications. To compare the classical with the Bayesian framework, we provide similar figures for the Bayesian regression estimates. We continue with the model choice criteria. To demonstrate the instability we use three methods of illustration. First, by printing the model choice criterion value for each scale factor f and for each bootstrap replication. Secondly, we calculate the median value of the distribution of the 100 model choice criteria estimates for each scaling scenario. Moreover, we estimate the 95 percent bootstrap confidence intervals. We illustrate the median values along with their confidence intervals for all models considered together. For the third method of illustration, we used Gaussian kernels to estimate the empirical densities of the model choice criteria for different scale factors f.

2.6.2 Pseudo logarithmic scale

Many of the presented plot use the pseudo logarithmic transformation for the transformation of the y axis. This transformation is a logarithmic transformation with base 2.71 for cases of $x > 1$. For cases were predictor x is close to zero, a smoothly transaction to the linear scale is used.

2.7 Case study: Bodyfat

We use the example of Bodyfat to illustrate the relationship of collinearity with model choice criteria in real data. We use the already modified dataset from Pavone et al. [2020], which is available on: `https://raw.githubusercontent.com/fpavone/ref-approach-paper/master/code/bodyfat.txt`. We added a new constant variable with repeated values of one.

The variable of interest, siri (Percent body fat using Siri's equation: 495/Density - 450), is defined as the amount of body fat, which is obtained by a complex and expensive procedure consisting in immersing a person in a water tank and carrying out different measurements and computations. Moreover, we obtain information for 13 variables which are anthropometric measurements. The body fat dataset contains measurements for 251 men. Because of the nature of the variables, we expect collinearity patterns among them. The source of data is Johnson [1996] and they are also reviewed by Heinze et al. [2018] and Pavone et al. [2020].

Discriptive statistics

We start by providing an overview of the Bodyfat dataset with an explanatory data analysis. We estimate the mean, standard deviation, skew, median, minimum and maximum values of each variable and

the response siri. For the variables **siri**, **abdomen**, **height_cm**, **weight_kg** and **hip** we additionally provide pair plots along with their Pairwise pearson correlation coefficients and their significance levels. Moreover, we estimate all the Pearson and Spearman correlation coefficients between all pairs of variables. Finally, we provide the boxplots of all the predictors of the dataset together.

Collinearity diagnostics

We consider all the possible model combinations from the Bodyfat data set. We quantify the collinearity between the models using the maximal equilibrated condition index. First, we estimate the equilibrated condition indexes and the variance decomposition proportions for models including the variables **abdomen**, **height_cm**, **weight_kg**, **hip** and **constant**. The distributions of the maximal equilibrated condition indexes for models with different number of predictors were illustrated as empirical densities. We used Gaussian kernel densities to illustrate the empirical densities of the maximal equilibrated condition indexes. For comparison, we scale and center the original dataset and we repeat the procedure.

Instability of estimates

We continue the analysis providing stability measurements for the regression estimates in classical and Bayesian framework while also for the stability of the model choice criteria. For this part of the analysis we only use eight models with **abdomen**, **height_cm**, **weight_kg**, **hip** and **constant** as predictors. We consider the predictors **weight_kg** and **hip** to be collinear and we use them to govern the severity of collinearity. To do that, we fit a linear regression model with variable **hip** being the response and **weight_kg** being the predictor. We extract the residuals of the predicted model and the scale factor f to modify the collinearity between the two predictors. Specifically, we first multiply the residuals with a scale factor f and subsequently, we add the scaled residuals to the expected values of the model fit. The spectrum of the scale factor f is 60 values from 0.000 015 6 to three. The new, scaled values of **weight_kg** predictor are saved in a temporary Bodyfat dataset and are used for the model fits. From each model, we extract the estimated coefficients and the model choice criteria.

For the next step we create 100 bootstrap samples. We sample for $n = 251$ with replacement from the Bodyfat data set. For each drawn sample, we follow the steps of scaling the residuals of **weight_kg** (predicted from **hip**) as described above and we fit the eight regression models for each different scale factor f. We end up with 100 estimates of regression coefficients from each model for each scale factor f. To illustrate the instability of the regression estimates for different cases of collinearity, we estimate the standard deviation of the 100 obtained estimates for each scale factor f. We plot the escalation of the standard deviation along the different scale factors for all the models that include **hip** as a predictor.

Apart from the regression estimates, the model choice criteria estimates for each model were also obtained. We illustrate the change of each model choice criterion for increasing scale factors f.

Model choice criteria

We fit a classical linear regression model and and Bayesian linear regression model for each possible model fit in the bodyfat dataset. We use `INLA` to fit the Bayesian regression models. For the Bayesian model fits, we used normal priors with zero mean ($\mu = 0$) and precision $\tau = 0.01$ (Table 4.7). We estimated the AIC, BIC, DIC, LML, LCPO and WAIC model choice criteria, the former two for classical regression models whereas the later four for the Bayesian model fits.

We investigated the relationship between model choice criteria and collinearity. The model choice criteria values of each model were ploted against their maximal equilibrated condition indexes. First we focus on the AIC plot only for the models with **abdomen**, **height_cm**, **weight_kg**, **hip** and

constant (or any subset of those) as predictors. At a second step, we provide plots including all the Bodyfat models and for all the model choice criteria. Models with different number of predictors are coloured differently, with lighter colours indicating more predictors. To model the relationship between model choice criteria and collinearity, we fit a simple regression model with AIC being the response and the maximal equilibrated condition index being the predictor. We categorize the models according to their number of predictors and we conduct a stratified analysis by fitting a linear model for each category. Additionally, we estimate the Spearman correlation coefficients and their level of significance for each category. The stratified linear model estimates, the Spearman correlation coefficients and their p-values are provided in tables.

Finally, we investigate the impact of each predictor to the model choice criteria evaluation. We construct plots of AIC model choice criterion per predictor. To do that, we separate all the models that have a specific predictor variable and we illustrate their AIC value along with their equilibrated condition index. The procedure is repeated for all the predictors. Because the predictor **abdomen** was found to be the most important regarding the AIC evaluation, we repeat the procedure without this predictor. For that, all models including **abdomen** as a predictor are removed and the procedure is repeated.

Link with existing literature

We analyse the Bodyfat data set in a similar way as Pavone et al. [2020] and Heinze et al. [2018] did. The intial data set is identical with the one Heinze et al. [2018] used. However, Heinze et al. [2018] defines a baseline model with the predictos **abdomen** and **height_cm**. We modify the initial Bodyfat dataset as in Pavone et al. [2020] by scaling and centering all the predictors of the dataset. The root-mean-square $\sqrt{\sum \frac{x^2}{n-1}}$ scaling method is used after the centering on the data. Only observations with positive response were kept. We estimate the stratified Spearman correlation coefficients and their level of significance for the relationship between the maximal equilibrated condition index and AIC.

2.8 R packages

2.8.1 New package

The package `Collinearity` which is available in `https://github.com/G-Kazantzidis/Collinearity.git` was developed to provide tools of collinearity diagnostics as described in Section 2.5.3. This package implements collinearity diagnostics methods suggested from Belsley [1991] and described in Section 2.5. Function `equilibrate_matrix` equilibrates the design matrix X to unit length. Function `vdp_svd` uses singular values decomposition to compute variance-decomposition proportions for a design matrix X. Generic function `Var_decom_mat` which supports both classes `matrix` and linear model objects (`lm`), provides the variance decomposition matrix along with condition indices. The user can also apply equilibration prior to the calculation of the variance decomposition proportions. This package was used for this master thesis project which is available on: `https://bitbucket.org/G_Kaza/master-thesis-2022/src/master/`.

2.8.2 Used

All analyses were performed in the R programming language using base packages. Specifically, for the analysis the packages used: `plot3D` to create 3D plots [Soetaert, 2021]. `ggplot2` for the production of all the figures [Wickham, 2016]. `RColorBrewer` for specific colour palettes for the figures [Neuwirth and Brewer, 2014]. `tableone` for the production of high quality summary tables [Yoshida and Bartel, 2020]. `xtable` for the construction of tables [Dahl et al., 2019]. `biostatUZH` for the construction of tables with regression results and convenient formatting of p-values and confidence intervals [Haile

et al., 2020]. `tidyverse` for efficient organization and manipulation of the used datasets [Wickham et al., 2019]. `INLA` for fitting the models in Bayesian framework. `scales` for including scientific-type scales in the created figures [Wickham and Seidel, 2020]. `gridExtra` for including grids inside the figures [Auguie, 2017]. `ggpubr` for a scientific appearance of the created figures [Kassambara, 2020]. `plotly` for the creation of interactive plots [Sievert, 2020]. `latex2exp` for the use of greek characters in ggplot objects [Meschiari, 2021]. `reshape2` for changing the format of the datasets [Wickham, 2007]. `ggridges` for creating preferable plots with ggplot2 [Wilke, 2021]. `viridis` for colour palettes used in the figures [Garnier, 2018]. Finally, `hrbrthemes` [Rudis, 2020].

Chapter 3

Simulation

3.1 Simulation scenarios

Figure 3.1 compiles information related to the different sets of parameters considered in the simulation. In general the same set of parameters were used both in classical and Bayesian framework apart from the different model choice criteria and the additional priors used for the Bayesian regression. To investigate the effect of the sample size on collinearity, two different sample sizes were considered, n = 20 and n = 300.

The unique parameter set of predictors used for this project is $(\beta_0, \beta_1, \beta_2, \beta_3)$. The predictors leading to those coefficients were generated once (for each sample size) and were used unchanged for the rest of the simulation.

Finally, in order to govern the amount of collinearity, we used 60 values between zero and one as the scale factor f. For the generations of the values of the scale factor, we first create sequence of equal steps beginning from 0.25 and until 1. To increase the accuracy of our estimates for cases of severe collinearity, we take the cube of the values of this sequence.

Table 3.1 provides the values of the parameters set for the simulations. The sample size is set to 20 and 300 observations. The coefficient for the intercept β_0 is set to one. The coefficients $\beta_1 = 3$ and $\beta_2 = 2$ for the predictors x_1 and x_2 respectively. The standard deviation $\sigma_y = 0.5$ and the scaling factor f for the errors covers the spectrum from 0 to 1.

Table 3.1: Parameters of the simulation.

	Parameter
Sample size	20, 300
β_0	1
β_1	3
β_2	2
γ_0	0
γ_1	1
σ_y	0.5
σ_x	0.5
Scaling factor f	from 0 to 1
Length of grid for f	60

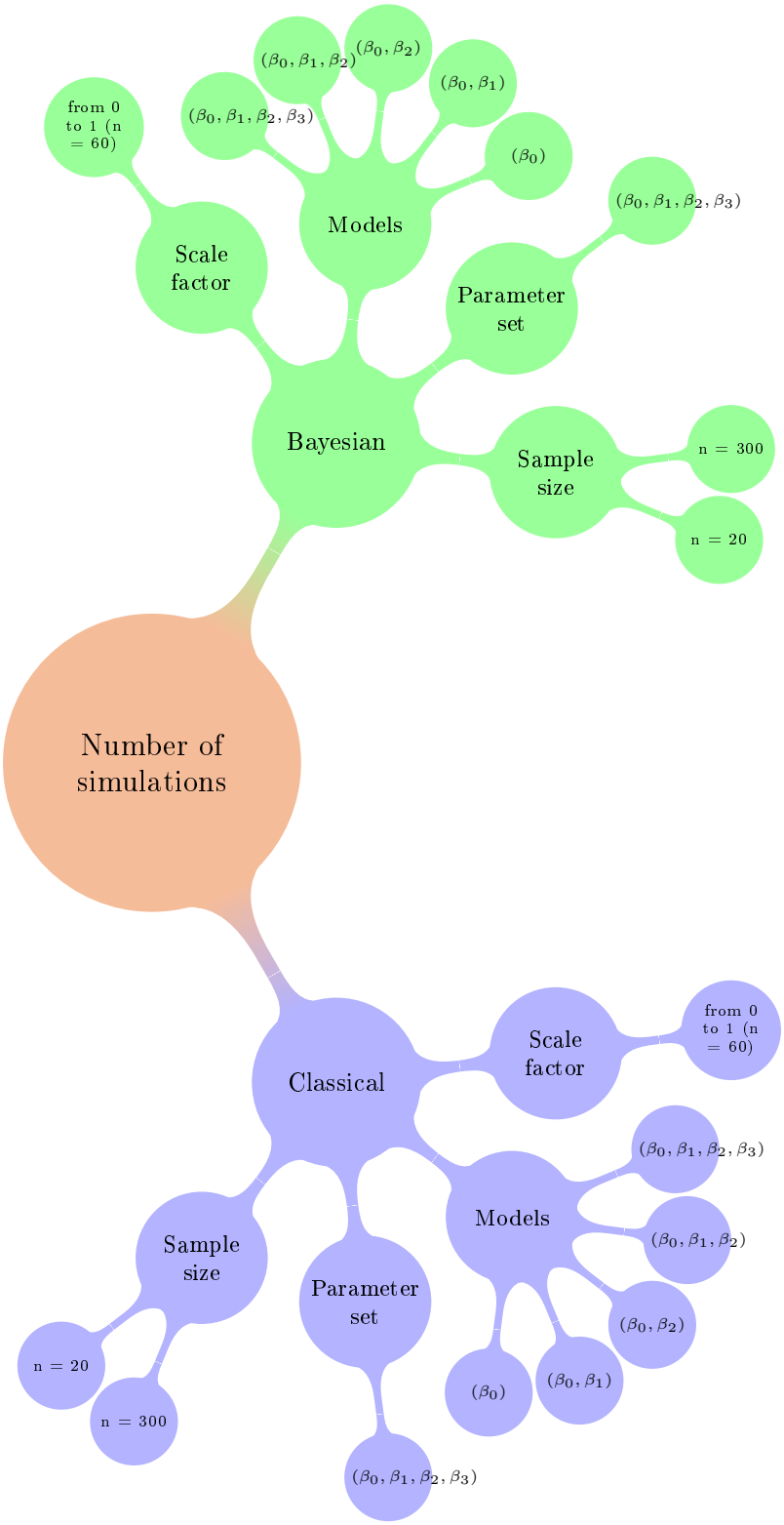

Figure 3.1: Factors included in the simulations.

Condition number

Figure 3.2 illustrates the relationship between the scale factor f, the corresponding condition number κ, condition index η_{max} and the equilibrated condition index $\tilde{\eta}_{max}$ for the model $(\beta_0, \beta_1, \beta_2)$.

In general, it can easily be seen that very small scale factors results to very high collinearity levels. When the scale factor becomes larger, the changes of the condition number and index become smaller. More specific, when the scale factor of the errors drops below 0.25, κ and η_{max} grow exponentially. In the plot, we witness a skyrocket increase of κ from 0.5×10^6 to 1.5×10^{10}. Finally, there is small difference between η_{max} and $\tilde{\eta}_{max}$ (blue and green lines match well), indicating that equilibration does not affect the condition indexes η_i.

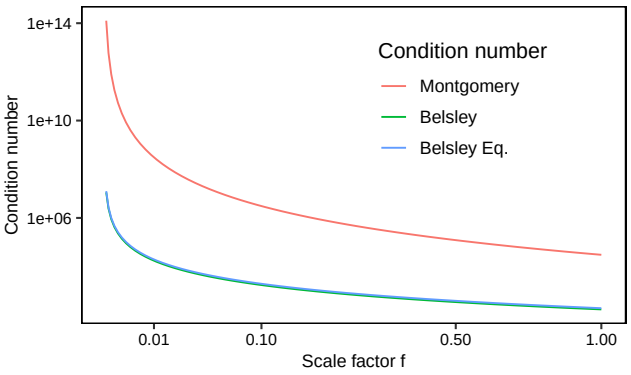

Figure 3.2: Relationship between the scale factor and the condition number. The x axis is the scale factor f (square root scale). The y axis is the condition number (logarithmic scale). The red line represents the condition number based on X^TX from Montgomery. The green line is the maximal (non-equilibrated) condition index of the design matrix X based on Belsley and the blue line represents the equilibrated condition index.

Coefficient estimates

For each different scale factor f, we fit all five regression models considered. From each model, we extracted the estimates from all the coefficients of the used model. We investigate the estimated coefficients under different cases of collinearity and for the different models considered. To improve our understanding in the impact of collinearity and the least squares estimates, we illustrate the estimated coefficients along both the scale factor f and the condition number of the model.

3.2 Instability of estimates

3.2.1 Small sample size

In order to measure the variability of the least square estimates, while also that of the model choice criteria, we contact a simulation study. We use the set of generated observations with sample size $n = 20$. We draw a bootstrap sample with same size as the original set of errors $n = 20$ and with replacement. We create a sequence of 60 scale factors from 0.0001 to 1 and we use them to scale the bootstrap sample. We use the scaled errors and the original predictor x_1 to create a second predictor variable x_2. We fit four linear models (β_0, β_1), (β_0, β_2), $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$ to all scaled versions of the same data set and we obtain the coefficient estimates and the AIC value. We replicate the whole procedure 100 times.

From the table of the outputs, we calculate the 2.5%, 50% and 97.5% percentiles of the 100 obtained estimates for each scale factor to create the 95% bootstrap confidence interval area. The raw outputs from the simulation are presented as lines in Figure 3.3 for the estimates and Figure 3.8 for the AIC. The mean estimates and their confidence interval area for the AIC are presented in Figure 3.9.

We use the Standard deviation to illustrate the instability of the least squares estimates. That is, the standard deviation of the estimates from the 100 bootstrap replications.

Figure 3.3 illustrates the instability of the estimated coefficient β_1 from the two models $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$ in association with the scale factor f for 100 bootstrap replications. The rest three

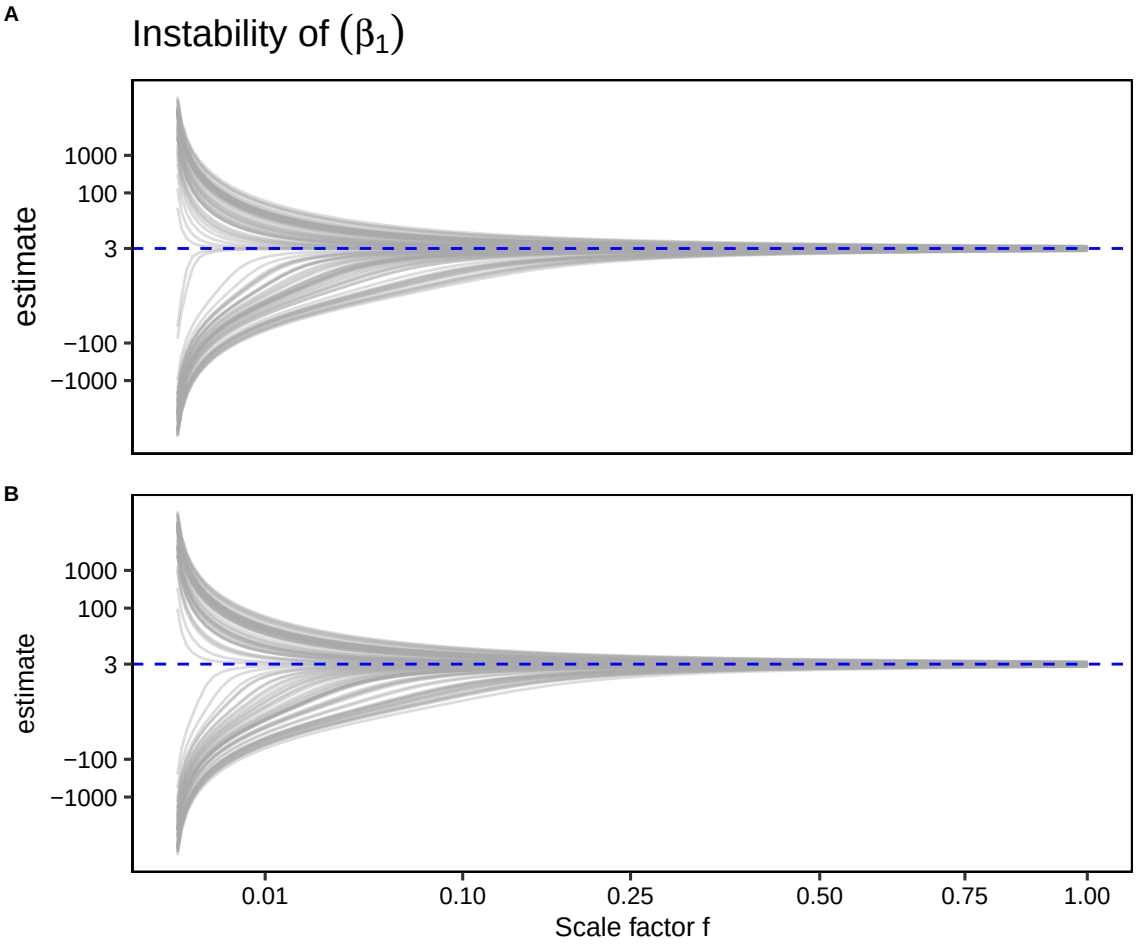

Figure 3.3: Instability of β_1 estimate for severe cases of collinearity. The horizontal axis indicates the values of the scale factor f in the square root scale. The vertical axis indicates the least square estimates in the pseudo logarithmic scale. The top plot (A) for the $(\beta_0, \beta_1, \beta_2)$ model and plot (B) is for the $(\beta_0, \beta_1, \beta_2, \beta_3)$ model. Each plot has 100 scenarios (gray lines).

models considered had stable regression coefficients for all different scale factors f. In general, the estimation of the coefficient is extremely unstable for small scale factors f and stabilizes as the scale factor f increases.

Specifically for scale factors very close to zero, the estimations of β_1 fluctuate between $-10\,000$ and $10\,000$ whereas the true value of the parameter is 3. For scale factors from zero to 0.2, the variability of the estimates gradually decreases. Both plots indicate a stable estimation for scale factors f greater than 0.25. Furthermore, whether the estimated coefficient for low scale factors f is close to $-\infty$ or ∞ is a matter of chance.

Figure 3.4, similar to Figure 3.3, illustrates the Bayesian regression estimates obtained from `INLA` for different cases of collinearity. The figure compiles information for models $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$. In agreement with the frequentist case, we observe increased instability of the estimates for scale factors close to 0.1. The instability of the estimated coefficient drops while the scale factor f approaches one. Comparing the two figures (for the Bayesian and frequentist approach), we realize that the amount of instability is considerably less in the case of the Bayesian framework. For the $(\beta_0, \beta_1, \beta_2)$ model, the largest observed estimate is 3.6946×10^4 whereas for the `INLA` estimation the largest observed value is 11. Likewise, the values of the largest estimates for β_1 in the $(\beta_0, \beta_1, \beta_2, \beta_3)$ model is 3.6791×10^4 for the least squares case in the frequentist format and 11 for the `INLA` estimation.

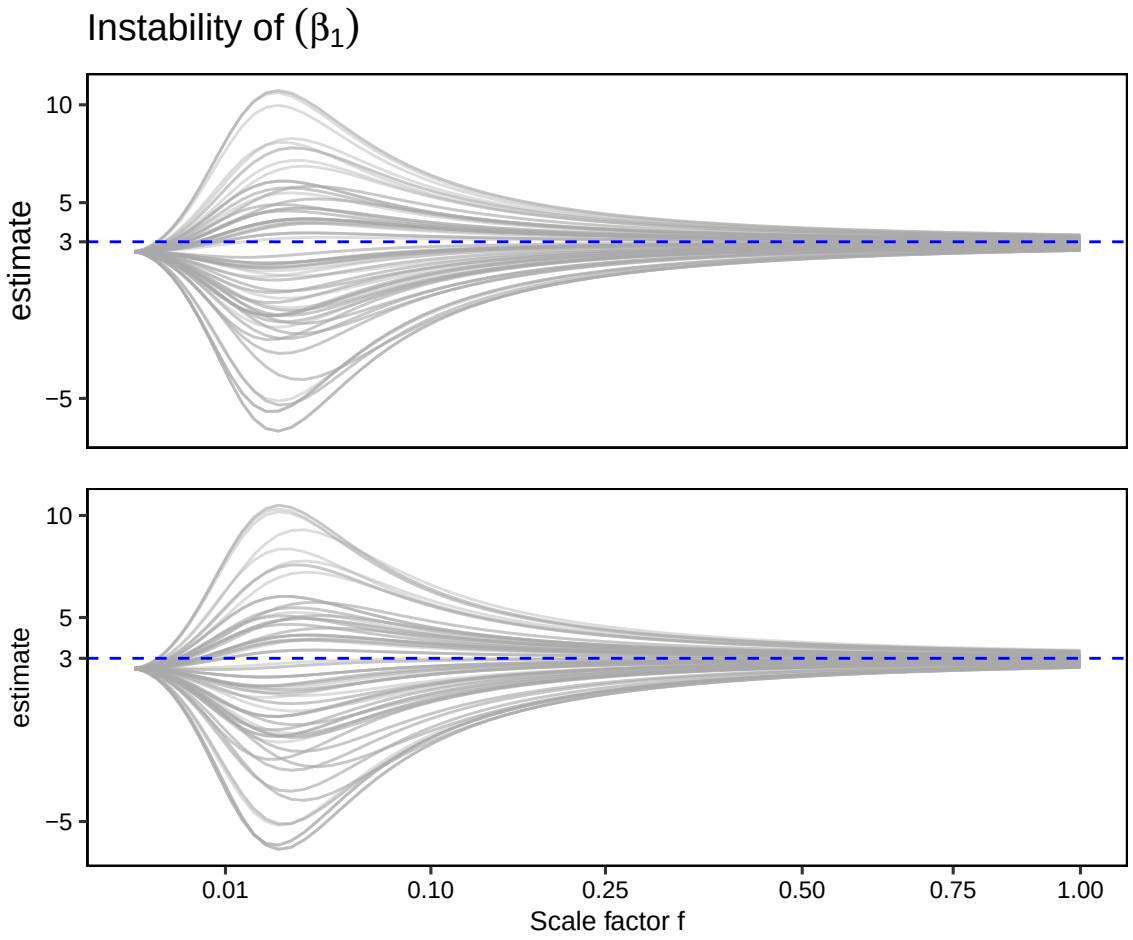

Figure 3.4: Instability of β_1 `INLA` estimate for different cases of collinearity. The horizontal axis indicates the values of the scale factor f in the square root scale. The vertical axis indicates the `INLA` regression estimates. The top plot for the $(\beta_0, \beta_1, \beta_2)$ model and plot is for the $(\beta_0, \beta_1, \beta_2, \beta_3)$ model. Each plot has 100 scenarios (gray lines).

3.2.2 Large sample size

We repeat the procedure as in section 3.2.1 with a large sample size $n = 300$.

Figure 3.5 depicts the instability of the β_1 estimates according to the scale factor f and for a sample size of n = 300. The two included plots represent the models $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$. The coefficient estimates remain stable for the scale factors f greater than 0.25. Increased instability is observed at the left side of the plots where the scale factor is close to zero.

Figure 3.6 is similar to Figure 3.5 but for the Bayesian framework. Both frameworks have unstable estimates for scale factors f less than 0.1. We see a rapid increase of the estimated coefficients immediately after scale factor f = 0 which peaks for an approximate scale factor 0.05. After that, the estimated coefficients begin to decrease slowly until they converge to the true value of the parameter $\beta_1 = 3$. The significant difference between the figures is the amount of instability that is observed. `INLA` estimates are more stable. The difference between the frequentist and Bayesian case (regarding the instability of estimates) is smaller in the case of large sample size.

Figure 3.7 compiles the estimates of the variation coefficient for the $(\beta_0, \beta_1, \beta_2)$ model and for increasing values of scale factor f. There are four differently coloured lines, one for each model. The lines represent the standard deviation as a measure of the instability of the regression estimates.

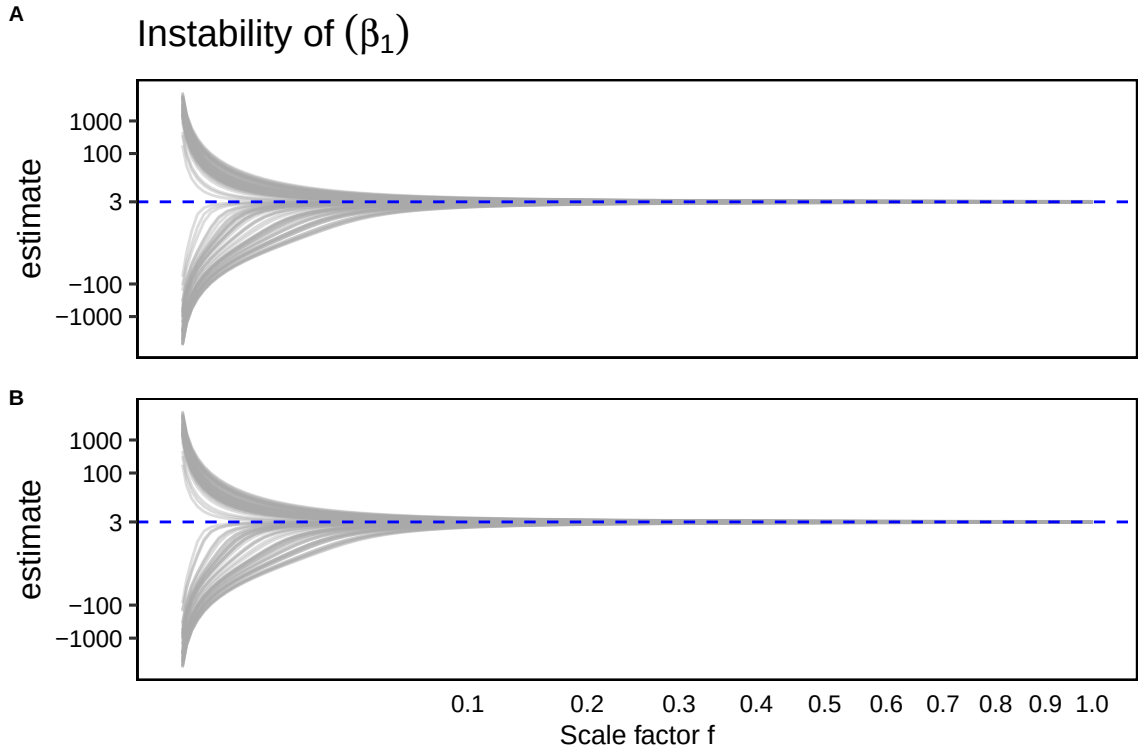

Figure 3.5: Instability of β_1 estimate for severe cases of collinearity. The horizontal axis indicates the values of the scale factor f in the square root scale. The vertical axis indicates the least square estimate in the pseudo logarithmic transformation scale. The top plot (A) for the $(\beta_0, \beta_1, \beta_2)$ model and plot (B) is for the $(\beta_0, \beta_1, \beta_2, \beta_3)$ model. Each plot has 100 scenarios (gray lines). The sample size is 300.

Overall, it can be seen that greater collinearity leads to greater standard deviation of the estimates.

More specific, for frwquentist models, the maximal standard deviation observed is more than 1000 for the small sample size and slightly less for the large sample size. In the Bayesian case, the maximal standard deviation for the large sample is less than 10, which is approximately 100 times smaller compared to the frequentist case. Both cases of sample size appear to have an exponential decrease until the approximate zero for scale factor f $= 1$ (note the square root transformed axis x and the pseudo-log transformed axis y).

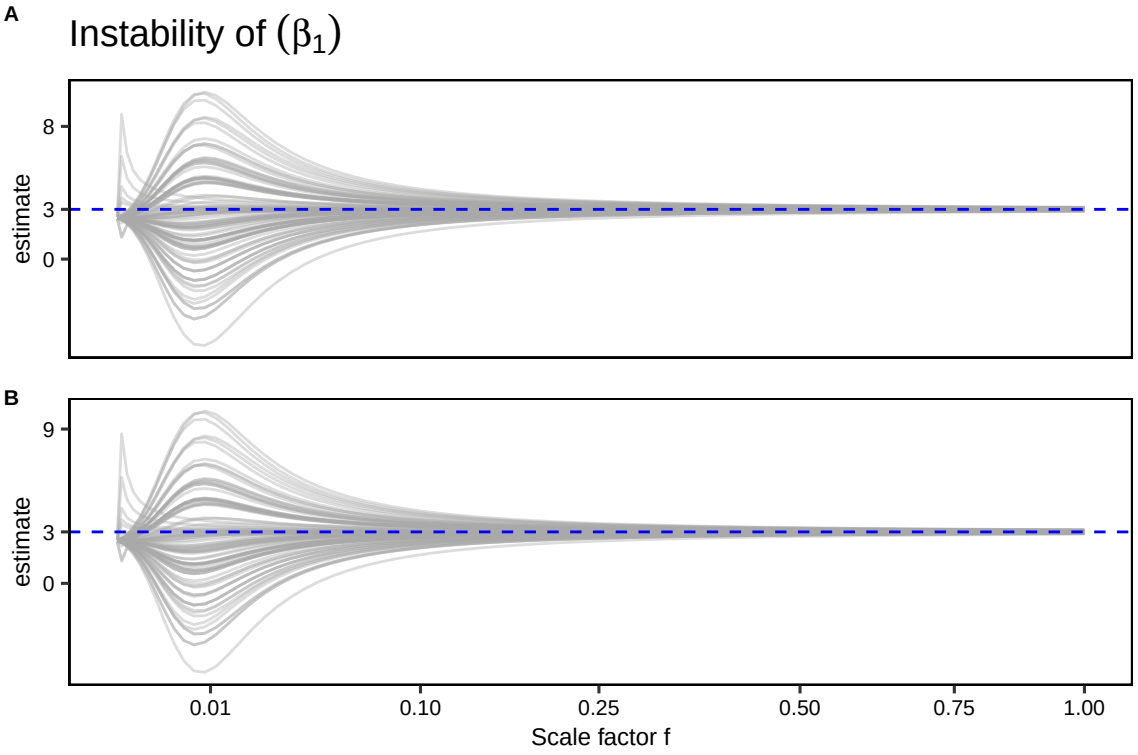

Figure 3.6: Instability of β_1 form `INLA` regression for different cases of collinearity. The horizontal axis indicates the values of the scale factor f in the square root scale. The vertical axis indicates the `INLA` estimate. The top plot for the $(\beta_0, \beta_1, \beta_2)$ model and plot is for the $(\beta_0, \beta_1, \beta_2, \beta_3)$ model. Each plot has 100 scenarios (gray lines). The sample size is 300.

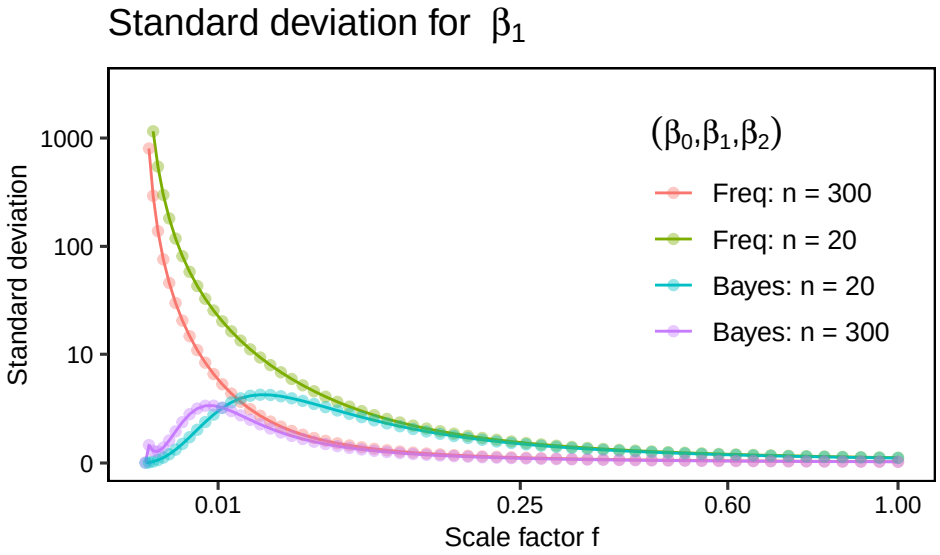

Figure 3.7: Standard deviation of the estimates for four models. The x axis represents the scale factor f (in the square root transformation scale). The vetrical axis marks the standard deviation value (in the pseudo log transformation scale) of estimate β_1 from the $(\beta_0, \beta_1, \beta_2)$ model. Colours represent different sample sizes for Bayesian or Frequentist models.

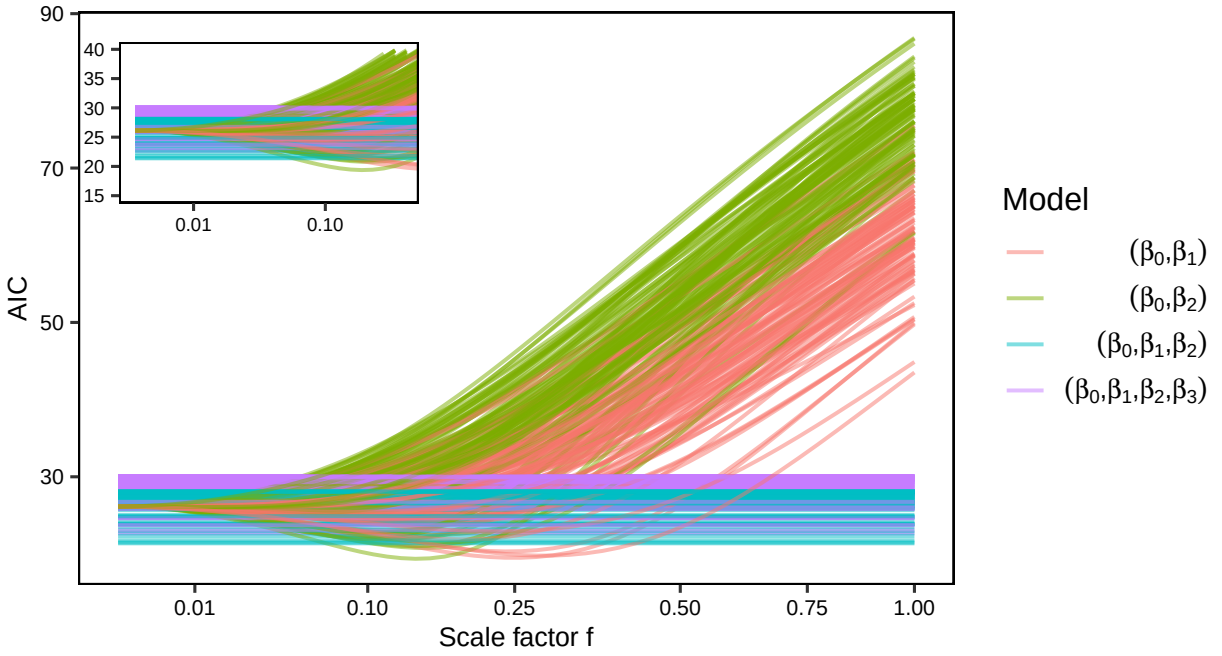

Figure 3.8: Instability of AIC for severe cases of collinearity. The horizontal axis indicates the values of the scale factor f. The vertical axis indicates the AIC value. The small plot within the figure enlarges the 0 to 0.20 scale factor region. The sample size is n = 20.

3.3 Model choice criteria

Small sample size (n = 20)

To investigate the instability of the model choice criteria, we used 100 bootstrap replications of length (20) and we estimated the corresponding model choice criteria. Figure 3.8 illustrates the instability of the AIC model selection criterion according to the scale factor f. Each replication is indicated with one line and each color signifies one model as is Table 2.7 (model (β_0) is excluded). The small plot within the figure is the enlarged left part of the plot. It can be seen that the AIC model estimate has some variability which is especially concerning in the left most area of the plot where the scale factor values are approaching zero.

Figure 3.9, similar with figure 3.8, illustrates the instability of the AIC model choice criterion for decreasing cases of collinearity. Here, the instability is expressed in terms of 95% bootstrap confidence intervals around the median. It is clear that for scale factors less than 0.5 the confidence intervals from different model choice criteria overlap. This indicates no significant difference for the AIC values between the different models. The AIC criterion fails to distinguish with evidence the best out of four models for cases of severe collinearity. Similar behavior is observed for the figure 3.10 and the BIC criterion.

At the left most area of the plot, we notice very narrow confidence intervals for the red and the green models. This is explained by the very small residuals those two models have when the scale factor is small.

Figure 3.11 illustrates the instability of the model choice criteria in the Bayesian framework for different levels of collinearity. Overall, the model choice criteria are unstable for small scale factors f. They fail to distinguish between the better model. All model choice criteria, DIC, WAIC, LCPO and LML have similar patterns for the five considered models. The two models including both predictors ($\beta_0, \beta_1, \beta_2$) and ($\beta_0, \beta_1, \beta_2, \beta_3$) have a constant evaluation from all model choice criteria used. The two models that include only one predictor, (β_0, β_1) and (β_0, β_2) have different model choice criteria

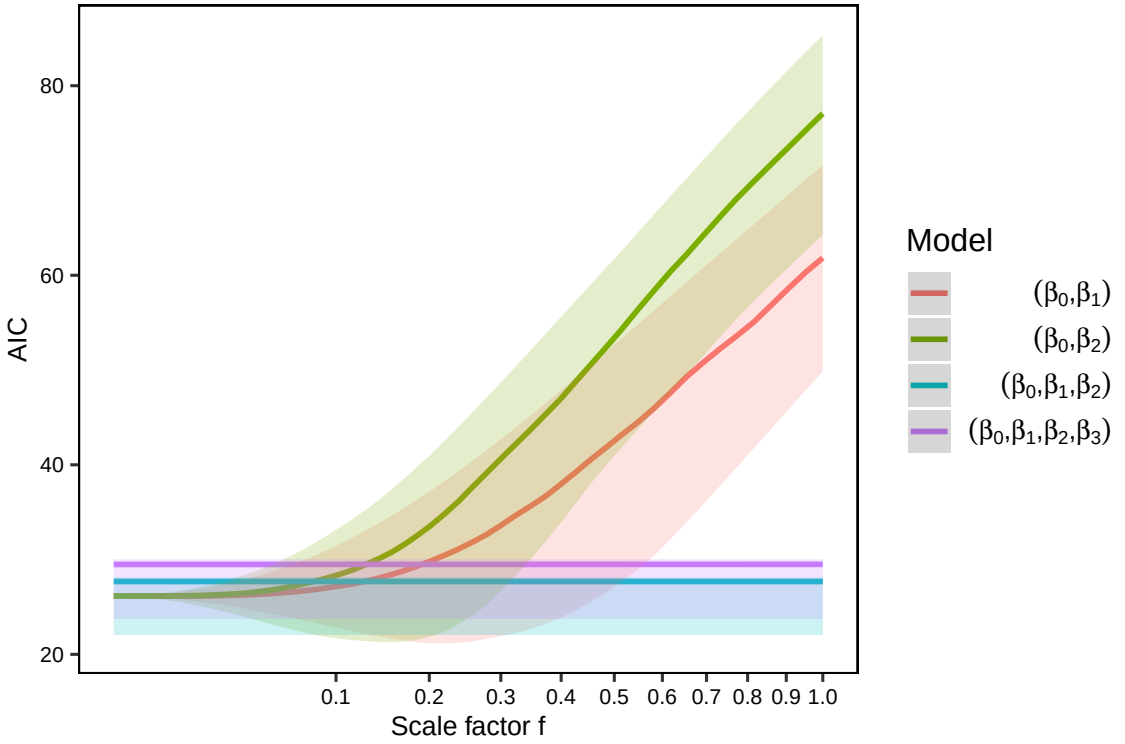

Figure 3.9: Instability of AIC for different cases of collinearity. The horizontal axis indicates the values of the scale factor f. The vertical axis indicates the AIC value for four different regression models. The thick lines represent the median AIC of each model whereas the transperant area represent the 95 percent bootstrap confidence interval area. The sample size is n = 20.

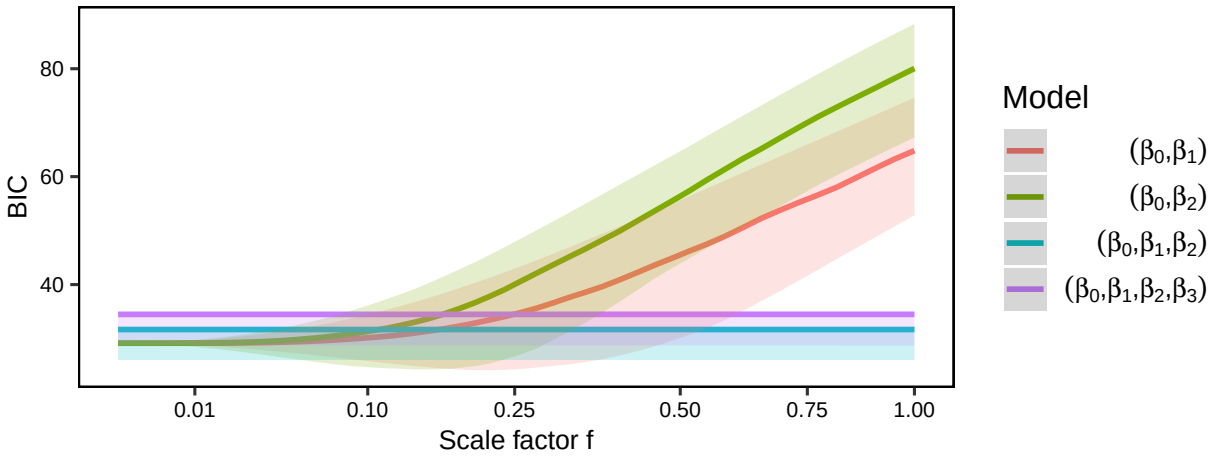

Figure 3.10: Instability of BIC for severe cases of collinearity. The horizontal axis indicates the scale factor f. The vertical axis indicates the BIC values. The thick lines represent the meidan BIC of each model for the specific scale factor whereas the transperant area represent the 95 percent bootstrap confidence interval area. Different models are marked by different colours. The sample size is n = 20.

values for different cases of collinearity. In detail, we observe a rise in the model choice criteria for increasing values of the scale factor f. The green line (β_0, β_2) has always a worse evaluation compared to the red line (β_0, β_1). However, the 95% bootstrap confidence intervals of the two models overlap throughout all the length of the scale factor f indicating that none of the model choice criteria provides

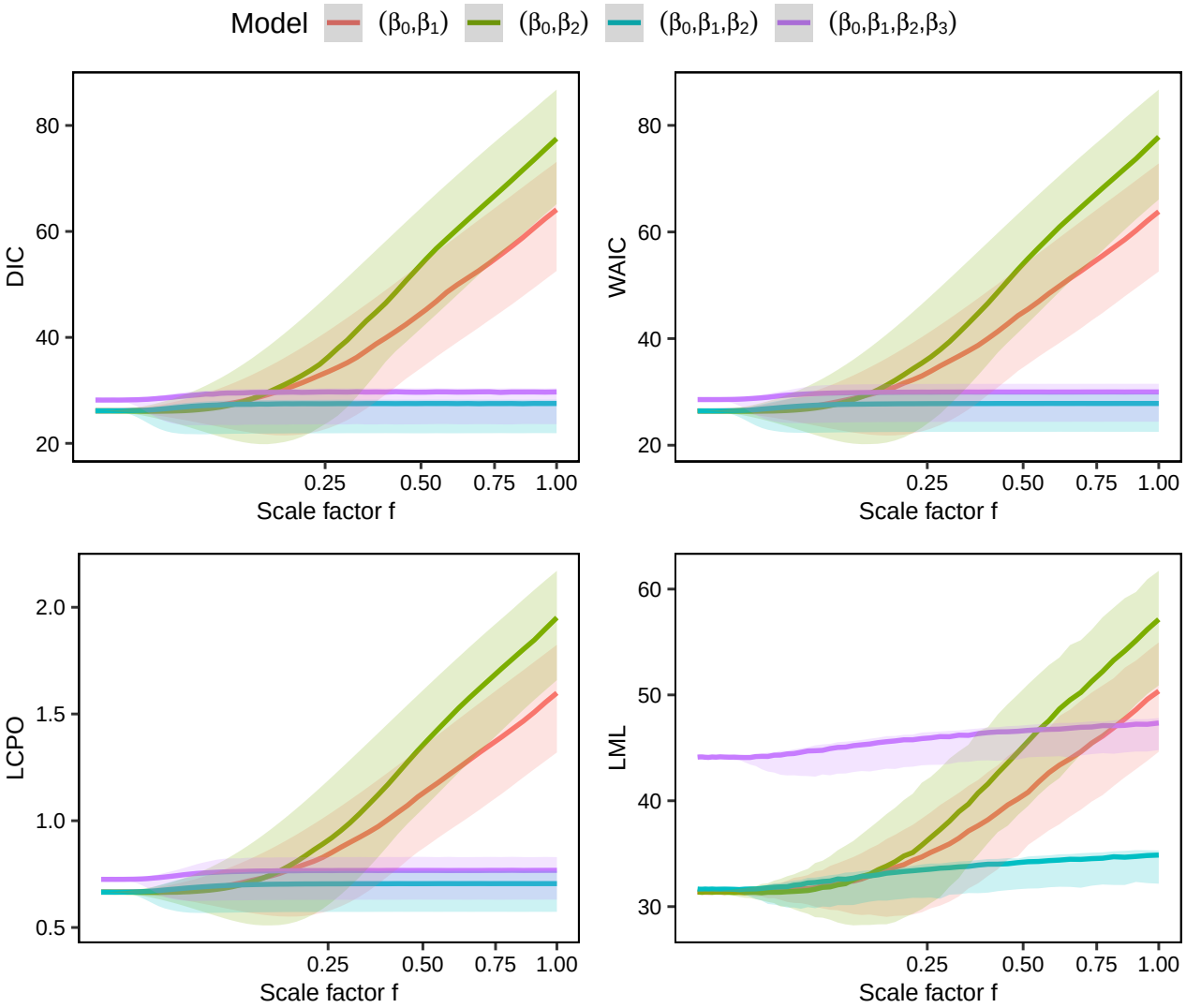

Figure 3.11: Instability of Bayesian model choice citeria for small sample size. Top left for DIC, top right for WAIC, bottom left for LCPO and bottom right for LML. The horizontal axis represents the scale factor f in the square root scale. Different colours represent different models. The simplest model representing the mean was omited from the plots. The sample size is n = 20.

significant evidence in favor of one of the two models. For scale factors greater than 0.5, DIC, WAIC and LCPO prefer the $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$ models with evidence. Nevertheless, there is no significant difference between the two regarding the three mentioned criteria. The logarithmic marginal likelihood model choice criterion is the only case where we have significant evidence in favor of a unique model. For scale factors greater than 0.75, LML provides evidence that the $(\beta_0, \beta_1, \beta_2)$ (true) model is significantly superior from the rest.

Figures 3.12 and 3.13 illustrate the empirical kernel densities of the AIC, BIC, LCPO, DIC, LML and WAIC estimated values for the four different models considered. First, the model choice criteria values of 60 different scale factors f were grouped in ten percentiles of the scale factor f as shown in the ticks of the vertical axis. For the estimation of the densities, we used a AIC Gaussian kernel with bandwidth 1, for DIC a bandwidth (bw) = 2, for LCPO a bw = 0.05, for WAIC a bw = 2 and for LML a bw = 1. The bandwidths were chosen after visual inspection of the plots. Furthermore, we

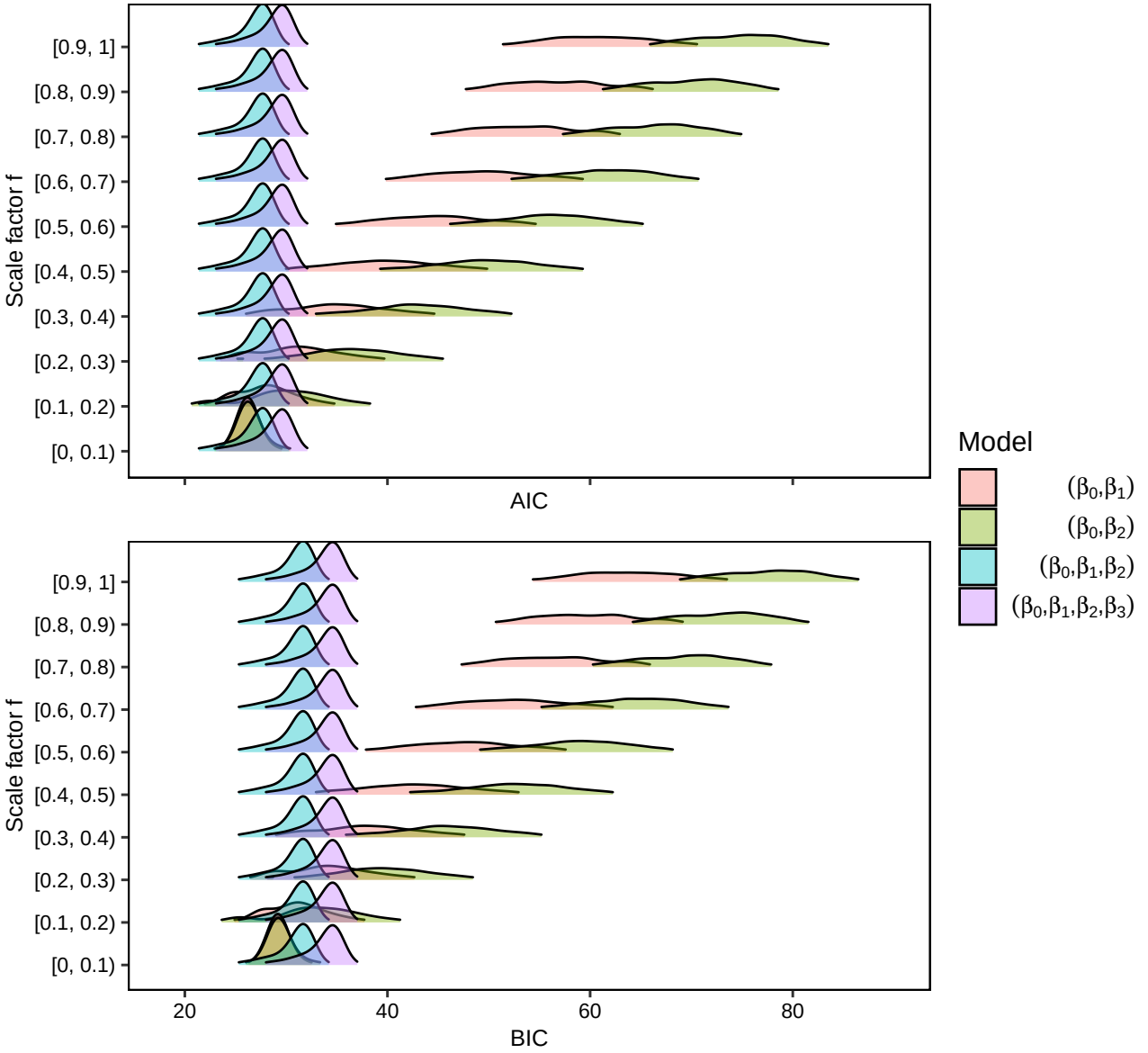

Figure 3.12: Densities of AIC and BIC criteria for four models considered (marked with different colours). The vertical axis represents the percentiles of the scale factor f. Sample size = 20

cut the estimated density to the 0.025 and 0.975 percentiles. Therefore, only the 95% of the density is presented. If two densities at the same height do not overlap, there is significant evidence at level 0.05 for the difference of the AIC value between the two models.

Large sample size (n = 300)

Figures 3.14 illustrate the instability of the AIC and BIC model choice criteria for decreasing levels of collinearity and for a sample size of n = 300. Compared with Figures 3.10 and 3.9 for small sample size, we notice the smaller 95% bootstrap confidence interval area. Moreover, we again observe the failure of the two frequentist criteria to select a model for cases of severe collinearity. The confidence intervals bands of all models overlap for scale factors between zero and 0.2, after which they drift apart. The AIC and BIC evaluation for the more complicated models remains stable for all scale factors f. Also,

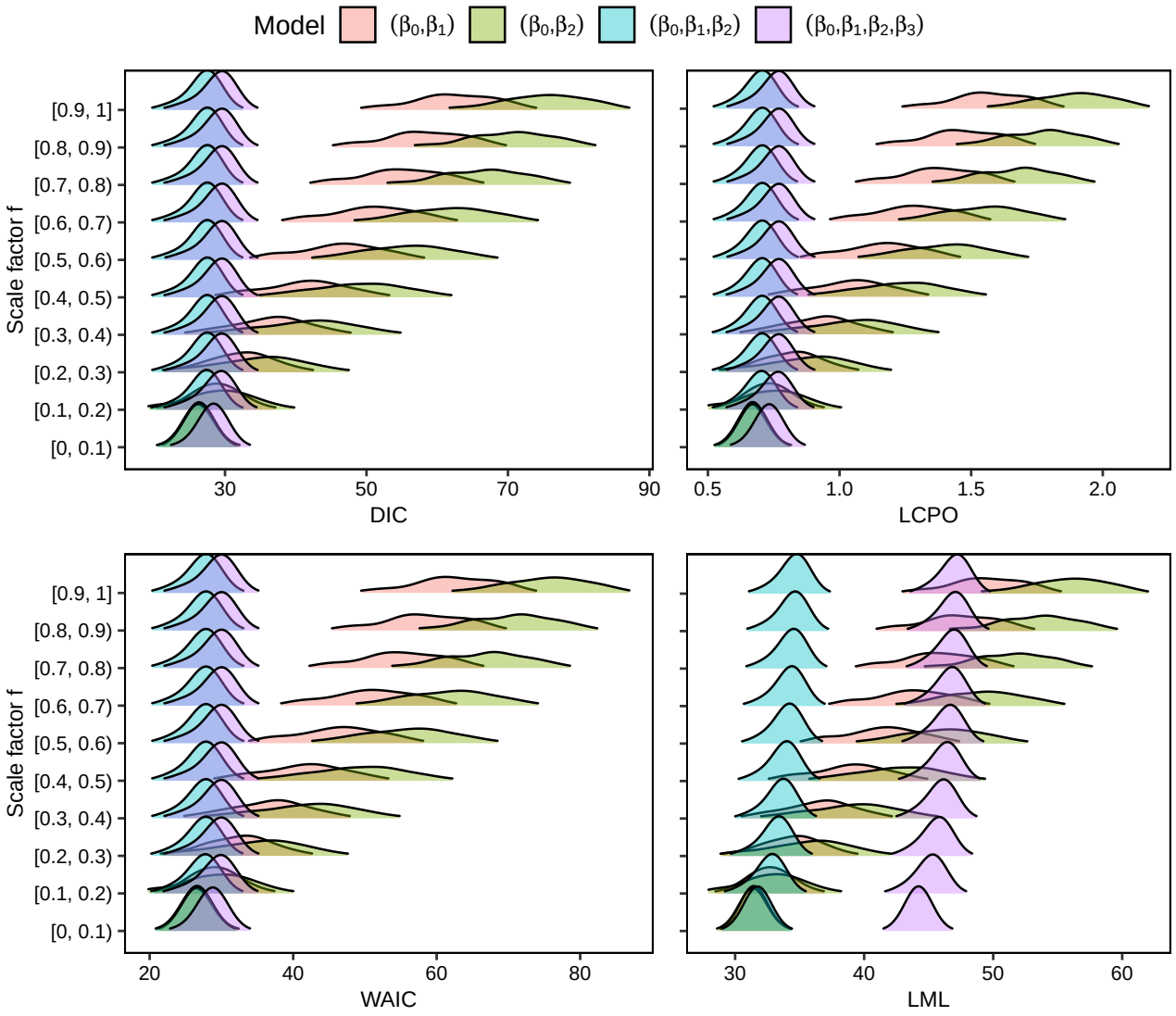

Figure 3.13: Kernel densities of Bayesian model choice criteria for increasing scale factor f. The sample size is n = 20. Different models are marked with different colour.

the confidence interval area around the two criteria is remarkably narrow.

Figure 3.15 illustrates the change of the Bayesian model choice criteria DIC, WAIC, LML and LCPO of four different models considered for different levels of collinearity. The model choice criteria agree regarding the selection of the best model when a best model can be identified. Similar to the frequentist framework, the criteria are unable to provide the better model in cases of severe collinearity. This can be observed in the left side of all plots. There, all the bootstrap confidence intervals for the different models overlap. The overlap of the confidence intervals suggests no significant differentiation between the considered models. For scale factors greater than 0.2 the overlap between the models stops. Specifically for the LML model choice criterion, all the four models have distinct evaluations and completely separated confidence interval bands. The criteria DIC, WAIC and LCPO suggest with equal evidence both $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$ models as the best ones (for f ≥ 0.2).

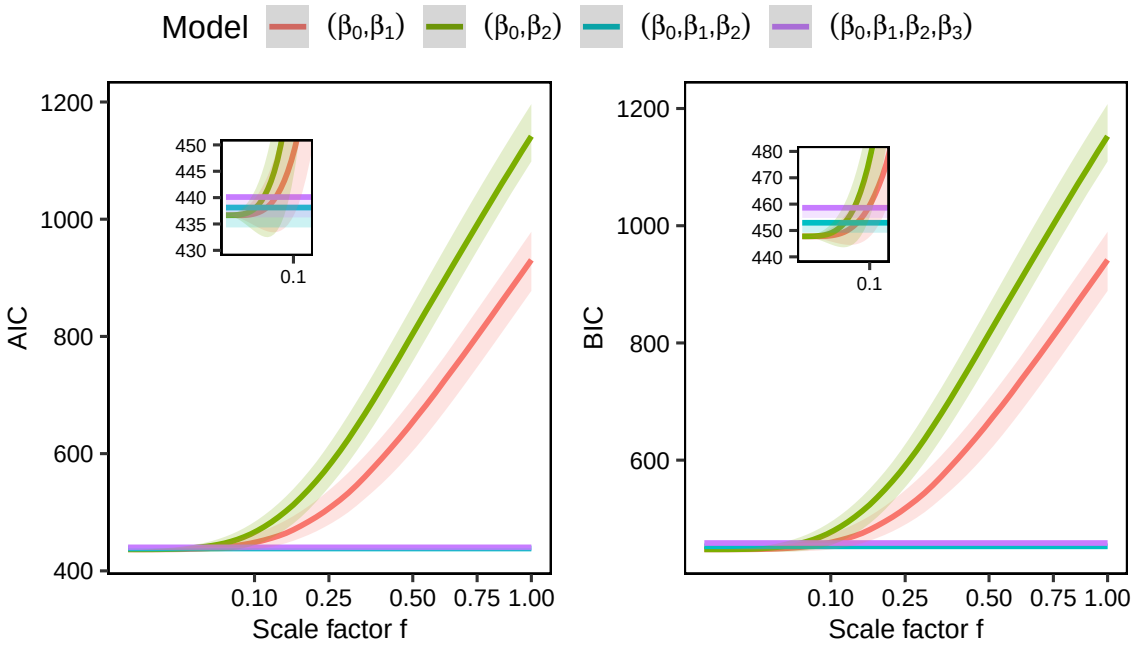

Figure 3.14: Instability of AIC and BIC for severe cases of collinearity. The horizontal axis indicates the values of the scale factor f (in the square root scale). The vertical axis indicates the AIC value. The small plot within the figure enlarges the 0 to 0.15 scale factor region. The sample size is n = 300.

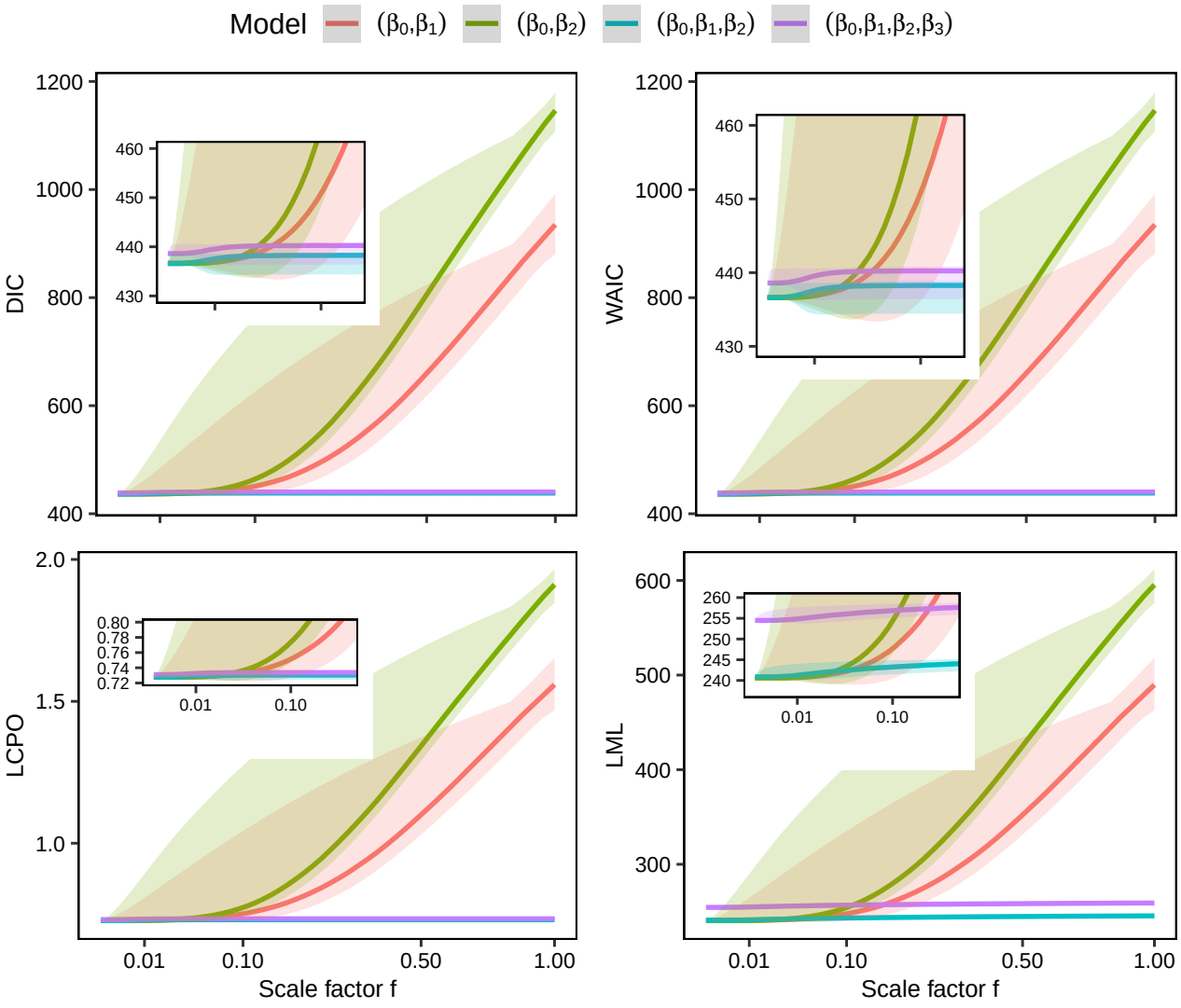

Figure 3.15: Instability of Bayesian model choice citeria. Top left for DIC, top right for WAIC, bottom left for LCPO and bottom right for LML. The horizontal axis represents the scale factor f in the square root scale. Different colours represent different models. The simplest model representing the mean was omited from the plots. The sample size is n = 300.

Chapter 4

Bodyfat Data

4.1 Discriptive statistics

The data contain n = 251 measurements for 13 variables. The response of interest is the percent body fat using Siri's equation. The included variables are **age** (in years), **weight_kg** (in kg), **height_cm** (is cm), **neck** circumference (in cm), **chest** circumference (in cm), **abdomen** circumference (in cm), **hip** circumference (in cm), **thigh** circumference (in cm), **knee** circumference (in cm), **ankle** circumference (in cm), **biceps** extended circumference (in cm), **forearm** circumference (in cm) and **wrist** circumference (in cm). All the variables are numeric. Summary statistics for the variables are provided in Table 4.1.

Heinze et al. [2018] used the unscaled version of this dataset. Here we investigate the differences in the collinearity diagnostics, estimates stability and general behavior and model selection criteria for different models in discrete cases of the raw data, the scaled data and the scaled and centered data. The different models considered and their notation can be found in Table 4.2.

An explanatory data analysis was performed for the Bodyfata dataset. We are interested in collinearity patterns within a subset of predictors. Figure 4.1 demonstrates pair plots, density plots and correlation coefficients (Pearson and Spearman) between some of the variables from the Bodyfat dataset. From the figure, we can distinguish high, positive correlation between the predictor **weight_kg** and **abdomen** with a correlation coefficient value of 0.874. The rest of the predictors included in the figure have smaller, non-alarming correlation values. The density plots for the included

Table 4.1: Summary statistics for Bodyfat data.

	Mean	SD	Median	Min	Max	Skew
siri	19.09	8.32	19.20	0.00	47.50	0.14
age	44.88	12.63	43.00	22.00	81.00	0.28
weight_kg	81.00	12.29	80.10	53.90	119.40	0.36
height_cm	178.61	6.67	178.00	163.00	197.00	0.09
neck	37.94	2.29	38.00	31.10	43.90	-0.04
chest	100.68	8.14	99.60	79.30	128.30	0.48
abdomen	92.33	10.22	90.90	69.40	126.20	0.41
hip	99.71	6.51	99.30	85.00	125.60	0.50
thigh	59.29	4.95	59.00	47.20	74.40	0.33
knee	38.55	2.32	38.50	33.00	46.00	0.26
ankle	23.08	1.65	22.80	19.10	33.90	2.27
biceps	32.22	2.92	32.00	24.80	39.10	0.04
forearm	28.66	2.02	28.70	21.00	34.90	-0.22
wrist	18.22	0.91	18.30	15.80	21.40	0.18

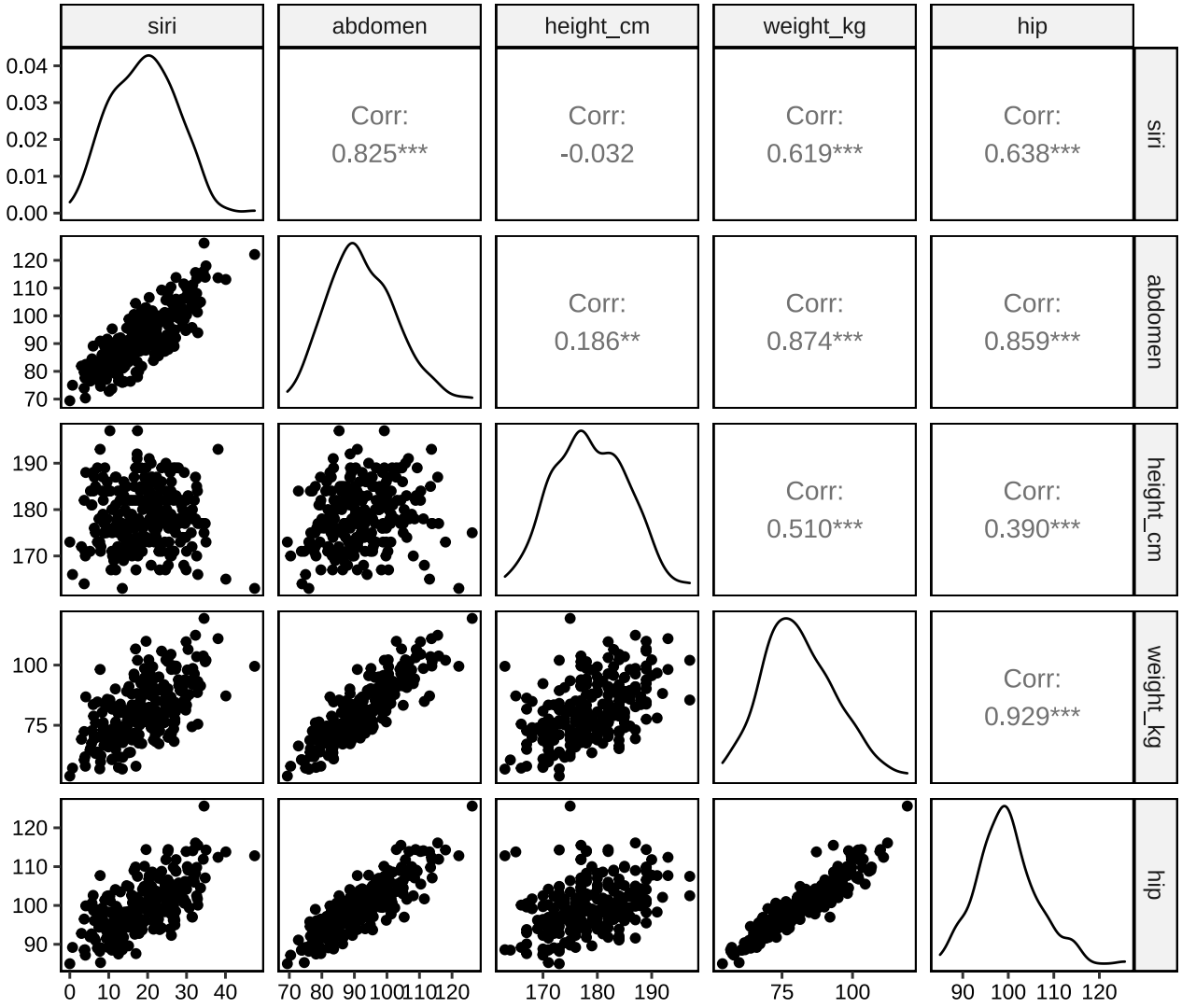

Figure 4.1: Pairs plot for variables abdomen, height_cm, weight_kg and siri from the Bodyfat data set.

variables indicate normally distributed predictors and response.

Figure 4.2 compiles information about all the correlation coefficients between the Bodyfat data predictors. Most of the correlation coefficients are below 0.8 (or above -0.8) indicating low concern about correlation. The highest observed absolute correlation coefficient value is 0.93 between the predictors **hip** and **weight_kg**. The second greatest correlation coefficient is 0.91 between variables **chest** and **abdomen**. In general, coefficients **weight_kg** and **hip** and **abdomen** seem to have greater values of the correlation coefficients compared with the rest of the predictors.

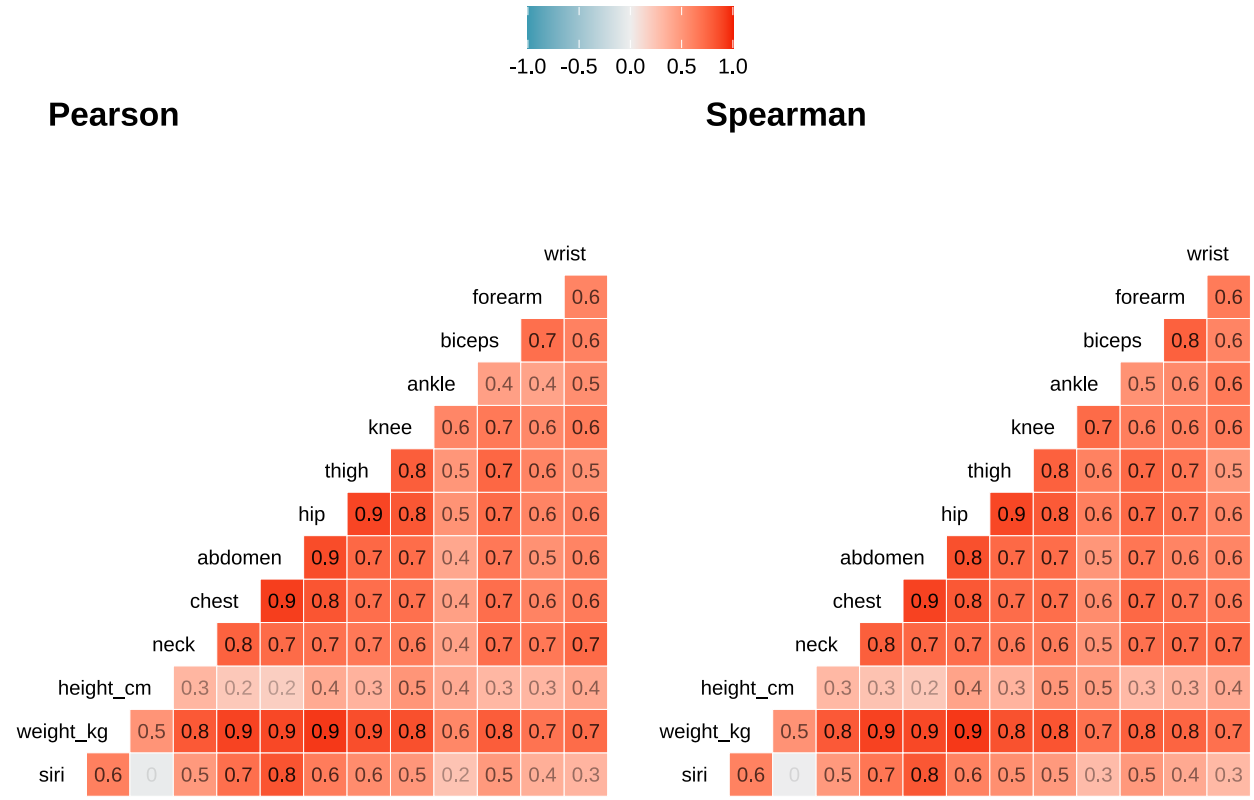

Figure 4.2: Pairwise Pearson and Spearman correlation coefficient plots for all the variables of the Bodyfat data. Transperancy indicates the magnitude of the correlation.

4.2 Collinearity diagnostics

Model fits

We conduct collinearity diagnostics for all the models from the Bodyfat dataset. However, we explicitly investigate the effect of scaling and centering for 8 specific models presented in Table 4.2 along with their notation. Table 4.4 contains the maximal equilibrated condition indexes $\tilde{\eta}_{max}$ and the AIC values of the considered models. Moreover, we specifically investigate the instability of the estimated parameters and that of the model choice criteria for 8 (different models) presented in Table 4.3. Finally, we estimated all the equilibrated condition indexes $\tilde{\eta}_i$ along with their variance decomposition proportions for 15 models (Tables 4.5 and 4.6 and Figure 4.3). All the combinations of models with the predictors **abdomen**, **height_cm**, **hip** and **weight_kg** were considered.

From Table 4.4 we can observe the effect of data transformations of scaling and centering on the estimates of AIC and $\tilde{\eta}_{max}$. Belsley [1991] suggests that the collinearity diagnosis should be performed before any transformation of the dataset.

Table 4.5 compiles collinearity diagnostics information for 15 regression models. The first column ($\tilde{\eta}_i$) provides the equilibrated condition indexes for each model. Moreover, the table includes the variance decomposition proportions for each $\tilde{\eta}_i$. Large variance decomposition proportions close to 1 indicate involvement of the predictor to collinearity. There is only one model with $\tilde{\eta}_{max} \leq 30$. Thus,

Table 4.2: Models considered for bodyfat data and their notation.

Notation	Model
(β_0)	siri $\sim \beta_0$
(β_0,β_w)	siri $\sim \beta_0 + \beta_w x_{weight}$
(β_0,β_h)	siri $\sim \beta_0 + \beta_h x_{height}$
(β_0,β_a)	siri $\sim \beta_0 + \beta_a x_{abdomen}$
$(\beta_0,\beta_h,\beta_w)$	siri $\sim \beta_0 + \beta_h x_{height} + \beta_w x_{weight}$
$(\beta_0,\beta_a,\beta_w)$	siri $\sim \beta_0 + \beta_a x_{abdomen} + \beta_w x_{weight}$
$(\beta_0,\beta_a,\beta_h)$	siri $\sim \beta_0 + \beta_a x_{abdomen} + \beta_h x_{height}$
$(\beta_0,\beta_a,\beta_w,\beta_h)$	siri $\sim \beta_0 + \beta_a x_{abdomen} + \beta_w x_{weight} + \beta_h\ x_{height}$

Table 4.3: Models considered for the research of collinearity in the Bodyfat dataset.

Notation	Model
(β_0,β_{hip})	siri $\sim \beta_0 + \beta_{hip} x_{hip}$
(β_0,β_w)	siri $\sim \beta_0 + \beta_w x_{weight}$
(β_0,β_{height})	siri $\sim \beta_0 + \beta_{height} x_{height}$
(β_0,β_a)	siri $\sim \beta_0 + \beta_a x_{abdomen}$
$(\beta_0,\beta_{hip},\beta_w)$	siri $\sim \beta_0 + \beta_{hip} x_{hip} + \beta_w x_{weight}$
$(\beta_0,\beta_{hip},\beta_{height})$	siri $\sim \beta_0 + \beta_{hip} x_{hip} + \beta_{height} x_{height}$
$(\beta_0,\beta_{hip},\beta_a)$	siri $\sim \beta_0 + \beta_{hip} x_{hip} + \beta_a x_{abdomen}$
$(\beta_0,\beta_{hip},\beta_a,\beta_w,\beta_{height})$	siri $\sim \beta_0 + \beta_{hip} x_{hip} + \beta_w x_{weight} + \beta_{height} x_{height} + \beta_a x_{abdomen}$

Table 4.4: AIC and maximal equilibrated condition index ($\tilde{\eta}_{max}$) for eight Bodyfat models for raw, scaled and centered data

	Raw		Scaled		Centered	
model	AIC	$\tilde{\eta}_{max}$	AIC	$\tilde{\eta}_{max}$	AIC	$\tilde{\eta}_{max}$
(β_0,β_{hip})	1'779		254		715	
(β_0,β_w)	1'660	13	135	13	596	1
(β_0,β_{height})	1'781	54	256	54	717	1
(β_0,β_a)	1'495	18	-30	18	431	1
$(\beta_0,\beta_{hip},\beta_w)$	1'585	72	60	72	521	2
$(\beta_0,\beta_{hip},\beta_{height})$	1'460	40	-65	40	396	4
$(\beta_0,\beta_{hip},\beta_a)$	1'468	66	-57	66	404	1
$(\beta_0,\beta_{hip},\beta_a,\beta_w,\beta_{height})$	1'460	123	-65	123	396	6

the rest 14 models are considered to have harmful collinearity.

Table 4.6 is similar to Table 4.5. The difference is that the former has been centered prior to the estimation of $\tilde{\eta}_i$ and the variance decomposition proportions. Here, all of the models have $\tilde{\eta}_{max} \leq 30$. This significant decrease of the condition indexes is a result of the centering of the dataset.

Figure 4.3 illustrates the variance decomposition proportions of different predictors from the 15 different models. Each vertical line represents one model and each dot marks the eigen value of one predictor. Different dot shapes represent the 5 predictors including the constant variable. The lower part of the plot presents the maximal equilibrated condition index of each model (arranged in increasing order). According to Belsley [1991], regression models with maximal equilibrated condition index $\tilde{\eta}_{max} \geq 30$ should be considered to have harmful collinearity. The horizontal dashed red line marks the threshold of 30. It can be seen that most of the models are diagnosed with high collinearity.

Figure 4.4 illustrates the densities of the maximal equilibrated condition indexes $\tilde{\eta}_{max}$ for all the

Table 4.5: 15 models for the Bodyfat data set (Figure 4.3), their equilibrated condition indexes ($\tilde{\eta}_i$) and their variance decomposition proportions matrices. The data are not modified.

Model	Eq. Cond. Ind.	weight_kg	height_cm	abdomen	hip	constant
siri ~ weight_kg + height_cm + abdomen + hip + c	1.00	0.00	0.00	0.00	0.00	0.00
	18.09	0.04	0.01	0.01	0.00	0.01
	44.42	0.15	0.04	0.37	0.00	0.00
	93.87	0.01	0.25	0.35	0.61	0.00
	160.20	0.80	0.70	0.27	0.39	0.99
siri ~ height_cm + abdomen + hip + c	1.00		0.00	0.00	0.00	0.00
	22.35		0.03	0.20	0.00	0.04
	70.24		0.20	0.29	0.34	0.74
	85.49		0.78	0.51	0.65	0.23
siri ~ weight_kg + abdomen + hip + c	1.00	0.00		0.00	0.00	0.00
	18.21	0.09		0.01	0.00	0.06
	45.55	0.31		0.97	0.01	0.02
	103.73	0.60		0.02	0.99	0.92
siri ~ abdomen + hip + c	1.00			0.00	0.00	0.00
	22.26			0.26	0.00	0.28
	63.77			0.74	1.00	0.72
siri ~ weight_kg + height_cm + hip + c	1.00	0.00	0.00		0.00	0.00
	17.19	0.12	0.01		0.00	0.01
	65.35	0.09	0.52		0.22	0.04
	127.63	0.80	0.48		0.78	0.95
siri ~ height_cm + hip + c	1.00		0.00		0.00	0.00
	35.10		0.06		0.97	0.12
	66.31		0.94		0.02	0.88
siri ~ weight_kg + hip + c	1.00	0.00			0.00	0.00
	16.18	0.15			0.00	0.06
	88.95	0.85			1.00	0.94
siri ~ hip + c	1.00				0.00	0.00
	30.74				1.00	1.00
siri ~ weight_kg + height_cm + abdomen + c	1.00	0.00	0.00	0.00		0.00
	16.19	0.07	0.01	0.01		0.01
	39.73	0.25	0.04	0.38		0.01
	123.34	0.68	0.95	0.60		0.98
siri ~ height_cm + abdomen + c	1.00		0.00	0.00		0.00
	19.62		0.03	1.00		0.03
	65.69		0.97	0.00		0.97
siri ~ weight_kg + abdomen + c	1.00	0.00		0.00		0.00
	16.03	0.17		0.01		0.57
	40.36	0.83		0.99		0.43
siri ~ abdomen + c	1.00			0.00		0.00
	18.17			1.00		1.00
siri ~ weight_kg + height_cm + c	1.00	0.00	0.00			0.00
	14.89	0.81	0.01			0.02
	72.28	0.18	0.99			0.98
siri ~ height_cm + c	1.00		0.00			0.00
	53.69		1.00			1.00
siri ~ weight_kg + c	1.00	0.01				0.01
	13.29	0.99				0.99

models created from the Bodyfat data set. The densities are seperated according to the number of predictors used for the model fit. All the 13 available predictors were used and a constant. We used Gaussian kernel densities for the creation of the empirical densities plot. The bandwidth was set to

Table 4.6: 15 models for the Bodyfat data set (Figure 4.3), their equilibrated condition indexes ($\tilde{\eta}_i$) and their variance decomposition proportions matrices. The data have been centered.

Model	Eq. Cond. Ind.	weight_kg	height_cm	abdomen	hip	constant
siri ~ weight_kg + height_cm + abdomen + hip + c	1.00	0.01	0.01	0.01	0.01	0.00
	1.73	0.00	0.00	0.00	0.00	1.00
	1.86	0.00	0.43	0.02	0.00	0.00
	5.10	0.00	0.13	0.55	0.57	0.00
	8.12	0.99	0.43	0.41	0.42	0.00
siri ~ height_cm + abdomen + hip + c	1.00		0.05	0.05	0.05	0.00
	1.42		0.00	0.00	0.00	1.00
	1.53		0.73	0.04	0.00	0.00
	4.21		0.21	0.91	0.95	0.00
siri ~ weight_kg + abdomen + hip + c	1.00	0.01		0.03	0.01	0.00
	1.67	0.00		0.00	0.00	1.00
	4.24	0.08		0.95	0.18	0.00
	6.31	0.91		0.02	0.80	0.00
siri ~ abdomen + hip + c	1.00			0.07	0.07	0.00
	1.36			0.00	0.00	1.00
	3.64			0.93	0.93	0.00
siri ~ weight_kg + height_cm + hip + c	1.00	0.02	0.06		0.02	0.00
	1.50	0.00	0.00		0.00	1.00
	1.81	0.01	0.78		0.03	0.00
	6.10	0.97	0.15		0.95	0.00
siri ~ height_cm + hip + c	1.00		0.30		0.30	0.00
	1.18		0.00		0.00	1.00
	1.51		0.69		0.69	0.00
siri ~ weight_kg + hip + c	1.00	0.04			0.04	0.00
	1.39	0.00			0.00	1.00
	5.23	0.96			0.96	0.00
siri ~ hip + c	1.00				1.00	0.00
	1.00				0.00	1.00
siri ~ weight_kg + height_cm + abdomen + c	1.00	0.02	0.04	0.02		0.00
	1.45	0.00	0.00	0.00		1.00
	1.58	0.00	0.41	0.04		0.00
	5.89	0.97	0.56	0.93		0.00
siri ~ height_cm + abdomen + c	1.00		0.41	0.41		0.00
	1.09		0.00	0.00		1.00
	1.21		0.59	0.59		0.00
siri ~ weight_kg + abdomen + c	1.00	0.06		0.06		0.00
	1.37	0.00		0.00		1.00
	3.86	0.94		0.94		0.00
siri ~ abdomen + c	1.00			1.00		0.00
	1.00			0.00		1.00
siri ~ weight_kg + height_cm + c	1.00	0.24	0.24			0.00
	1.23	0.00	0.00			1.00
	1.76	0.76	0.76			0.00
siri ~ height_cm + c	1.00		1.00			0.00
	1.00		0.00			1.00
siri ~ weight_kg + c	1.00	0.00				1.00
	1.00	1.00				0.00

one. Models with severe $\tilde{\eta}_{max}$ (≥ 2500) were excluded from the plot (eight in total). It can be seen, that models with less predictors have lower $\tilde{\eta}_{max}$. Moreover, most of the models have an $\tilde{\eta}_{max} \geq 30$. Thus, they are considered to have harmful collinearity patterns.

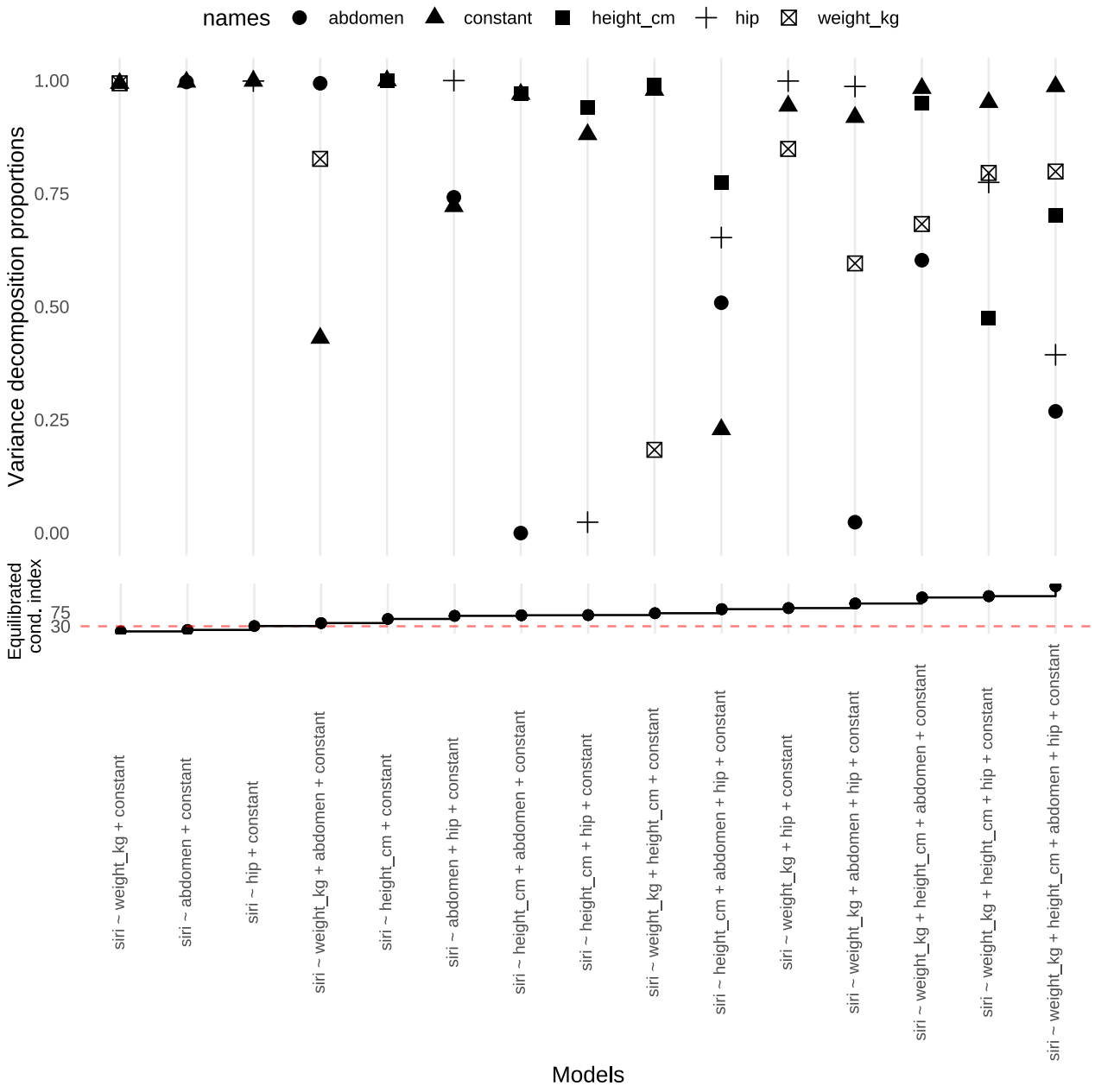

Figure 4.3: Variance decompossition proportions for the maximal equilibrated condition index for different bodyfat models. The top plot demonstrates the variance decompossition proportions along with their corresponding predictors (marked with different shapes) for each model. The bottom plot indicates the value of the maximal equilibrated condition index of each model.

4.3 Instability of estimates

To investigate the instability of the regression estimates we considered eight models using the four variables, **weight_kg**, **height_cm**, **hip** and **abdomen** as predictors (Table 4.3). We research the instability of the different parameter estimates from the different model fits.

We consider the variables **hip** and **weight_kg** to be collinear. We fit the linear model (4.1)

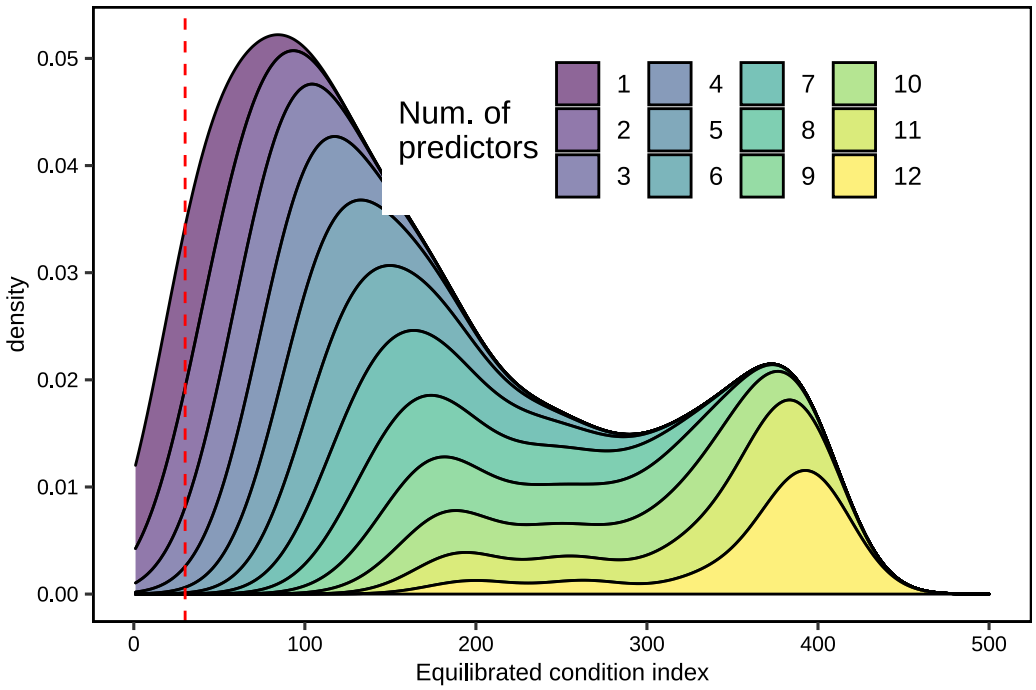

Figure 4.4: Gaussian kernel densities of maximal equilibrated condition indexes for all posible models from the Bodyfat dataset. The bandwidth is set to one. Different colours represent different number of predictors.

$$\text{hip} \sim \text{weight\_kg} \tag{4.1}$$

with response being the variable **hip** and predictor being the variable **weight_kg**. We use the residuals of this linear regression and the scale factor f, to govern the amount of collinearity between the two variables **hip** and **weight_kg**. The scaled (from the scale factor f) residuals are added to the expected values of 4.1. We repeat the procedure for 60 different scale factors from zero to three.

We replicate the above procedure 100 times using bootstrap samples of sample size 251 and replacement. We use the bootstrap replications to research the instability of the predicted estimates. The priors used for the Bayesian estimation are available in Table 4.7.

Figure 4.5 illustrates the standard deviation of the parameter estimates of **hip** from different models. The standard deviation represents the instability of the estimated parameters from the 100 bootstrap samples and for the 60 different cases of scale factors f. It is immediately clear that the two models (pink and brown) have significantly high standard deviation for small scale factors f. Both models include the predictors hip and weight predictors, the relationship of which was used to define the amount of collinearity. The rest thre models include either the predictor hip or weight but not both. Thus, it can be concluded that for increasing collinearity between hip and weight, the instability of the hip parameter is rising exponentially.

The right plot of the figure illustrates the effect of the sample size on the instability of the estimated parameters. Once again, the instability of the estimate is represented as the standard deviation of the **hip** estimates of the model ($\beta_0\beta_{hip}\beta_{weight}$). Five different sample sizes from 50 to the maximal of 251 were considered. Overall, for small scale factors f close to zero, we observe great instability for all sample sizes with the standard deviation being more than 1000. While the scale factor f increases, there is a exponential decrease of the standard deviation of the estimates. The standard deviations

Table 4.7: Prior parameters for Bodyfat models in INLA.

Coefficient	Mean	Precision
β_0	0	0.01
$\beta_{abdomen}$	0	0.01
β_{age}	0	0.01
β_{ankle}	0	0.01
β_{biceps}	0	0.01
β_{chest}	0	0.01
β_{forarm}	0	0.01
$\beta_{height_{cm}}$	0	0.01
β_{hip}	0	0.01
β_{knee}	0	0.01
β_{neck}	0	0.01
β_{thigh}	0	0.01
$\beta_{weight_{kg}}$	0	0.01
β_{wrist}	0	0.01

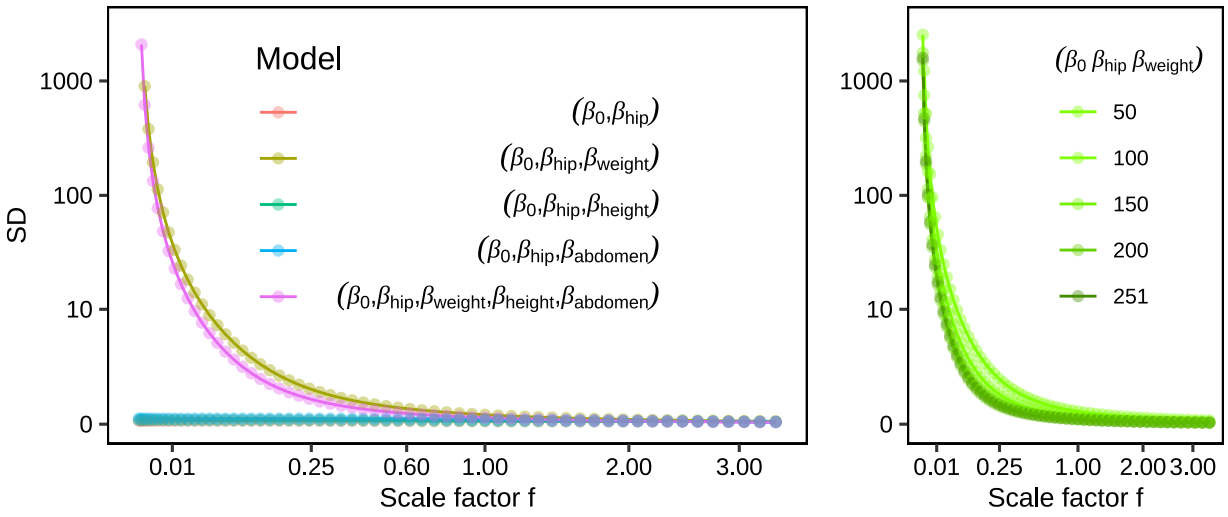

Figure 4.5: Standard deviation of the linear regression estimates for predictor variable **hip** from five different models (left) and for the $(\beta_0, \beta_{hip}, \beta_{weight})$ model for different sample sizes (right). The x axis is the scale factor f. y axis is the standard deviation of the 100 parameter estimates.

approximate zero for very large scale factors f close to one. Larger sample sizes have slightly smaller standard deviations within all the spectrum of scale factors.

Figure 4.6 illustrates the evolution of the **hip** regression estimate instability for increasing cases of collinearity. From left to right there are five plots arranged for increasing sample size. Each gray line represents one bootstrap replication. At the far left plot, for the smallest sample size of 50, we observe severe instability of the hip estimate especially in the left size of the plot where the scale factor f is close to zero. The estimations cover a huge width of values from less than $-10\,000$ to above $10\,000$.

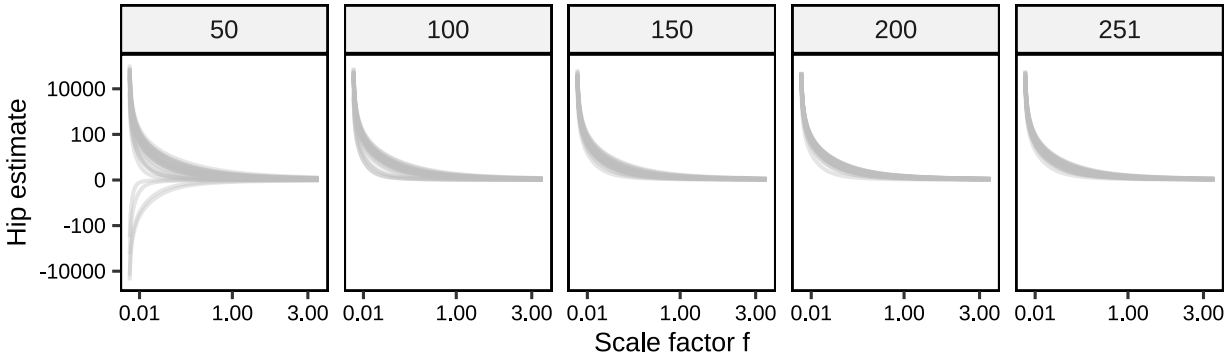

Figure 4.6: Instability of hip estimate for different sample sizes. The x axis is the scale factor f and the y axis is the linear regression estimate of hip predictor. The subtitles of the plots indicate the sample size of the bootstrap replications.

4.4 Model choice criteria

Figure 4.7 illustrates the instability of the model choice criteria for eight different models as in Table 4.3. The model choice criteria estimates were obtained from the bootstrap replications as described in the previous section (4.3).

From all the plots, the pink and the purple model have the smallest model choice criteria values at the left side of the plot. While the scale factor f increases, the two models distinguish from each other with the pink one indicating the smallest values. This pattern is observed for all the model choice criteria.

The purple, blue and brown models indicate a stable increase for all the model choice criteria and for scale factors greater than 0.5. In detail, the model choice criteria for the purple model are stable until scale factor of 0.5 after which they have a small and stable increase until the end of the plot (scale factor f = 3). The blue model is also stable until a scale factor f equal to 0.5 followed by an stable increase. In the case of the blue model, the overall change of the model choice criteria is greater. Finally, the model choice criteria for the brown models seem to be stable until a scale factor f of 1.5. Then they witness a slight increase. The model choice criteria for the rest of the models remain unchanged for all the spectrum of the scale factor f. None of the model choice criteria is able to significantly distinguish between the purple and the pink models with the smaller evaluations.

Figures 4.8 and 4.9 include all the possible models from the bodyfat data set. They illustrate the relationship between the model choice criteria and the maximal equilibrated condition indexes $\tilde{\eta}_{max}$ for each model. Specifically, there are 8192 models from different combinations of 13 variables. The models are categorized by the number or used predictors (from one to 13). Each colour represents a different category. Lighter colours indicate more predictors.

In general, different models have different levels of collinearity. Using the threshold of 30 to describe the collinearity as "concerning", we conclude that the majority of the available models have a concerning amount of collinearity. As in Figure 4.4, we observe that models with more predictors have greater amounts of collinearity. Regarding the model choice criteria values, we observe a wide spectrum of values. For all the model choice criteria we can distinguish three chunks of models. Moreover, none of the model choice criteria is able to suggest a unique model as the best one.

At a second step, we illustrate the relationship between AIC and $\tilde{\eta}_{max}$ for all models that include one predictor (Figure 4.10). It is clear that predictor **abdomen** in the top left plot has a significant effect to the AIC evaluation. To investigate closer the impact of the rest predictors to AIC, we exclude **abdomen** and we repeat the procedure in Figure 4.11.

Table 4.8 compiles information for the Spearman correlation coefficient estimates between the

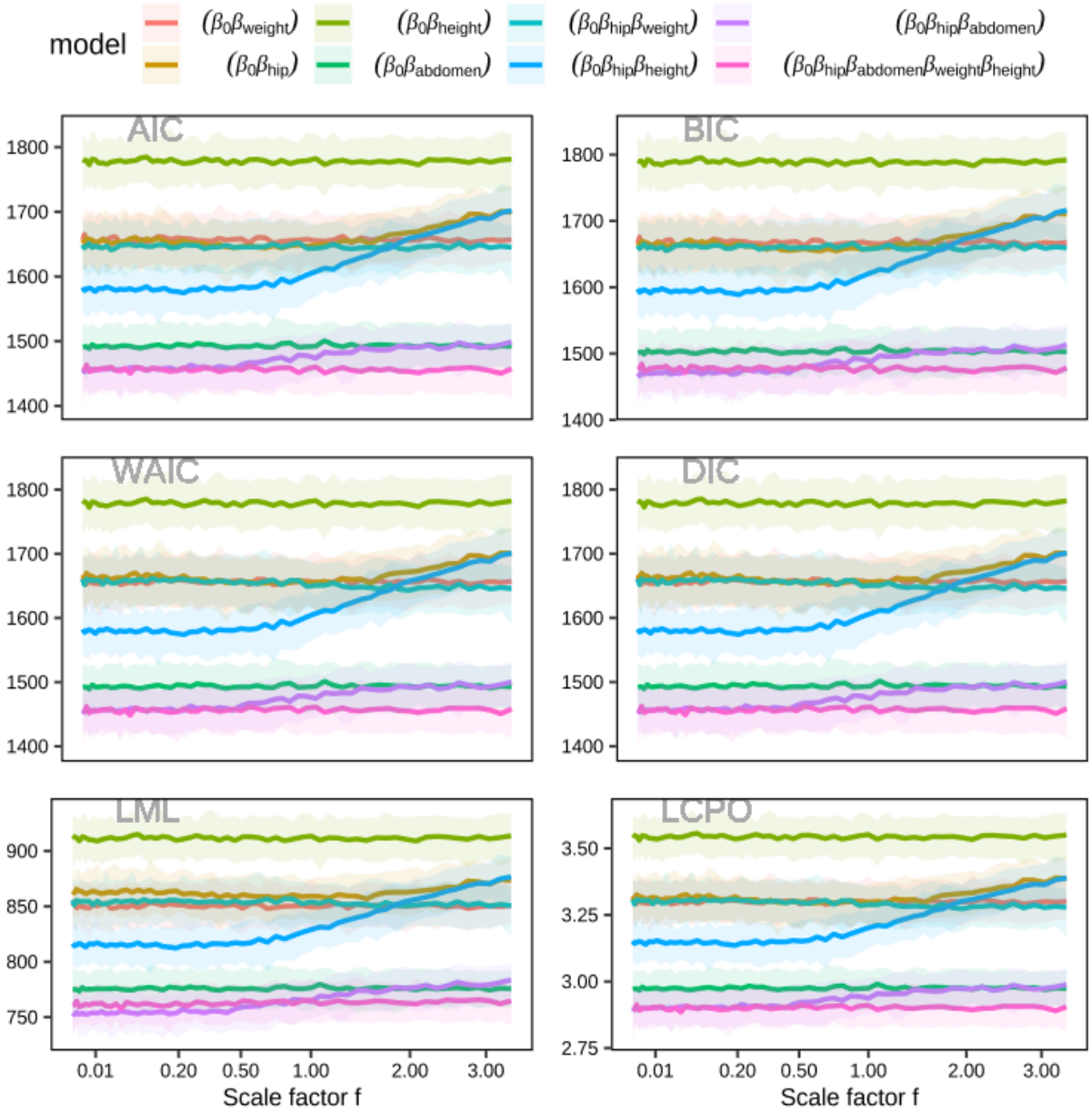

Figure 4.7: Instability of model choice criteria for eight different Bodyfat models. The sample size of the 100 bootstrap samples is 251. The x-axis is the scale factor f and the vertical axis is the value of the model choice criterion.

maximal equilibrated condition index and the model choice criteria as in Figure 4.8. Each column represents a different model choice criterion. For each model choice criteria, the value of the Spearman correlation estimate and its p-value are available. The rows represent the number of predictors from one to 12. The Spearman correlation coefficient estimates for models with 4, 5, 6 and 7 predictors are negative with strong evidence (for all model choice criteria except from LML). This negative correlation suggests that when collinearity increases, the model choice evaluation is decreasing. The rest of the strata regarding the number of predictors have positive correlation coefficients but with no significant p-values.

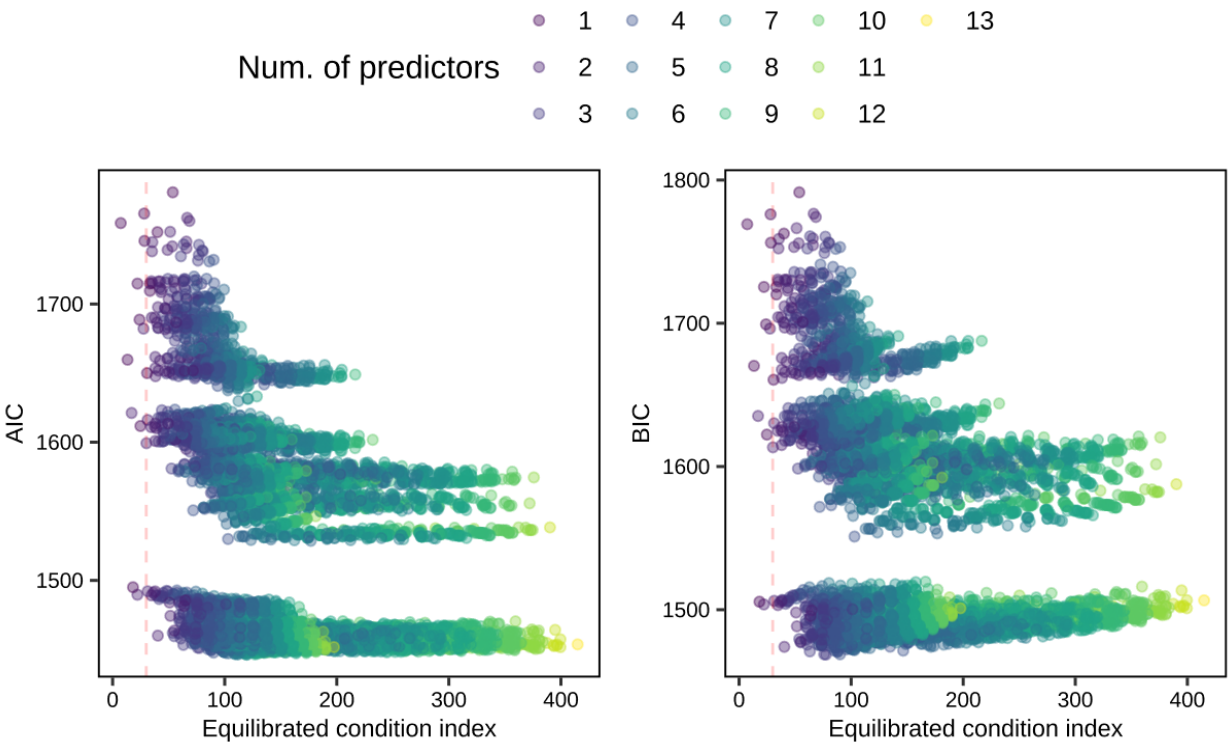

Figure 4.8: AIC, BIC and maximal equilibrated condition index for all possible model fits from the bodyfat data. The vertical axis is the AIC or BIC value for the each corresponding model. The horizontal axis represent the maximal equilibrated condition index of each model. The vertical red dashed line marks the equilibrated condition index of 30. Different colours identify different number of predictors.

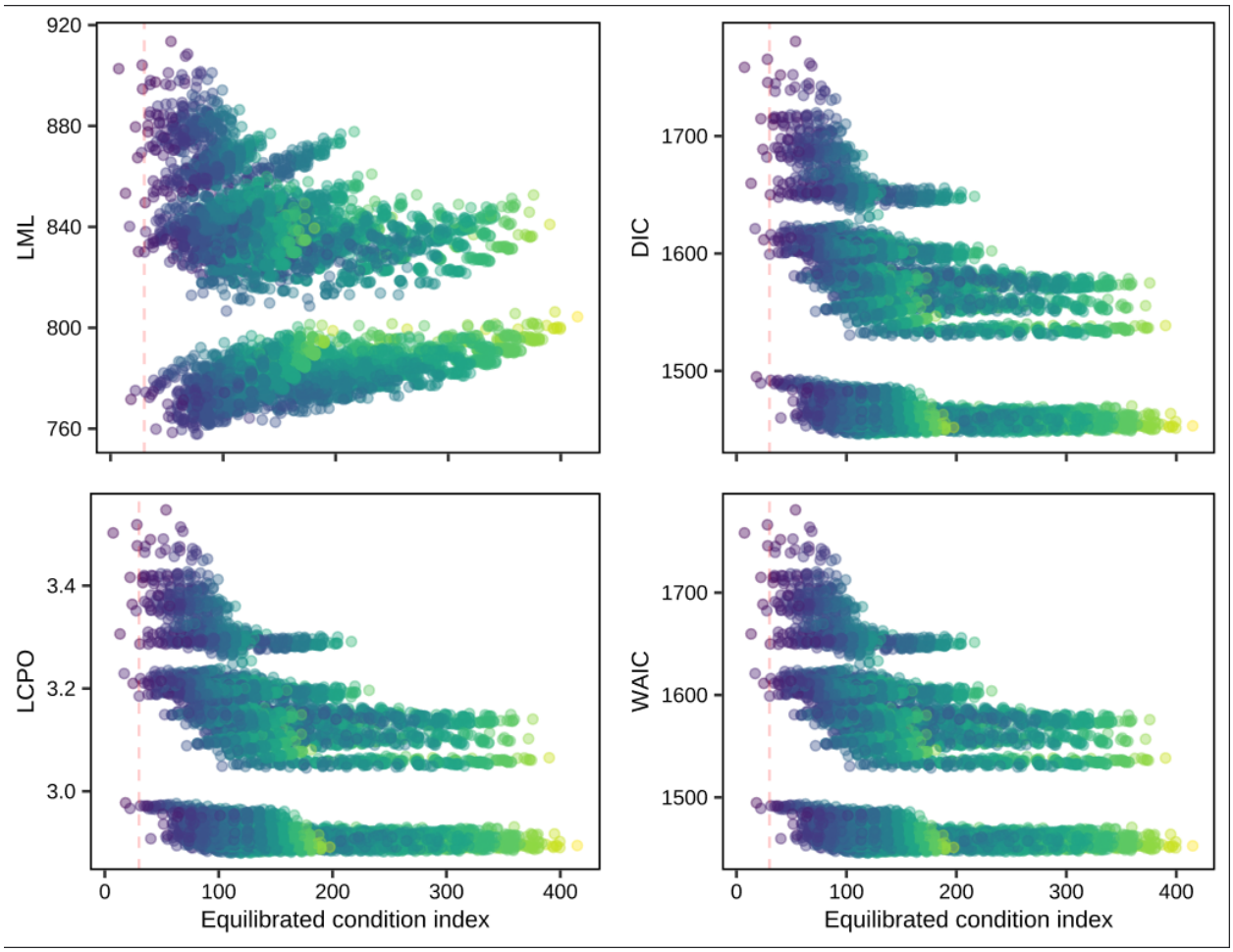

Figure 4.9: See Figure 4.8.

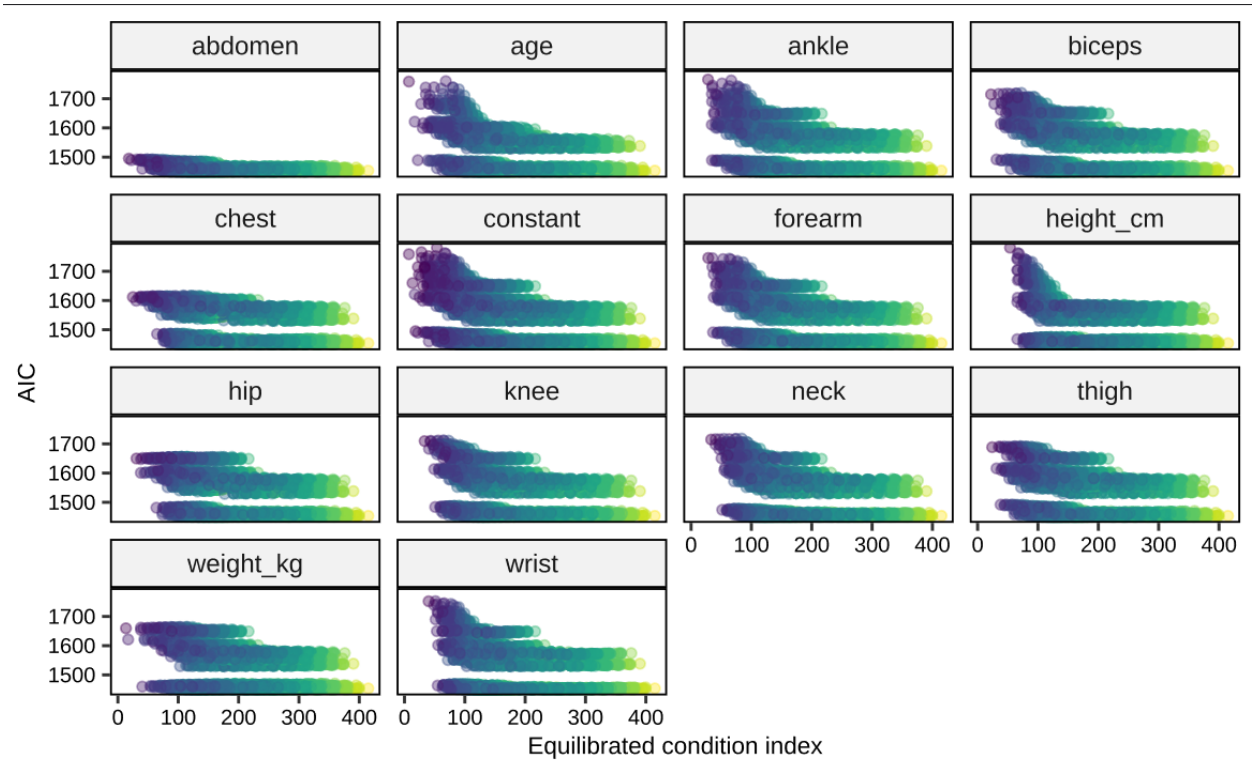

Figure 4.10: AIC and maximal equilibrated condition index for all possible model fits from the bodyfat data. The vertical axis is the AIC value for the each corresponding model. The horizontal axis represent the maximal equilibrated condition index of each model. Different colours identify different number of predictors. Lighter colours indicate more predictors. Each subplot contains all the models where the coresponding predictor is present.

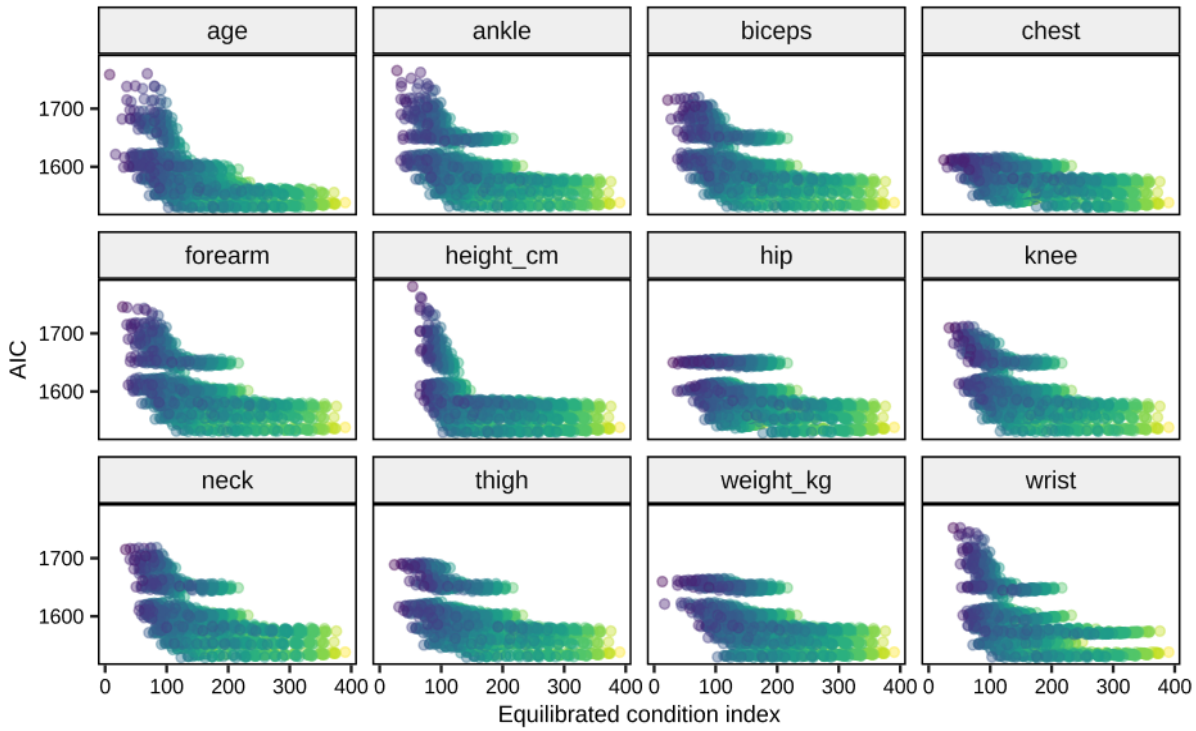

Figure 4.11: Similar with Figure 4.10 without abdomen variable.

Table 4.8: Spearman correlation coefficients ρ for the relationship between the model choice criteria and the maximal equilibrated condition index for each category of number of predictors. Second coloumn N indicates the number of models within the category of predictors.

		AIC		BIC	
Num. of preds	N	ρ	p-val	ρ	p-val
1	13	0.37	0.21	0.37	0.21
2	78	0.09	0.42	0.09	0.42
3	286	-0.11	0.068	-0.11	0.068
4	715	-0.15	< 0.0001	-0.15	< 0.0001
5	1287	-0.14	< 0.0001	-0.14	< 0.0001
6	1716	-0.11	< 0.0001	-0.11	< 0.0001
7	1716	-0.05	0.033	-0.05	0.033
8	1287	0.02	0.59	0.02	0.59
9	715	0.09	0.015	0.09	0.015
10	286	0.16	0.005	0.16	0.005
11	78	0.19	0.091	0.19	0.091
12	13	0.07	0.82	0.07	0.82

		WAIC		DIC		LCPO		LML	
Num. of preds	N	ρ	p-val	ρ	p-val	ρ	p-val	ρ	p-val
1	13	0.34	0.25	0.37	0.21	0.34	0.25	0.37	0.21
2	78	0.09	0.43	0.09	0.42	0.09	0.44	0.08	0.46
3	286	-0.11	0.065	-0.11	0.067	-0.11	0.064	-0.10	0.094
4	715	-0.15	< 0.0001	-0.15	< 0.0001	-0.15	< 0.0001	-0.13	0.0004
5	1287	-0.14	< 0.0001	-0.14	< 0.0001	-0.14	< 0.0001	-0.11	0.0001
6	1716	-0.10	< 0.0001	-0.10	< 0.0001	-0.10	< 0.0001	-0.06	0.012
7	1716	-0.05	0.032	-0.05	0.026	-0.05	0.031	0.00	0.99
8	1287	0.01	0.76	0.01	0.84	0.01	0.78	0.07	0.01
9	715	0.07	0.049	0.07	0.059	0.07	0.053	0.16	< 0.0001
10	286	0.13	0.025	0.13	0.029	0.13	0.03	0.27	< 0.0001
11	78	0.14	0.21	0.14	0.22	0.14	0.24	0.36	0.001
12	13	0.01	0.98	0.05	0.88	0.04	0.89	0.51	0.078

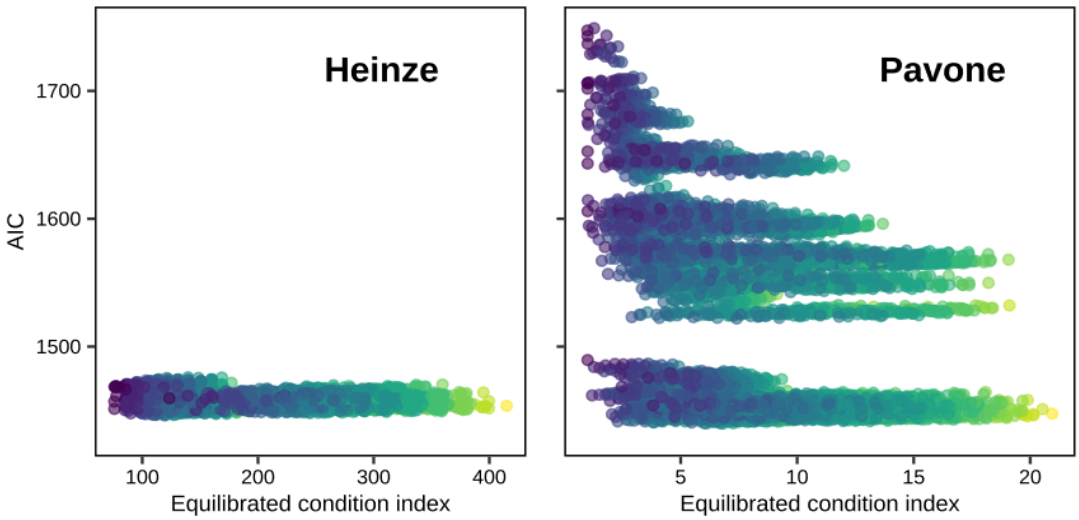

Figure 4.12: See Figure 4.8. AIC for models from Pavone and Hainze data.

4.5 Relevance of findings

Table 4.9 is similar with Table 4.8 after we apply the transformation from Heinze et al. [2018] and Pavone et al. [2020]. We notice that the left table has only positive correlation coefficients indicating a positive assotiation between AIC and collinearity.

Table 4.9: Spearman correlation coefficient ρ for the relationship between maximal equilibrated condition index and AIC for all models from the Bodyfata data set. The models are categorized according to the number of predictors used. Column 'N' indicates the number of model fits in each category of predictors. The data have been modified according to [Heinze et al., 2018] (left) and [Pavone et al., 2020] (right) in Figure 4.12

Num. of preds.	N	ρ	p-value
3	11	0.14	0.69
4	55	0.01	0.94
5	165	0.01	0.91
6	330	0.03	0.56
7	462	0.09	0.065
8	462	0.15	0.001
9	330	0.20	0.0002
10	165	0.25	0.001
11	55	0.27	0.046
12	11	0.21	0.54

Num. of preds.	N	ρ	p-value
1	13	-0.82	0.0006
2	78	-0.24	0.031
3	286	-0.29	< 0.0001
4	715	-0.29	< 0.0001
5	1287	-0.28	< 0.0001
6	1716	-0.24	< 0.0001
7	1716	-0.17	< 0.0001
8	1287	-0.08	0.003
9	715	0.01	0.70
10	286	0.10	0.08
11	78	0.16	0.17
12	13	0.12	0.71

Chapter 5

Conclusions

We have investigated the behavior of classical and Bayesian model choice criteria in different scenarios of collinearity levels and we have shown the impact of collinearity on them. We applied the singular value decomposition method with equilibration from Belsley [1991] to quantify the amount of collinearity in different models. As shown by Belsley [1991] this method is numerically stable and comparable between models. Although a threshold for a harmful amount of collinearity is unknown, it is easy to identify large singular values within each model. Figure 4.4 shows that a larger number of predictors leads to higher collinearity. Therefore, researchers should be especially concerned about collinearity when many predictors are considered.

To understand the behavior of model choice criteria under collinearity, we ran a simulation for different scenarios of collinearity between the predictors while also taking into account different sample sizes in both classical and Bayesian framework. Finally, we used a case study based on the Bodyfat data set which is well known from literature [Pavone et al., 2020, Heinze et al., 2018]. To investigate how unstable the model choice criteria are due to collinearity and also to quantify the instability of the regression estimates in classical and Bayesian framework, we extended both the simulation and the case study.

Figure 2.6 and Tables 2.2 – 2.6 demonstrate an explicit example where the eight predictors of a linear model linearly describe the ninth. Although this example shows that there is no correlation between the predictors, Belsley's approach clearly identifies the collinearity problems. We confirm the recommendation put forward by Belsley [1991] that the singular values decomposition together with the equilibrated condition indexes is the preferable method of collinearity diagnostics. It is a numerical stable and comparable approach for estimating the level of collinearity. Moreover, the variance decomposition proportions do indicate which predictors are involved in collinearity.

Pavone et al. [2020] and Heinze et al. [2018] apply two different model selection methods both of which include AIC as the basic model choice criterion. We provide results indicating the instability of AIC both in the simulation and in the Bodyfat case study. Table 4.9 and Figure 4.12 demonstrate the uncertainty of the AIC criterion regarding the selection of the best model in both studies.

Figures 4.8 – 4.9 and Table 4.8 indicate a negative and positive association between the amount of collinearity and the model choice criteria values depending on the number of predictors considered. This finding certifies that model choice criteria can be sensitive to collinearity. This demonstrates the relevance of collinearity diagnostics as a part of the model selection procedures.

We have shown the advantage of the Bayesian framework (compared to classical) for cases of severe collinearity. Figures 3.3, 3.4, 3.5 and 3.6 show the instability of the estimates for cases of severe collinearity. Although instability of estimates is present in both classical and Bayesian framework, the later has significantly smaller magnitude. We showed that Bayesian models are more stable especially in cases of high collinearity. Figures 3.7 also clarifies the stability of the Bayesian estimates.

We illustrate that the sample size has a beneficial impact on the estimation stability and model

choice criteria evaluation when collinearity is severe. The comparison of Figures 3.9 and 3.14 demonstrates the difference of the AIC stability between the case of small and large sample size. We found similar effects for the impact of sample size on other model choice criteria. These findings agree with previous research on this topic.

The R package 'Collinearity' that implements the suggested methods for collinearity diagnostics has been provided on github (https://github.com/G-Kazantzidis/Collinearity.git). This is a free R package that implements the singular value decomposition for a predictor matrix allowing for the option of equilibration. The package also provides functions for the calculation of the variance decomposition proportions that indicate which variables are involved in collinearity. With this package, a useful tool for collinearity diagnostics for model selection approaches is now available.

In this project, we used integrated nested Laplace approximation (INLA) [Rue et al., 2009] to show the negative effects of collinearity on Bayesian models. Because this program is based on approximate Bayesian inference, convergence diagnostics are not required. Moreover, model fitting in INLA is very fast, efficiently decreasing the time of simulations. Ideally, JAGS and Stan [Stan Development Team, 2021] should also be used for the assessment of collinearity to Bayesian model choice criteria and estimates.

For the Bodyfat data, we estimated the amount of collinearity (measured as the maximal equilibrated condition index) for each possible model. In addition, we estimated all model choice criteria for each of these models. Finally, we calculated the Spearman correlation coefficients between these two. These estimates contain information which could be investigated in greater detail in the future.

Finally, we only focus on the impact of collinearity in classical and Bayesian multiple linear regression models. Since collinearity only depends on the predictors, the phenomenon is not limited to linear models. More general models such as generalized linear models or survival regression models can also include collinear predictors. Therefore, it is important that the impact of collinearity is assessed in both linear and generalized linear models.

One of the main strengths of this project is the association of identification methods of harmful patterns of collinearity along with methods of model selection. Although there are many approaches for collinearity diagnostics and also many more for model selection, the existence of one method that combines both is not yet available. Here, we have shown that the inclusion of both collinearity diagnostics and model choice criteria in model selection methods is essential, especially for models with many predictors. Also, we presented a case study were we apply both collinearity diagnostics and model selection methods in two separate steps, demonstrating the importance of each.

Secondly, we provide empirical evidence for the behavior of model choice criteria under collinearity in practice. The importance of this information is very useful for model selection methods based on model choice criteria. Apart from the behavior of model choice criteria, we provide a new handful tool for the assessment of collinearity diagnostics and collinearity comparisons between different considered models. Having proved that model choice criteria are unstable for cases of severe collinearity, increases the value of such tool that can be used from the researcher to filter out models with high levels of collinearity.

Appendix A

Appendix

- `Collinearity` R package: https://github.com/G-Kazantzidis/Collinearity.git
- Reproducible code: https://bitbucket.org/G_Kaza/master-thesis-2022/src/master/. All the contents of folder 'Final' are required for the compilation of the file. The executable file is 'MSc_Report.Rnw'. It is located in the folder named 'Final'.

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] ggpmisc_0.4.5      ggpp_0.4.3          lme4_1.1-27.1       ggforce_0.3.3
##  [5] showtext_0.9-4     showtextdb_3.0      sysfonts_0.8.5      ggtext_0.1.1
##  [9] ggrepel_0.9.1      tableone_0.13.0     shrink_1.2.1        hrbrthemes_0.8.0
## [13] viridis_0.6.2      viridisLite_0.4.0   ggridges_0.5.3      reshape2_1.4.4
## [17] latex2exp_0.5.0    scales_1.1.1        plot3D_1.4          doParallel_1.0.16
## [21] iterators_1.0.13   INLA_21.11.22       sp_1.4-6            foreach_1.5.1
## [25] Matrix_1.3-4       ggpubr_0.4.0        plotly_4.10.0       Collinearity_1.1.1
## [29] GGally_2.1.2       RColorBrewer_1.1-2  ellipse_0.4.2       forcats_0.5.1
## [33] stringr_1.4.0      dplyr_1.0.7         purrr_0.3.4         readr_2.1.1
## [37] tidyr_1.1.4        tibble_3.1.6        ggplot2_3.3.5       tidyverse_1.3.1
## [41] xtable_1.8-4       knitr_1.37
##
```

```
## loaded via a namespace (and not attached):
##   [1] utf8_1.2.2           rms_6.2-0            tidyselect_1.1.1
##   [4] htmlwidgets_1.5.4    grid_4.1.2           pROC_1.18.0
##   [7] munsell_0.5.0        codetools_0.2-18     future_1.23.0
##  [10] misc3d_0.9-1         withr_2.4.3          colorspace_2.0-2
##  [13] highr_0.9            rstudioapi_0.13      stats4_4.1.2
##  [16] ggsignif_0.6.3       Rttf2pt1_1.3.9       listenv_0.8.0
##  [19] labeling_0.4.2       polyclip_1.10-0      farver_2.1.0
##  [22] parallelly_1.30.0    vctrs_0.3.8          generics_0.1.1
##  [25] TH.data_1.1-0        ipred_0.9-12         xfun_0.29
##  [28] R6_2.5.1             biostatUZH_1.8.0     reshape_0.8.8
##  [31] assertthat_0.2.1     multcomp_1.4-18      nnet_7.3-16
##  [34] gtable_0.3.0         globals_0.14.0       conquer_1.2.1
##  [37] sandwich_3.0-1       timeDate_3043.102    rlang_0.4.12
##  [40] MatrixModels_0.5-0   systemfonts_1.0.3    splines_4.1.2
##  [43] rstatix_0.7.0        extrafontdb_1.0      lazyeval_0.2.2
##  [46] ModelMetrics_1.2.2.2 broom_0.7.11         checkmate_2.0.0
##  [49] abind_1.4-5          modelr_0.1.8         backports_1.4.1
##  [52] Hmisc_4.6-0          caret_6.0-90         gridtext_0.1.4
##  [55] extrafont_0.17       tools_4.1.2          lava_1.6.10
##  [58] tcltk_4.1.2          ellipsis_0.3.2       proxy_0.4-26
##  [61] Rcpp_1.0.7           plyr_1.8.6           progress_1.2.2
##  [64] base64enc_0.1-3      prettyunits_1.1.1    rpart_4.1-15
##  [67] cowplot_1.1.1        zoo_1.8-9            haven_2.4.3
##  [70] cluster_2.1.2        fs_1.5.2             survey_4.1-1
##  [73] magrittr_2.0.1       data.table_1.14.2    SparseM_1.81
##  [76] reprex_2.0.1         mvtnorm_1.1-3        matrixStats_0.61.0
##  [79] hms_1.1.1            evaluate_0.14        jpeg_0.1-9
##  [82] mfp_1.5.2.1          readxl_1.3.1         gridExtra_2.3
##  [85] compiler_4.1.2       crayon_1.4.2         minqa_1.2.4
##  [88] htmltools_0.5.2      tzdb_0.2.0           Formula_1.2-4
##  [91] lubridate_1.8.0      DBI_1.1.2            tweenr_1.0.2
##  [94] dbplyr_2.1.1         MASS_7.3-54          boot_1.3-28
##  [97] car_3.0-12           cli_3.1.0            mitools_2.4
## [100] gower_0.2.2          pkgconfig_2.0.3      foreign_0.8-81
## [103] recipes_0.1.17       xml2_1.3.3           prodlim_2019.11.13
## [106] rvest_1.0.2          digest_0.6.29        rmarkdown_2.11
## [109] cellranger_1.1.0     htmlTable_2.4.0      gdtools_0.2.3
## [112] curl_4.3.2           quantreg_5.86        nloptr_1.2.2.3
## [115] lifecycle_1.0.1      nlme_3.1-153         jsonlite_1.7.2
## [118] carData_3.0-5        fansi_0.5.0          labelled_2.9.0
## [121] pillar_1.6.4         lattice_0.20-45      fastmap_1.1.0
## [124] httr_1.4.2           survival_3.2-13      glue_1.6.0
## [127] png_0.1-7            class_7.3-19         stringi_1.7.6
## [130] polspline_1.1.19     latticeExtra_0.6-29  e1071_1.7-9
## [133] future.apply_1.8.1
```

List of Figures

List of Tables

Bibliography

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. 1, 3

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978. 1, 4

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: SeriesB (Statistical Methodology)*, 64(4):583–639, 2002. 1, 4

Sumio Watanabe and Manfred Opper. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12), 2010. 1, 4

LI Pettit. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):175–184, 1990. 1, 4

Seymour Gkisser. *Predictive inference: an introduction*. Chapman and Hall/CRC, 2017. 1, 5

Malgorzata Roos and Leonhard Held. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278, 2011. 1, 5

Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. 1, 5

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001. 1, 5

Virgilio Gómez-Rubio and Haavard Rue. Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051, 2018. 1, 5

Virgilio Gómez-Rubio, Roger S Bivand, and Haavard Rue. Estimating spatial econometrics models with integrated nested Laplace approximation. *Mathematics*, 9(17):2044, 2021. 1, 5

Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. Applied multiple regression. *Correlation Analysis for the Behavioral Sciences*, 2, 1983. 1

Ronald R Hocking. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. John Wiley & Sons, 2013. 1

John Neter, Michael H Kutner, Christopher J Nachtsheim, William Wasserman, et al. Applied Linear Statistical Models. 1996. 1

Barbara G Tabachnick and Linda S Fidell. Using multivariate statistics. Northridge. *Cal.: Harper Collins*, 1996. 1

Norman R Draper and Harry Smith. *Applied Regression Analysis*, volume 326. John Wiley & Sons, 1998. 1

Samprit Chatterjee, AS Hadi, and B Price. Regression analysis by example. *Inc., New York*, 2000. 1

Michael H Graham. Confronting multicollinearity in ecological multiple regression. *Ecology*, 84(11): 2809–2815, 2003. 1

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 5 edition, 2021. 1, 2, 12, 15

David A Belsley. *Conditioning diagnostics: Collinearity and Weak Data in Regression*. Number 519.536 B452. John Wiley & Sons, 1991. 1, 2, 12, 13, 14, 19, 37, 38, 53

Michael B Morrissey and Graeme D Ruxton. Multiple regression is not multiple regressions: the meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology*, 10(3), 2018. 2

Jan-Michael Becker, Christian M Ringle, Marko Sarstedt, and Franziska Völckner. How collinearity affects mixture regression results. *Marketing Letters*, 26(4):643–659, 2015. 2

Achmad Efendi and Effrihan. A simulation study on Bayesian ridge regression models for several collinearity levels. In *AIP Conference Proceedings*, volume 1913, page 020031. AIP Publishing LLC, 2017. 2

Roman Salmerón Gómez, José García Pérez, María Del Mar López Martín, and Catalina García García. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. *Journal of Applied Statistics*, 43(10):1831–1849, 2016. 2

Habshah Midi, Saroje Kumar Sarkar, and Sohel Rana. Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3):253–267, 2010. 2

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection: a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018. 2, 17, 19, 35, 51, 53, 61

Federico Pavone, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Using reference models in variable selection. *arXiv preprint arXiv:2004.13118*, 2020. 2, 17, 19, 51, 53, 61

Xavier Basagaña and Jose Barrera-Gómez. Reflection on modern methods: visualizing the effects of collinearity in distributed lag models. *International Journal of Epidemiology*, 2021. 2

M. U. Imdad and M. Aslam. *mctest: Multicollinearity Diagnostic Measures*, 2020. URL https://CRAN.R-project.org/package=mctest. R package version 1.3.1. 2

M. Imdadullah, M. Aslam, and S. Altaf. mctest: An R package for deteection of collinearity among regressors. *The R Journal*, 8(2):499–509, 2016. URL https://journal.r-project.org/archive/2016/RJ-2016-062/index.html. 2

M. U. Imdad, M. Aslam, S. Altaf, and A. Munir. Some new diagnostics of multicollinearity in linear regression model. *Sains Malaysiana*, 48(9):2051–2060, 2019. URL http://dx.doi.org/10.17576/jsm-2019-4809-26. 2

Chen Lin, Kevin Wang, and Samuel Mueller. mcvis: A new framework for collinearity discovery, diagnostic and visualization. *Journal of Computational and Graphical Statistics*, In Press, 2020. doi: 10.1080/10618600.2020.1779729. 2

Claudia García, Román Salmerón Gómez, and Catalina B García. Choice of the ridge factor from the correlation matrix determinant. *Journal of Statistical Computation and Simulation*, 89(2):211–231, 2019. 2

Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013. 4

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. 4, 5

Mervyn Stone. An asymptotic equivalence of choice of model by cross–validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47, 1977. 5, 6

Roger W Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1), 1996. 17

Karline Soetaert. *plot3D: Plotting Multi–Dimensional Data*, 2021. URL https://CRAN.R-project.org/package=plot3D. R package version 1.4. 19

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org. 19

Erich Neuwirth and R Color Brewer. ColorBrewer palettes. *R package version*, 1, 2014. 19

Kazuki Yoshida and Alexander Bartel. *tableone: Create Table 1 to Describe Baseline Characteristics with or without Propensity Score Weights*, 2020. URL https://CRAN.R-project.org/package=tableone. R package version 0.12.0. 19

David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. URL https://CRAN.R-project.org/package=xtable. R package version 1.8-4. 19

Sarah R. Haile, Leonhard Held, Sebastian Meyer, Sina Rueeger, Kaspar Rufibach, and Simon Schwab. *biostatUZH: Misc Tools of the Department of Biostatistics, EBPI, University of Zurich*, 2020. URL https://R-Forge.R-project.org/projects/ebuzh/. R package version 1.8.0/r99. 19

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Fran**c**ois, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the `tidyverse`. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686. 20

Hadley Wickham and Dana Seidel. *scales: Scale Functions for Visualization*, 2020. URL https://CRAN.R-project.org/package=scales. R package version 1.1.1. 20

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3. 20

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL https://CRAN.R-project.org/package=ggpubr. R package version 0.4.0. 20

Carson Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall CRC, 2020. ISBN 9781138331457. URL https://plotly-r.com. 20

Stefano Meschiari. *latex2exp: Use LaTeX Expressions in Plots*, 2021. URL https://CRAN.R-project.org/package=latex2exp. R package version 0.5.0. 20

Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12): 1–20, 2007. URL http://www.jstatsoft.org/v21/i12/. 20

Claus O. Wilke. *ggridges: Ridgeline Plots in 'ggplot2'*, 2021. URL https://CRAN.R-project.org/package=ggridges. R package version 0.5.3. 20

Simon Garnier. *viridis: Default Color Maps from 'matplotlib'*, 2018. URL https://CRAN.R-project.org/package=viridis. R package version 0.5.1. 20

Bob Rudis. *hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'*, 2020. URL https://CRAN.R-project.org/package=hrbrthemes. R package version 0.8.0. 20

Haavard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. 54

Stan Development Team. RStan: the R interface to Stan, 2021. URL https://mc-stan.org/. R package version 2.21.3. 54