Draft: Novel Approach to Knot Placement for Natural Spline Regressions

Jo Inge Arnes, Tonje Braaten, and Alexander Horsch

January 2023

1 Introduction

For natural cubic spline regressions, the two most common approaches to placing knots are by uniform distances or equal-sized quantiles along the independent variable—conventions unrelated to inherent properties of the sample data. For example, biological reasons rarely dictate that the relationship between an independent and a dependent variable must align with uniform distances or quantiles. Nevertheless, small quantiles result in knots being close to each other, especially in the denser regions of the independent variable's distribution. If the knots are sufficiently close, regressions can readily fit a model to the sample data closely. However, the issue then becomes a high risk of overfitting, meaning that the model matches one sample well but not the other samples from the population in general. Therefore, keeping the number of knots low is desirable, yielding models that fit less exactly but generalize better.

Several measures for estimating the relative quality of a model exist that favor lower numbers of knots. Examples are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Unfortunately, when we are limited to placing knots at uniform distances or quantiles, having a low number of knots may fail to include locations essential to fit the model to the underlying curve's turning points and curvatures. Furthermore, we risk placing knots in positions that do not substantially improve the fit or can contribute to overfitting in denser subintervals.

Against this background, we present a novel approach to placing knots for natural spline regression models. The approach empirically shows improved results compared to placing knots at uniform distances or quantiles for a comparable knot count.

2 Novel approach to placing knots

We now describe our novel approach to choosing knot locations for natural cubic splines, which has two main steps. The first step is to find a model with a high number of knots that closely fits the sample data. The second step is to systematically remove knots until an acceptable low count has been reached. The aim is to find a better model than we get from placing knots by quantiles for a low number of knots.

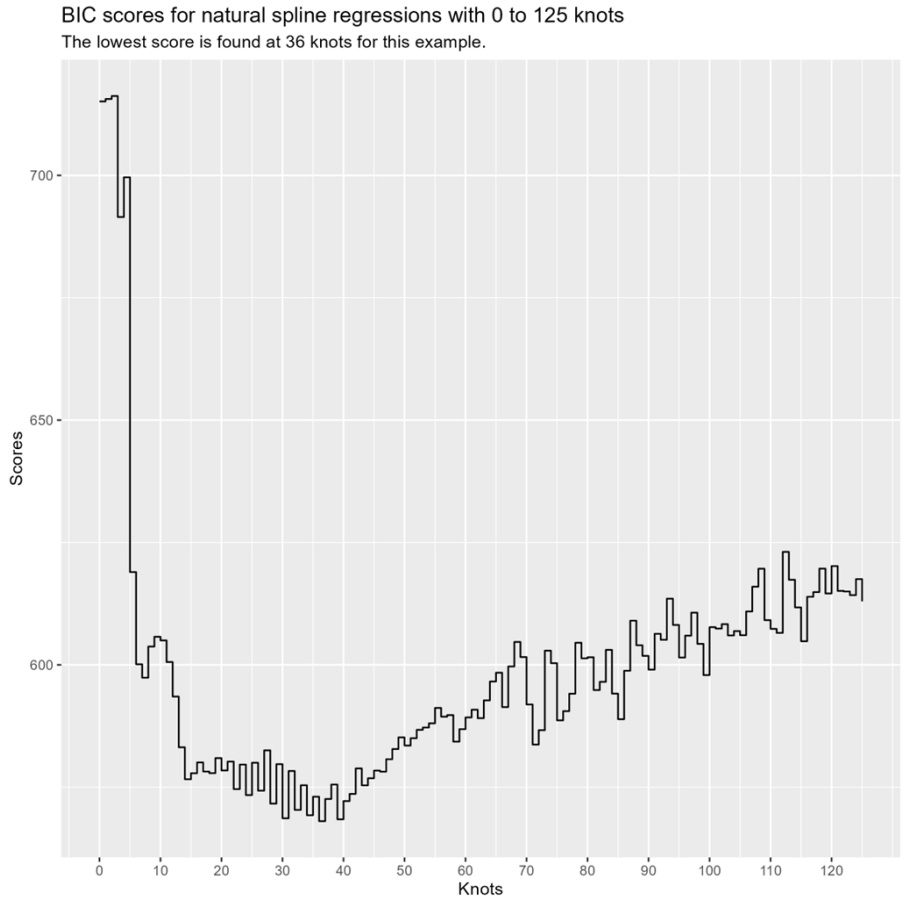
2.1 Finding a start model

In the first step, we compare a series of models with zero to a relatively high number of knots, placed by quantiles, to find a good starting model. For example, we can set the highest number of knots equal to $k = (n/2)$ rounded down to the nearest integer. A quantitative measure for estimating the models' relative quality, which also considers the knot count, is used to compare the models. AIC is one such measure that can be used, BIC is another. We have chosen to use BIC, because it more strongly penalizes higher knot counts than AIC. For both AIC and BIC, lower scores are better.

Figure X shows how increasing the number of knots typically leads to progressively lower BICs, with several local minima, before reaching a global minimum. Beyond this number of knots, the BIC starts increasing gradually. We choose the model with the lowest BIC as our starting model.

Some of the knots in the model will likely be closer to the underlying optimal knot locations than other knots in the same model. However, because the quantiles used to place knots are tied to a smaller sample from a larger population, we are unlikely to have found the mathematically optimal locations for the underlying curve for the population. The optimal location may be at a value not present in our sample. Our start model additionally has a relatively high number of knots and is likely overfitted. We

still consider this model a good starting point for improving the model through reducing the number of knots.



2.1.1 Alternative start models

Alternative approaches to finding a start model were tried and gave inconsistent results compared to the above-described method. For example, we can find a start model based on uniform distances between knots. If the distances are short enough, the curve can closely match the sample data. Such short, equally distanced knots can lead to empty subintervals between knots, even for uniform distributions. Thus, empty subintervals are merged with non-empty ones. Another alternative is letting the knots of the start model coincide with the distinct values for the independent variable. We also tried using uniform sized quantiles with higher numbers of knots without the process of finding the model with lowest BIC, e.g., by setting number of knots to half of the number of distinct values for the independent variable. This method of finding a start model generally showed worse results than the other three alternatives. An issue for these three approaches is non-convergence, which often happens for high number of knots due to low variability for the dependent variable in some regions of the independent variable.

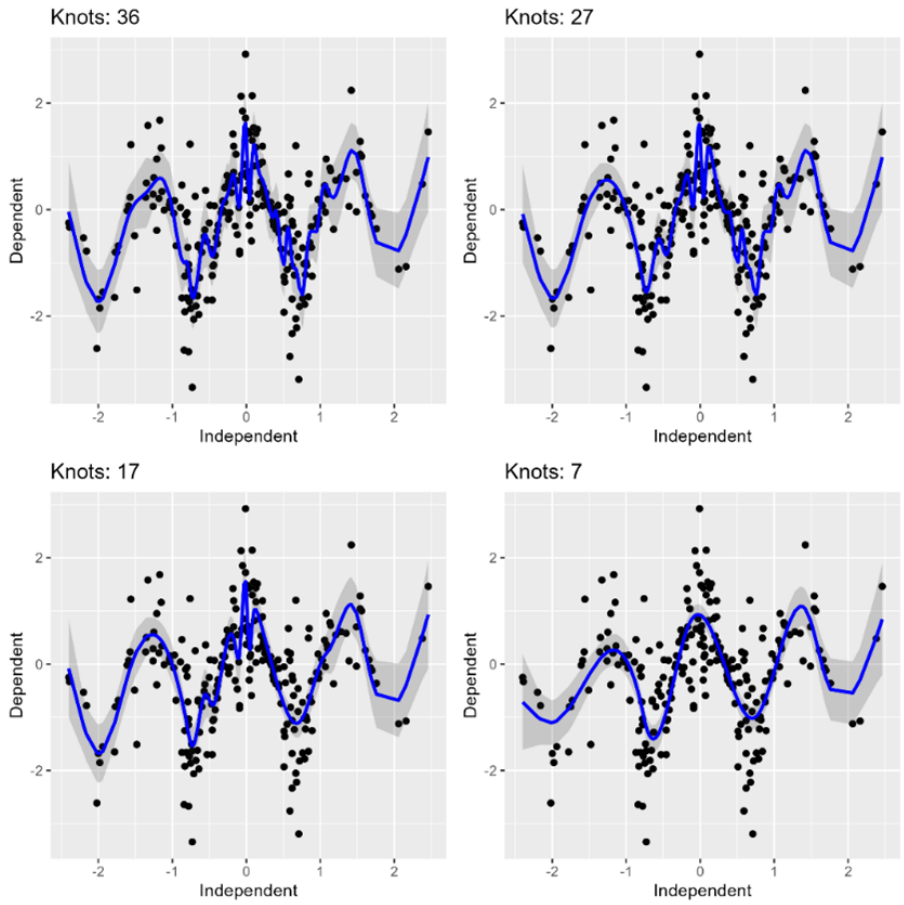
2.2 Systematically removing knots

After generating a start model, the next step is to systematically reduce the number of knots. A lower number of knots reduces the risk of overfitting. The rationale behind the approach is that some of the start model's knots probably are located relatively close to positions needed for a good fit to the underlying curve, given it has a relatively high knot count and the best BIC for knots placed by quantiles, for the given sample. At the same time, we assume that several of the knots in the start model are redundant, since their locations are not based on any relationships between the independent and dependent variables.

The aim is to remove knots in order of redundancy. We do this through an iterative process where a single knot is removed per step. For each step, we remove the knot that leaves the best model score compared to the other alternatives. Here, we are comparing models with equal knot counts. Therefore, it may be relevant to consider other measures than when models with different numbers of knots are

compared. We use AIC as default because the alternative measures such as RSS and F-statistics experimentally gave worse results. The iteration continues until all knots have been removed, not counting the two boundary knots. A maximum knot count, k_{max}, has been set in advance, and from the models with a knot count $k \leq k_{max}$, the one with the lowest BIC is chosen.

Figure Y shows the effect of gradually removing individual knots.



3 Data generator

A data generator was designed and implemented for generating the synthetic data sets used in the experiments. It produces pseudorandom samples by applying three user-defined functions:

- The independent variable's statistical distribution
- The curve for the dependent variable's population means
- The variance, or error distribution, around the dependent variable's means

First, the data generator draws a sample for the independent variable, X, from a given statistical distribution. For repeatability, the user can optionally set the seed used internally by the pseudorandom number generator. Next, the generator computes values representing the population mean for the dependent variable, Y, at each value of X in the sample. These values lie on a curve defined by a mathematical function taking values of X as an argument. Lastly, the generator computes the values for Y by adding the variance or error, E, to the means. The error values are drawn from a user-defined distribution, which alternatively can be heteroscedastic. For example, we can scale the error distribution's variance by a factor of the given value for X, which can be relevant for ratio-valued variables. Finally, the X and Y values are rounded to a chosen accuracy, simulating the limits of the measurement method. The resulting data set includes both the rounded and unrounded values for X, Y, and E.

4 Method

4.1 Comparing information criterion scores

We start the experiment by generating a set containing $m = 500$ samples of size $n = 250$. Then, the standard and the novel approach to placing knots are applied to each sample, denoted as A_0 and A_1, respectively. Next, the information criterion scores for the two resulting models are computed, in our case, BIC. We group the results for the sample set into three categories by the pairwise difference in BIC means, $BIC_{A_0} - BIC_{A_1}$, where scores that differ $\leq |1|$ are considered equally good:

- Better (B)
- Equally good (E)
- Worse (W)

Next, we test the BIC scores for A_0 and A_1 for normality with a Shapiro-Wilk test. If passed, a two-tailed paired samples t-test for the mean is conducted with $alpha = 0.05$. The null hypothesis, H_0, is that there is no difference between A_0 and A_1. If the normality test is not passed, we instead use bootstrapping to compute 95% CI for the difference in means for A_0 and A_1 with $R = 1000$ bootstrap replicates.

4.2 Correlations between BIC and actual goodness of fit

When deciding which model to use, we generally make a choice based on the information criterion score. The question is then how well the score indicates the actual quality, or goodness of fit, to the true underlying curve function. For example, an overfitted model can have a better information criterion score than a model that matches the underlying curve better. Information criterion scores, such as BIC, additionally favor lower knot counts. However, in our case, we do not need further to adjust the quality measure to different knot counts because all candidate models are within the acceptable range. Mean squared error (MSE) is used as the goodness of fit measure in the paper:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{1}$$

Two MSEs are computed per sample in the experiments, denoted as MSE_1 and MSE_2. MSE_1 is the goodness of fit to the true curve given the original sample's observations. We compute the MSE between the model's predicted values, $\hat{Y}$, and the corresponding values for the underlying true curve, Y. MSE_2, is the same for, $n = 500$, evenly spaced values for the independent variable, instead of the original sample observations. The interval is defined by the original sample's lower and upper independent variable observations, to avoid predicting dependent variable values for any independent variable values outside the original sample's boundaries. The reason behind the use of two MSEs, is the distinction between approximating the curve function and predicting an outcome. The values for independent variables are usually not uniformly distributed, so the samples have a higher density of observations in some regions than others. Consequently, a model that has a good fit to the underlying curve in the denser regions can predict outcomes for the dependent variable well *on average*. We can add more knots in the denser regions to achieve this, which is what happens when quantiles are used as a knot placement strategy. The splines can then more closely adjust to the curve of the observed values in these denser regions. This comes at a higher risk of overfitting in those regions as well as worse curve approximation in sparser regions.

In the experiments, the regression models for A_0 and A_1 per single sample are treated as a pair. The spline regressions for A_0 and A_1 are pairwise applied to each sample in the sample set. The difference between the MSE for the model pairs for A_0 and A_1 are denoted $MSE_{A0} - MSE_{A1}$. A pair of models can have unequal numbers of knots, but because all knot counts are within an acceptable range, $0 \leq k \leq 4$, we do not penalize higher counts. The relationship between differences in MSE and differences in BIC in our experiments should in the ideal case also be linear in our correlation experiments.

The Pearson correlation coefficients, r, are reported together with accompanying plots showing the relationship between differences in BIC and differences in MSEs. If the models are not overfitted, the correlations should be relatively strong.

5 Experiments and Results

5.1 Distributions, variance, and accuracy

In our experiments, the independent variable, X, is lognormally distributed (Eq. 2). The variance for the dependent variable, Y, is homoscedastic and normally distributed (Eq. 3). We did make initial experiments with heteroscedasticity, but we did not experience any altered results with respect to the conclusions of the experiments. Lastly, the values for X and Y in the sample are rounded to simulate an accuracy of two decimal places.

$$X \sim Lognormal(0.3, 0.4) \tag{2}$$

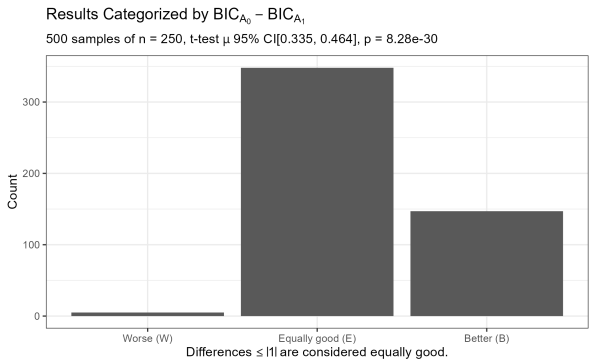
$$Y \sim N(0, 0.1) \tag{3}$$

5.2 Curve functions

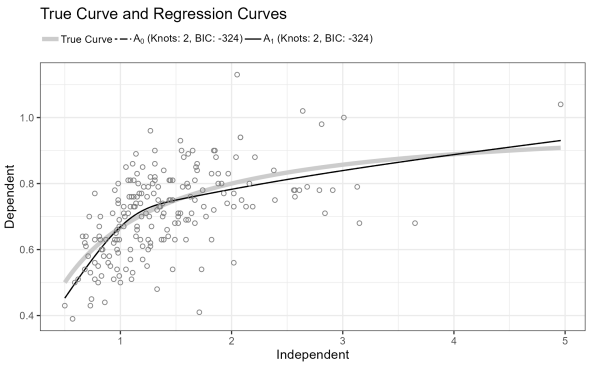
In this section, we report results for five different curve functions. Different types of curve functions were assessed in the experiments. Our approach showed significantly better results for some cases, and non-significant differences for others. None of the results were significantly worse. There is not much difference between A_0 and A_1 for the first two reported curve functions. However, the improvements are more evident for the last three curve functions.

5.2.1 Yield-Loss

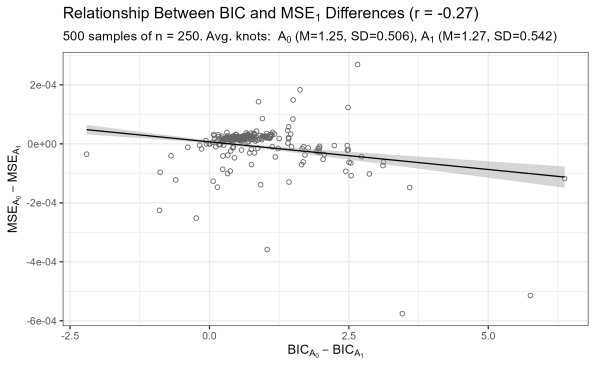
$$f(x) = \frac{x}{\frac{1}{2} + x} \tag{4}$$



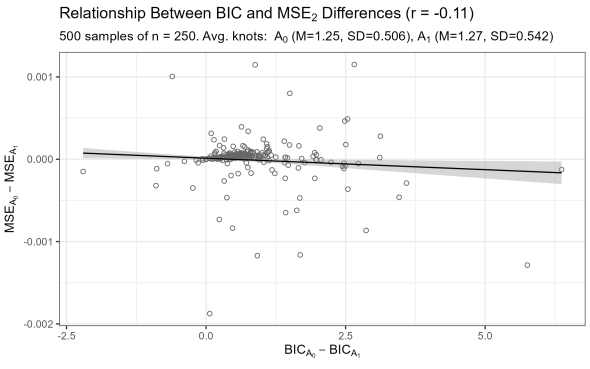
(a) The differences in BIC scores categorized



(b) The A_0 and A_1 curve approximations are identical



(c) The differences are small, which affects the BIC and MSE_1 diff. correlation



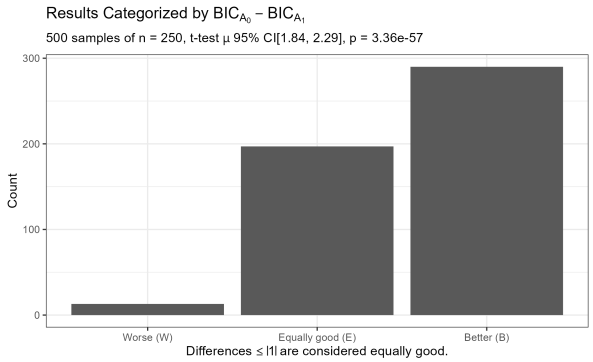
(d) The differences are small, very low BIC and MSE_2 diff. correlation

Figure 1: The figure illustrating the Yield-Loss function experiment results

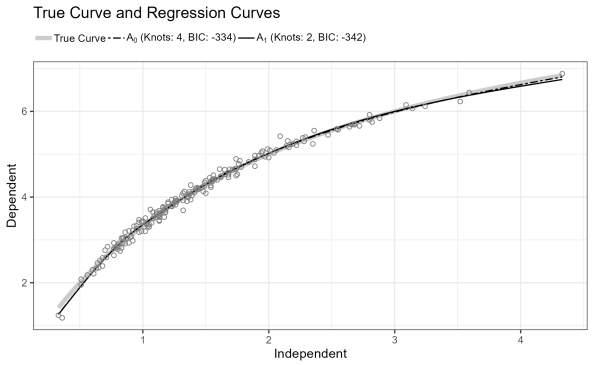
5.2.2 Michaelis-Menten

Michaelis-Menten equation with fictitious parameter values: $K_m = 2.00\mu M$, and $V_{max} = 10.00\mu mol\,s^{-1}$.

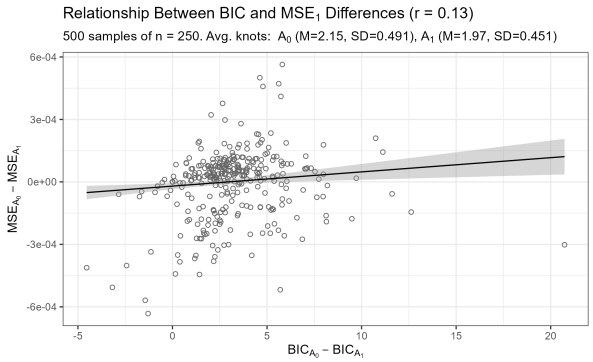
$$f(x) = \frac{V_{max}x}{K_m + x} \tag{5}$$



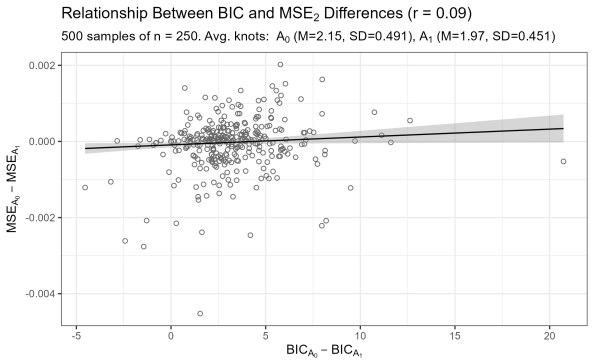
(a) The differences in BIC scores categorized



(b) Almost similar curves, but A_1 has two less knots



(c) Low difference in true fit, because lower knot counts is the reason for A_1's lower BIC scores



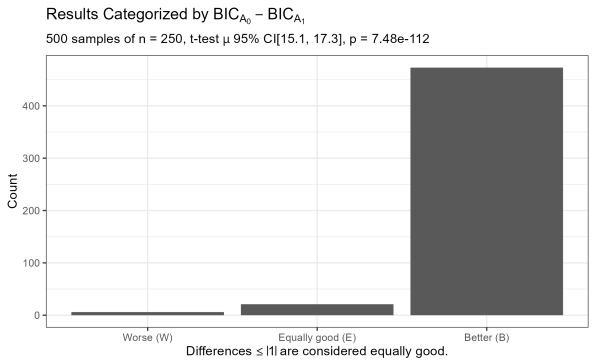
(d) Here, we see the same as for 2c

Figure 2: The Michaelis-Menten function experiment results

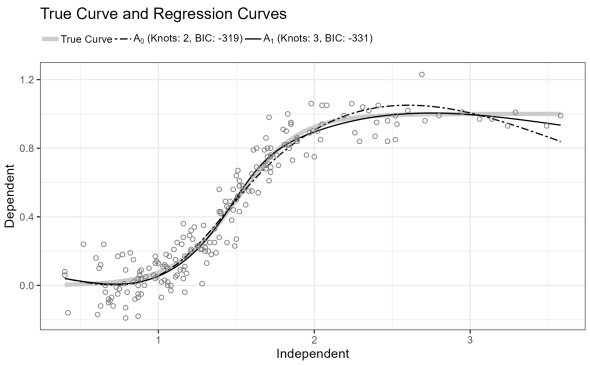
The curves for the regressions models are almost similar but knot counts are lower for our approach.

5.2.3 Logistic

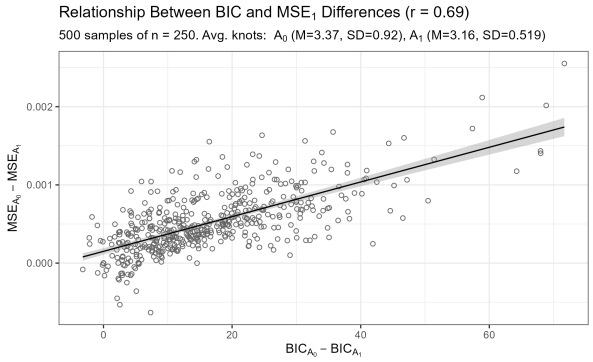
$$f(x) = \frac{1}{1 + \exp(-\frac{x-\frac{3}{2}}{\frac{1}{5}})} \tag{6}$$



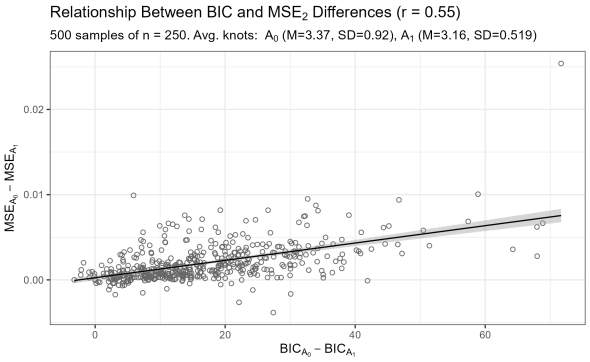
(a) The differences in BIC scores categorized



(b) The A_1 curve is better in this example



(c) The correlation is stronger in this case

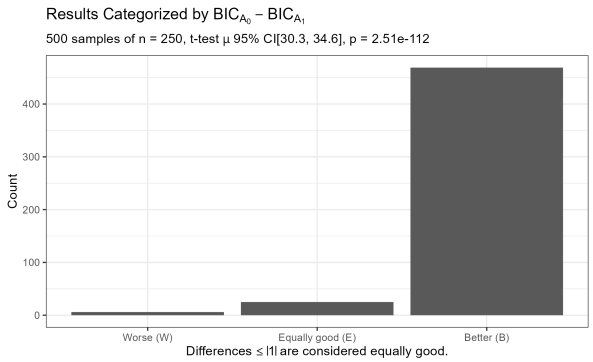


(d) The findings are in accordance with 3c

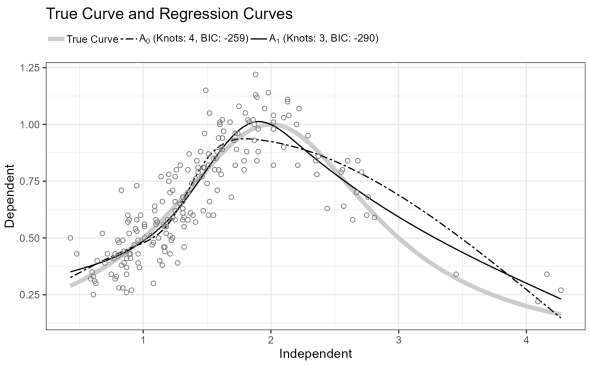
Figure 3: The figure illustrating the logistic function experiment results

5.2.4 Runge

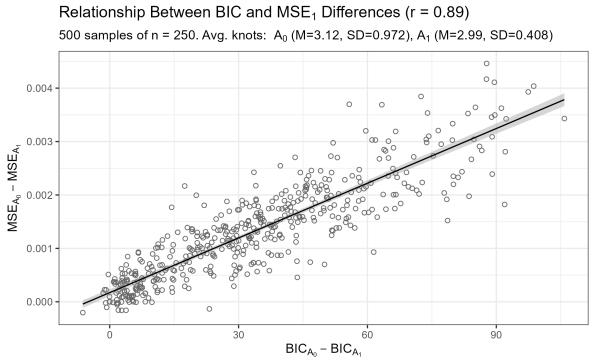
Runge functions are commonly used to demonstrate the Runge's phenomenon, but here the curve is translated so that the central peak is at $x = 2$.



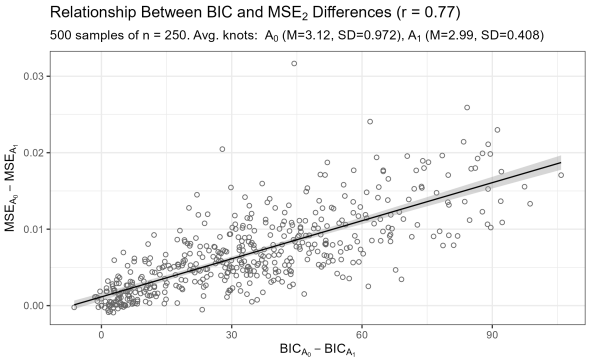
(a) The differences in BIC scores categorized



(b) Here, the A_1 curve is better



(c) Very strong correlation



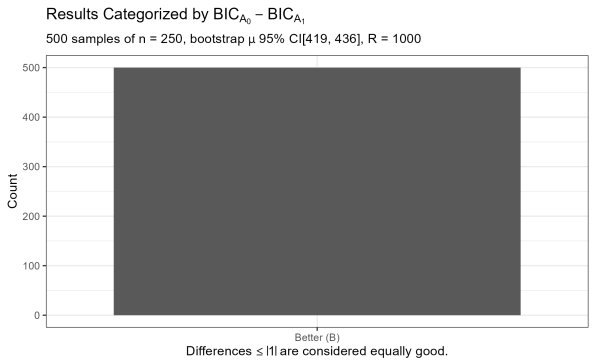
(d) The findings are in accordance with 4c

Figure 4: The figure illustrating the Runge function experiment results

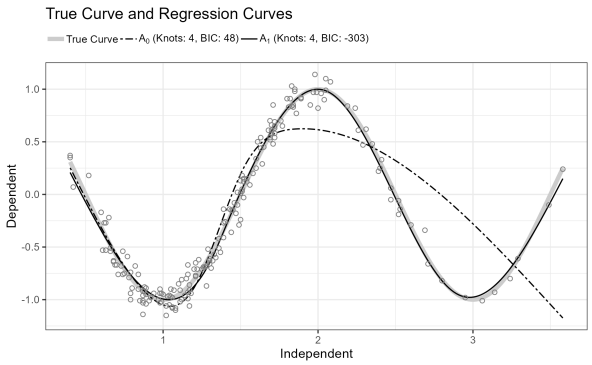
5.2.5 Trigonometric

A simple cosine function

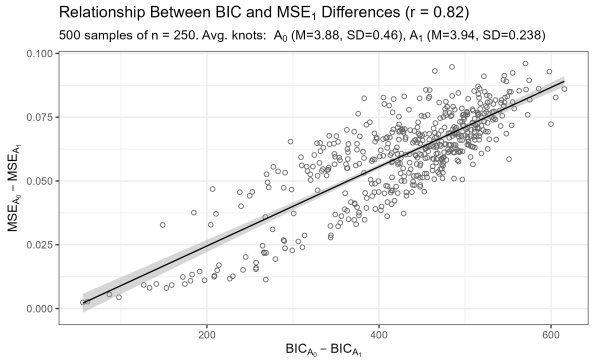
$$f(x) = \cos(\pi x) \tag{7}$$



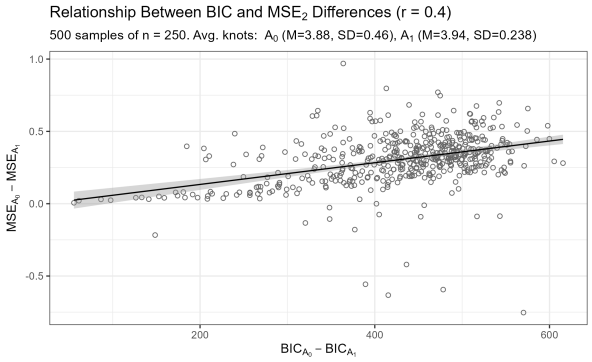
(a) A_1 has better BIC score for all samples



(b) The A_1 curve approximations is better



(c) The correlation is very strong between a better BIC and a better fit for MSE_1



(d) The correlation for MSE_2 is a bit weaker, but the plot shows that both BICs and MSE_2 are better for A_1

Figure 5: The figure illustrating the trigonometric function experiment results

6 Discussion

7 Conclusion

8 References