

The value of Bayesian predictive projection for variable selection: An example selecting lifestyle predictors of young adult well-being

Adam Bartonicek (✉ adam.bartonicek@postgrad.otago.ac.nz)

University of Otago <https://orcid.org/0000-0002-4528-5543>

Shay Ruby Wickham

University of Otago

Narun Pat

University of Otago

Tamlin S. Conner

University of Otago

Research article

Keywords: Variable selection, Prediction, Inference, Psychological well-being, Young adults, Health habits, Sleep, Diet, Exercise

Posted Date: January 11th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-28472/v2>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Public Health on April 9th, 2021. See the published version at <https://doi.org/10.1186/s12889-021-10690-3>.

Abstract

Background

Variable selection is an important issue in many fields such as public health and psychology. Researchers often gather data on many variables of interest and then are faced with two challenging goals: building an accurate model with few predictors, and making probabilistic statements (inference) about this model. Unfortunately, it is currently difficult to attain these goals with the two most popular methods for variable selection methods: stepwise selection and LASSO. The aim of the present study was to demonstrate the use predictive projection feature selection – a novel Bayesian variable selection method that delivers both predictive power and inference. We apply predictive projection to a sample of New Zealand young adults, use it to build a compact model for predicting well-being, and compare it to other variable selection methods.

Methods

The sample consisted of 791 young adults (ages 18 to 25, 71.7% female) from New Zealand who had taken part in the Dunedin Daily Life Study in 2013-2014. Participants completed a 13-day online daily diary assessment of their well-being and a range of lifestyle variables (e.g., sleep, physical activity, diet variables). The participants' diary data was averaged across days and analyzed cross-sectionally to identify predictors of average flourishing. Predictive projection was used to select as few predictors as necessary to approximate the predictive accuracy of a reference model with all 28 predictors. Predictive projection was also compared to other variable selection methods, including stepwise selection and LASSO.

Results

Three predictors were sufficient to approximate the predictions of the reference model: higher sleep quality, less trouble concentrating, and more servings of fruit. The performance of the projected submodel generalized well. Compared to other variable selection methods, predictive projection lead to models with either matching or slightly worse performance but with much fewer predictors.

Conclusion

Predictive projection was used to efficiently arrive at a compact model with good predictive accuracy. The predictors selected into the submodel – felt refreshed after waking up, had less trouble concentrating, and ate more servings of fruit – were all theoretically meaningful. Our findings have important implications for applications of variable selection in health research.

Background

Variable selection is an important topic in public health, well-being, and other fields. Researchers often collect data on large numbers of variables (predictors) that are plausibly related to the outcome of

interest, and then try to find an optimal subset of the predictors that can maximally predict the outcome measure (Heinze et al., 2018; Höge et al., 2018; Müller et al., 2013). For example, in well-being research, researchers may collect data on many demographic, lifestyle, and psychological variables, and then look to build a compact model with fewer variables that can accurately predict the participants' self-reported well-being. Importantly, the model should not only explain the data at hand well, but its performance should generalize to new, out-of-sample data. Additionally, researchers are typically not only after an accurate "black box" model that predicts well but is hard to understand, but instead want to be able to interpret the model and make probabilistic statements about it – how uncertain is the selection, how variable is the model's performance, and, perhaps most importantly, how strong and reliable are the relationships between the selected predictors and the outcome (Taylor & Tibshirani, 2015). Ideally, variable selection should simultaneously deliver on these two goals: predictive power and inference.

Variable selection is necessary because the exact set of predictors to be included in the model is rarely known. A key assumption in traditional regression modelling is that the set of predictors to be included in the model is known a-priori. However, across many fields, the theory is rarely as strong as to determine the set of predictors to include in the model exactly, and there are choices to be made (Heinze et al., 2018; Kadane & Lazar, 2004; van Erp et al., 2019). Under favourable circumstances, variable selection can be done by fitting all possible models and keeping the best one. This is possible when the number of candidate predictors is small, however, with many predictors, this best-subset approach quickly becomes computationally expensive – with k predictors, there are 2^k models that need to be fitted and evaluated (e.g. 1,024 models with 10 predictors, 1,048,576 models with 20 predictors; Heinze et al., 2018; James, Witten, Hastie, & Tibshirani, 2013). Specialized variable selection methods are thus necessary to reduce the computational burden and make the task of finding the optimal subset of predictors possible.

Currently, there are two widely used variable selection methods: stepwise selection (Efroymson, 1960), and the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996). Stepwise selection is the more traditional method of the two, and is characterized by building up the model step-by-step, each time either adding or subtracting a predictor based on some pre-specified criterion (typically either p -values or information criteria such as the AIC or BIC; Whittingham, Stephens, Bradbury, & Freckleton, 2006). LASSO is a newer, popular machine learning method, in which models are typically fitted via the efficient least-angle regression (LARS) algorithm (Efron et al., 2004). LARS resembles forward stepwise regression except that instead of adding or removing predictors wholesale, the predictor slopes are continuously increased towards their least-squares solution. A key additional feature of LASSO is that the method places a constraint or a limited "budget" on the sum of the absolute predictor slopes (Tibshirani, 2019), via the penalization parameter λ . The constraint on the sum of the predictor slopes produces shrinkage – the slopes are shrunk towards zero when compared to classical least squares estimates, and this in turn leads to better out-of-sample predictive performance (Zou & Hastie, 2005). The optimal value of λ is typically tuned through k -fold cross-validation – a procedure in which the data are split into k parts, and then $k - 1$ parts are repeatedly used for model fitting, with the leftover part always used for model evaluation (Arlot & Celisse, 2010; Hawkins et al., 2003; Koul et al., 2018).

Unfortunately, at the present moment, it is difficult to obtain predictive power and inference simultaneously with stepwise selection and LASSO. Stepwise selection has major issues when it comes to predictive power because it suffers from *overfitting* – it is liable to select predictors that fit to pure noise in the data, and shows poor out-of-sample predictive performance as a result (Flom & Cassell, 2007; James et al., 2013; Thompson, 1995; Whittingham et al., 2006). Overfitting is a well-known problem in machine learning, and occurs because extra predictors always make models more flexible and allow them to fit the data better (Hawkins, 2004; Helwig, 2017). Therefore, even completely irrelevant predictors can ostensibly improve a model's fit. Overfit models seem to explain the data at hand well (as indicated by, e.g., high R^2 , low RMSE, and good performance metrics overall), however, this performance does not generalize, and so overfit models end up performing poorly when faced with new data. Importantly, some authors have argued that overfitting may be one of the less-known causes of replication crisis, in that some research findings may fail to replicate not because of any ill-will of the researchers but because they are supported by fragile trends idiosyncratic to the training data (Yarkoni & Westfall, 2017). Overfitting is especially a problem with stepwise selection when p-values are used as criterion for adding or subtracting predictors (Flom & Cassell, 2007; James et al., 2013; Thompson, 1995; Whittingham et al., 2006). In contrast, LASSO is set up to actively combat overfitting – by penalizing the sum of the absolute slopes and using cross-validation to determine the optimal degree of shrinkage, LASSO models produce more restrained predictions that generalize well even to new data. However, both stepwise regression and LASSO are lacking when it comes to inference. In stepwise selection, classical inferential tools such as p-values and confidence intervals are sometimes used to summarize the final model – however, these are invalid because they ignore the selection procedure (Flom & Cassell, 2007; Kuhn & Johnson, 2013; Smith, 2018; Steyerberg et al., 2001; Whittingham et al., 2006). In LASSO, classical inferential tools are not available by default, and while some methods for inference with have been developed recently, they are fairly involved, and the area is still undergoing development (Benjamini, 2010; Taylor & Tibshirani, 2015).

Predictive projection feature selection is a novel method for variable selection within the Bayesian framework (Piironen et al., 2018) that delivers both predictive performance and inference. The method consists of two steps. At the first step, a flexible reference model is fitted, using all available candidate predictors. At the second step, smaller submodels are sequentially fitted to approximate the reference model's predictions using projection. After that, it is possible to simply select the smallest possible submodel which makes predictions "similar enough" to those of the reference model (i.e., with performance matching the reference model within some uncertainty bound, such as one standard error; Piironen et al., 2018). To avoid overfitting, the submodels are compared to the reference model on cross-validated prediction accuracy, via the efficient Bayesian approximation of leave-one-out cross-validation: Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO; Vehtari, Gelman, & Gabry, 2015; 2017). There are three key advantages to the predictive projection feature selection method: it selects a parsimonious model with good predictive accuracy, is robust to overfitting, and produces a valid posterior distribution that can be used for inference just like in any other Bayesian posterior (Piironen et al., 2018).

Applying Bayesian predictive projection: Predictors of well-being

In the present paper, we applied predictive projection feature selection on a dataset related to well-being. The issue of variable selection arises frequently in well-being research. Determining which factors are associated with greater well-being is important because there is robust evidence showing that psychological well-being is linked with slew of positive outcomes including better physical health and longevity (Diener & Chan, 2011; Kern et al., 2014; Keyes, 2007; Keyes et al., 2010; Keyes & Simoes, 2012; Pressman & Cohen, 2005). Past research has identified predictors of well-being on many levels, including socioeconomic trends and policies, such as population density, income inequality, and strength of social welfare systems (Diener et al., 2018; Huppert & So, 2013; Li & Kanazawa, 2016; Oishi et al., 2011; Tay & Kuykendall, 2013), community and family factors, such as childhood socioeconomic status (SES; Piquart & Sörensen, 2000) Particularly interesting to well-being researchers are individual-level lifestyle factors such as sleep, diet, and exercise, as these can be modified and can thus be target for intervention studies. Among these health habits, sleep quality has been consistently shown as one of the strongest predictors of well-being (Kawada et al., 2009; Pilcher et al., 1997; Prendergast, Schofield, et al., 2016; Ridner et al., 2016; Wallace et al., 2017). Diet quality has also been linked with well-being – specifically, fruit and vegetable consumption has been shown to be associated with greater well-being in observational (Adams & Colner, 2008; Blanchflower et al., 2013; Conner et al., 2015; Jacka et al., 2010; Moreno-Agostino et al., 2019; Piqueras et al., 2011; Prendergast, Schofield, et al., 2016), prospective (Jacka et al., 2011), micro-longitudinal or daily diary (Jacka et al., 2011; Liao et al., 2018; Wahl et al., 2017; White et al., 2013) and experimental studies (Conner et al., 2017; Jacka et al., 2017). Conversely, poor diet as indicated by soft drink and fast-food consumption has also been linked to lower well-being (Hong & Peltzer, 2017; Moreno-Agostino et al., 2019; Prendergast, Schofield, et al., 2016). Finally, exercise has been linked to greater well-being (Diener et al., 2018; Goodwin, 2003; Hassmén et al., 2000; Prendergast, Schofield, et al., 2016). The question arises though, which of these lifestyle factors, or possibly other demographically related factors (such as BMI, SES, age, gender, etc.) offers the most compact and efficient model predicting well-being?

In the present study, we demonstrate the use of predictive projection feature selection in constructing a compact model for predicting well-being, using data from a sample of 791 young adults. The participants took part in the 2013 and 2014 waves of the Daily Life Study, a daily diary study of the health and well-being of young adults in Dunedin, New Zealand. Well-being was surveyed every day for 13 days using the 8-item Flourishing Scale (Diener et al., 2010) adapted for daily measurement. We first fitted a Bayesian multiple regression reference model, predicting average daily flourishing cross-sectionally from 28 candidate well-being predictors, that included demographic and background variables (e.g., age, gender, childhood SES, BMI), and an extensive range of lifestyle variables assessed in the daily diary related to stress (stress, most stressful event of the day), somatic symptoms (e.g. tiredness, lack of ability to concentrate), diet-related variables (e.g. fruit and vegetable consumption), and health habit variables (e.g. sleep quality/quantity, exercise). After that, we used predictive projection to find a smallest possible submodel that would predict average daily flourishing almost as well as the reference model. Finally, we compared predictive projection to stepwise selection and LASSO.

Method

Software and packages

Throughout our entire workflow, we used the programming language R, version 3.6.3 (R Core Team, 2019). For general-purpose Bayesian modelling, we used the brms package (Bürkner, 2017), which provides an interface to the state-of-the-art Bayesian statistical programming language Stan (Carpenter et al., 2017). For predictive projection feature selection, we used the projpred package (Piironen et al., 2018). For additional packages used, see Appendix A.

Participants and procedure

Participants came from the second two waves (2013/2014) of the Daily Life Study, a daily diary study assessing the psychological well-being and daily health habits in a large sample of young adults living in Dunedin, New Zealand (total $n = 821$, 71.7% female). Daily Life Study was run across four years from 2011 to 2014, however, we selected participants from the 2013/2014 waves only because the information that was collected changed over time and fewer variables of interest were collected during the 2011/2012 waves (for example, no diet-related variables). Most of the participants were University of Otago students, and all were between 17-25 years old ($m = 19.73$, $sd = 1.73$). Majority of the participants were New Zealand European/pākehā (77.5%), followed by Asian (10.6%), and Māori/Pacifica participants (5.4%). Participants of all other ethnicities made up 6.4% of the sample and were aggregated into one category. The participants had first completed an initial survey on demographics, and then were tracked across 13 days via an online daily diary completed each night between 3-8 pm. Participants also attended a single clinic visit during which height and weight was recorded and used to compute BMI.

Measures, data cleaning, and preprocessing

A list of measures with reliabilities and descriptive statistics can be found in the Supplementary Materials. We used the following demographic variables ($n = 5$) from the initial survey: age, gender, ethnicity, self-reported childhood socioeconomic status (SES; Griskevicius, Delton, Robertson, & Tybur, 2011), and a measure of body mass index (BMI). From the daily survey, we took the daily flourishing scale items (Diener et al., 2010), as well as all stress self-assessment, somatic self-assessment, diet-related, and health habit variables (total $n = 22$) that were measured for the entirety of 2013-2014 and could be theoretically linked with flourishing. While most of the 22 variables assessed the participants' state on the day of reporting, some of the variables also assessed the participants' state on the night before. Of the 22 variables, the stress self-assessment variables ($n = 2$) were: felt stressed today, most stressful event today. The somatic self-assessment variables ($n = 5$) were: felt tired today, felt rundown today, felt cold or flu today, had hangover today, and had trouble concentrating today. The diet-related ($n = 11$) variables were: servings of fruit today, servings of vegetables today, servings of sweets today, servings of soft drink today, servings of chips today, servings of fruit last night, servings of vegetables last night, servings of sweets last night, servings of soft drink last night, servings of chips last night, and standard drinks of alcohol last night. The health habits ($n = 4$) were: hours slept last night (sleep

quantity), felt refreshed after waking up today (sleep quality), minutes physically active today, and minutes spent in nature today. Most variables were reported on a 5-point Likert scale, (0 = not at all, 1 = a little, 2 = somewhat, 3 = moderately, 4 = very). Drinking alcohol, sleep quantity, time spent in nature, physical activity, and the diet related variables were freely reported in the specified units (standard drinks, hours, and minutes, respectively, and servings).

Figure 1 presents a flow diagram of data cleaning. We excluded those participants who had provided fewer than 7 diary records (out of 13 possible) across the course of the study ($n = 25$). Across all daily variables, we dropped the first two days of observations to account for initial elevation bias – the tendency of participants to over-report symptoms in the beginning of longitudinal studies (Shrout et al., 2018). To make sure that excluding the first two days did not bias our results, we ran a sensitivity analysis (see Supplementary Materials). Additionally, there were 0.5% of missing values across all 29 daily variables (8 flourishing items + 22 lifestyle habits). All variables except for minutes spent in nature today were missing less than 1.5% of all values. The minutes spent in nature variable was missing 4.35% of all values, and there were several participants for whom all or majority ($> 50\%$) of values were missing. Because of the number of missing values, we decided to drop the variable, we again ran sensitivity analysis to show that its exclusion did not affect the selection (see Supplementary Materials). After excluding the time spent in nature variable, the daily ($n = 21$) and demographic ($n = 5$) made up a total of 26 variables. Because the ethnicity variable had four levels, this added two additional dummy predictors, and so there were in the end 28 predictors, including dummy variables. We did not impute 0.4% of missing values that were left after excluding the time spent nature variable, as the values for each variable would be averaged across days for each participant. As for demographic variables, there were four participants missing their BMI information, and one participant missing two out of the three of the childhood SES items. These five participants were dropped from the analysis. Thus, the final sample consisted of 791 participants. Dropping the 30 participants from our had little effect on the demographic variables. The mean age in the final sample was 19.72 ($sd = 1.73$), majority of participants were female (72.3%). Most of the participants were New Zealand European/pākehā (77.5%), followed by Asian (10.8%), Māori/Pacifica (5.4%) participants, with the rest making up 6.3% of the sample.

The eight flourishing scale items were averaged into one daily flourishing variable. Following that, all variables from the daily survey were averaged across days (including the newly created daily flourishing variable). Additionally, childhood SES was averaged across items as well. Finally, all continuous variables were centred and scaled to 1-unit standard deviation.

Statistical analyses and modeling

Prior to fitting any models, we randomly split our data into training (75%; $n = 593$) and test set (25%, $n = 198$). The training data was used to fit all models. The test data was used to validate the models' predictions on independent data. All relevant statistics are reported for both the training and test data.

We first fitted a Bayesian multiple regression as the reference model, predicting average daily flourishing from all 28 candidate well-being predictors (21 daily predictors, 5 demographic predictors, with 2 additional dummy predictors for ethnicity). We used weakly informative normal priors for the predictor slopes and the intercept (normal: mean = 0, sd = 1). For model standard deviation, we also used a weakly informative prior (half-normal: mean = 0, sd = 1). As for sampling, we ran four chains of 4,000 iterations, with 2,000 iterations of warm-up and 2,000 iterations of sampling each. The chains were run in parallel to speed up convergence.

After fitting the reference model, we used predictive projection feature selection to fit a submodel with fewer predictors that would give similar predictions to the full model. To implement predictive projection, we used the `projpred` package (Piironen et al., 2018). The variables were entered into the submodels using L1 search (`projpred` default for > 20 variables) and the submodels' predictive performance was evaluated and compared using expected log predictive density (ELPD) obtained through PSIS-LOO cv (Pareto-smoothed importance sampling leave-one-out cross-validation; Vehtari, Gelman, & Gabry, 2016). For the optimal submodel, we chose the smallest submodel that had ELPD within 1 standard error of the reference model (1SE-submodel; `projpred` default). That is, we chose the smallest submodel which was expected to perform the same as/outperform the reference model with at least 16% probability (and perform worse with 84% probability).

Finally, we compared the predictive performance of predictive projection feature selection to five other models: the original reference model (Bayesian multiple regression), a frequentist multiple regression model with all predictors, frequentist stepwise selection model using p-values, frequentist stepwise selection model using Akaike Information criterion (Akaike, 1974), LASSO with minimum cross-validated RMSE (min-LASSO), and LASSO with cross-validated RMSE within 1 SE of the minimum (1SE-LASSO). R does not provide a default function for stepwise selection using p-values so we used a publicly available R code that implements an SPSS-like stepwise selection with p-values (Meys, 2009). Besides the projected 1SE-submodel described above, we also included in the comparison a projected submodel with cross-validated predictive performance set to match reference model, that is, the smallest possible projection submodel for which there was at least 50% probability that it would perform as well as/better than the reference model. In total, eight models were compared. To evaluate the performance of all models, we used root mean squared error (RMSE) and Bayesian ρ . Bayesian ρ is a generalization of classical ρ which has the advantage that it can incorporate posterior uncertainty and remains bounded below 1 even in the presence of strong priors and weak data (Gelman et al., 2019); here we just use it as a convenient summary of predictive performance that we could compare: a) between training and test data, and b) across models on test data only.

Results

The reference model converged well, with no divergent transitions and good \hat{r} values (all ≈ 1), indicating that the chains mixed well (Gelman, Andrew & Rubin, 1992). All parameters had a good effective sample size (all > 5,000). The model passed simple posterior predictive checks and visual checks of residuals did

not reveal any heteroscedasticity or gross non-linear trends (see Supplementary Materials). There were few outliers with very low average daily flourishing in the data, however, when we evaluated the model via PSIS-LOO cv, we found no evidence of these observations having a disproportionate influence on the model fit, as indicated by satisfactory Pareto-k values (all “good”, $k < 0.5$; see Supplementary Materials). On the training data, the reference model had a Bayesian R^2 of 0.359 (0.308 – 0.405, 95% credible interval; CI) and a model standard deviation/RMSE of 0.821 (0.774 – 0.870 95% CI), indicating a moderately good predictive performance.

Using the predictive projection, we found that a submodel with only 3 predictors made predictions that were similar enough to those of the reference model with all 28 predictors, based on the 1 SE rule. The sensibility of the 1 SE rule was confirmed with a visual check of the feature selection trajectory, which showed that predictive performance did seem to stop improving after 3 predictors (see Figure 2A). On the training data, the 1SE-submodel had a Bayesian R^2 of 0.267 (0.213-0.323, 95% CI) and a model standard deviation/RMSE of 0.878 (0.826 – 0.931 95% CI). The 1SE-submodel’s predictions were strongly correlated with the reference model’s predictions (Pearson $r = 0.897$, see Figure 2B). The three predictors in the optimal submodel, in order as they were entered into the submodel, the three predictors were: felt refreshed after waking up today (0.377, 0.306 – 0.453 95% CI), had trouble concentrating today (-0.211, -0.281 – -0.138 95% CI), and servings of fruit today (0.132, 0.065 – 0.200 95% CI; see Figure 3A). We tested the submodel’s performance on the independent test data and it performed well, with an observed Bayesian R^2 of 0.253 and an observed RMSE of 0.884 (see Figure 3B).

The results of the comparison of variable selection methods are shown in Table 1. The overall trend was that the largest models had the best predictive accuracy, as assessed by test data R^2 and RMSE: frequentist multiple regression reference model (28 predictors; R^2 : 0.332, RMSE: 0.858), reference model (28 predictors; R^2 : 0.331, RMSE: 0.858) and the min-LASSO (23 predictors; R^2 : 0.283, RMSE: 0.857). Based on Bayesian R^2 alone, the models that performed the best were: frequentist multiple regression (0.332), the reference model (0.331), AIC-stepwise (0.315), matched-submodel (0.284), min-LASSO (0.283), p-stepwise (0.275), 1SE-submodel (0.253), 1SE-LASSO (0.139). Based on RMSE, the models that performed the best were: the reference model (0.858), the frequentist multiple regression (0.858), min-LASSO (0.857), matched-submodel, p-stepwise (0.864), p-stepwise (0.871), AIC-stepwise (0.871), 1SE-submodel (0.883), 1SE-LASSO (0.893). In terms of the numbers of selected predictors, predictive projection produced the simplest models, with 1SE-submodel containing only 3 predictors and the matched projection submodel containing 6 predictors, fewer than any of the other methods apart from 1SE-LASSO (4 predictors).

Discussion

Using predictive projection feature selection, we found that only three of the 28 candidate well-being predictors were sufficient to approximate the predictions of a large reference model with all 28 predictors. Specifically, participants who reported feeling more refreshed after waking up, having less trouble concentrating, and eating more servings of fruit scored highest in their average daily flourishing. We

tested the optimal 3-predictor submodel on test data and found that its predictive performance generalized well to new data, with performance measured by RMSE and Bayesian R^2 only slightly worse than on the training data and well-within the training data uncertainty bounds. Lastly, we also found that, when comparing predictive projection to other variable selection methods, larger models generally tended to have better predictive accuracy on test data, however, predictive projection tended to produce smaller models, and the projection submodel matched to the reference model performed better than stepwise selection based on p-values, with fewer predictors.

The projected 1SE-submodel had somewhat worse predictive accuracy on test data than most of the other variable selection methods, as measured by RMSE and Bayesian R^2 , however it achieved this performance with much fewer selected predictors than the other methods. In fact, the 1SE-submodel had better test data performance (as indicated by both RMSE and R^2) than the second smallest model – 1SE-LASSO – which included an additional predictor. The other models included at least twice as many predictors as the projected 1SE-submodel or more. Interestingly, 1SE-LASSO had a particularly low Bayesian R^2 , however, this was due to strong shrinkage – the model predicted only a narrow range of outcomes which lead to very small variance in predictions (the numerator in Bayesian R^2 , see Supplementary Materials). The reason why large models performed the best in our study is most likely because we had used a relatively large training sample ($n = 593$). Overfitting is less of a concern when the ratio of training observations to the number of predictors is large (McNeish, 2015), as was the case in our study. The more regularizing methods, namely the 1SE predictive projection and 1SE LASSO, may have performed comparatively better, if the training sample had been smaller. Still, predictive projection produced a well-performing model, considering it had contained only 3 predictors.

It is also important to highlight that while the two stepwise models and the min-LASSO model had higher test-data predictive accuracy than the projected 1SE-submodel, none of these models can be readily used for inference. The stepwise models cannot be used for inference because the p-values and standard errors from these models are not adjusted for the selection and as such do not control for type-I error (Flom & Cassell, 2007; Smith, 2018). As for LASSO, the tools for inference with these models are still undergoing development (Benjamini, 2010; Taylor & Tibshirani, 2015). The projected submodels, on the other hand, have a valid posterior distribution (Piironen et al., 2018) and thus can be used for inference.

The fact that the submodel with only three predictors was sufficient to predict almost as much variation in average daily flourishing as the reference model with all 28 predictors does not suggest that the left out 25 predictors have no relationship with flourishing. Instead, our results suggest that the three selected predictors are the strongest predictors of average daily flourishing, at least among groups of correlated predictors. For example, inasmuch as diet is related to flourishing, our results suggest that fruit consumption may be the strongest indicator of good diet and flourishing, as indicated by the fact that it was the earliest from the group of diet-related predictors in the feature selection trajectory, and was the only diet-related predictor that was present in the optimal submodel. Likewise, while there were several correlated candidate predictors related to fatigue and somatic issues, the fact that having (less) trouble

concentrating was the first and only predictor from this group selected into the optimal submodel suggests that it may be the strongest somatic predictor of well-being.

The three predictors selected into the projected submodel are all theoretically meaningful within the field of well-being research. The first predictor selected into the submodel was sleep quality – how refreshed participants felt after waking up, on their average day. Sleep quality has been consistently shown to be one of the strongest predictors of well-being, especially in young adults, with poor sleep quality being strongly linked to poor mental health outcomes including symptoms of depression (Pilcher et al., 1997; Ridner et al., 2016; Wallace et al., 2017; S.-R. Wickham et al., 2020; Wilson et al., 2014). Additionally, while sleep quality has often been shown to be an important predictor of well-being, sleep quantity has not (Pilcher et al., 1997; Wallace et al., 2017), and this is congruent with our results – while sleep quality was entered into the submodels early along the feature selection trajectory, sleep quantity was only entered long after any improvement in predictive accuracy was shown, indicating a lack of predictive power. The second predictor selected into the submodel was how much trouble concentrating a participant had on their average day. This results is also meaningful – having trouble concentrating is one of the key symptoms of the major depressive disorder (MDD), and is often assessed by diagnostic scales, such as the popular CES-D scale (Kohout et al., 1993) and the DSSS scale (Hung et al., 2006). Finally, the third and the last predictor entered into the submodel was daytime fruit consumption. As discussed above, the fact that daytime fruit consumption was the first diet-related predictor to be selected in the feature selection trajectory and the only diet-related predictor that made it into the submodel suggests that it may be one of the strongest indicators of diet quality, as it relates to flourishing. Fruit and vegetable consumption has been previously shown to predict psychological well-being and flourishing (Blanchflower et al., 2013; Conner et al., 2015, 2017; Hong & Peltzer, 2017; Mujcic & Oswald, 2016; Prendergast, Mackay, et al., 2016). There were other diet related predictors entered into the submodel early along the feature selection trajectory, namely (lower) night-time softdrink consumption and daytime vegetable consumption as the 4th and 5th predictors, respectively. However, based on the 1 SE rule, fruit consumption only was sufficient to approximate the reference model's predictions, suggesting that fruit consumption may contain enough information about the quality of one's diet to make other diet-related predictors redundant. Additionally, there is evidence that raw fruit and vegetables are stronger predictors of well-being than cooked fruit and vegetables (Brookie et al., 2018; S.-R. Wickham et al., 2020), and since fruit is more often eaten raw, general fruit consumption may be a stronger indicator of good diet than vegetable consumption. Be it as it may, sleep quality, having trouble concentrating, and fruit consumption are all meaningful predictors within the well-being literature.

Our study has several limitations. First, the predictive performance of predictive projection and the other methods was evaluated on only one independent test set. As such, the RMSE and R^2 values we obtained in our study may be subject to sampling variation. While this is an issue, our main goal was to demonstrate the use of predictive projection and show that it is a viable alternative to the other methods – not to firmly prove that it has superior predictive performance. Even if other methods predict better, they are still lacking when it comes to inference. Future studies may use simulation to compare predictive

performance of predictive projection, LASSO, and stepwise selection, under different conditions such as sample size and distribution of effect sizes. Second, as was mentioned earlier, the fact that only three variables were selected into the optimal submodel does not mean that the other predictors are not related to well-being. Therefore, we cannot conclude that the three predictors are the only predictors of flourishing, only that they are sufficient to predict it with high degree of accuracy, similar to the model with all 28 predictors. Third, given that our data is observational, we cannot make causal claims about the predictors' influence. When we say that a predictor predicts higher average daily flourishing, we mean that participants with higher values in the predictor tend to report higher average daily flourishing. In concrete terms, poor sleep, lack of ability to concentrate, and poor diet, may not cause low well-being, but instead may be just indicators or even results of low well-being. While the lack of ability to make directional causal claims is certainly a limitation, analyses of observational data are important and necessary to identify possible targets of interventions to be investigated in follow up research. Fourth, our sample was relatively homogenous, consisting of young, college-age adults (age-range 17-25 years), and thus some of the findings may not generalize to different populations. For example, young adults tend to be at an increased risk for poor sleep quality (Lund et al., 2010), and thus it is possible that the association between sleep quality and well-being may be weaker in older population. Predictive projection may be applied in future to data from other cohorts to see whether there are age-related and other differences. Fifth, as our outcome measure, we used the Flourishing Scale (Diener et al., 2010), and while this is a popular measure of well-being, it is by far not the only one. There is an extensive amount of theoretical work comparing the different ways of measuring well-being that is outside the scope of the present article, however, it may be interesting to see how much the results from predictive projection generalize to other well-being scale. Finally, while our data comes from a micro-longitudinal daily diary study, we only analysed our data cross-sectionally. There are several reasons for why we did not analyse the trends within participants. First off, predictive projection feature selection is not yet implemented for mixed effects models (Piironen et al., 2018) and so at the present moment we do not have the ability to build within-participants models in the same way as we did across participants. Second, while we could use the three predictors from the submodel to fit a within-participant mixed effects model, there is no guarantee that the three strongest cross-sectional predictors of flourishing across participants will also be the strongest predictors of flourishing within-participants – in fact, it is rather unlikely (Simpson's paradox; see Wagner, 1982). As such, we believe that repeated measures analyses using mixed effects models are outside of the scope of this article.

Conclusion

We have demonstrated that how predictive projection feature selection can be used to build compact models with good predictive power that can also be used for inference. Specifically, we were able to accurately predict average daily flourishing across young adults with a model that used information from just three predictors: how refreshed the participants felt after waking up, how much trouble they had concentrating, and how many servings of fruit they ate on their average day. That is, we were able to approximate the predictive accuracy of a large model with all 28 predictors with a much smaller

submodel that contained only the three predictors. Compared with other variable selection methods, predictive projection performed adequately and produced much more parsimonious models. The findings from our model are congruent with the well-being literature, in that, for example, sleep quality is more strongly associated with better well-being than sleep quantity, that having trouble concentrating is related to lower well-being, and having good dietary habits (as indicated by fruit consumption) is related to higher well-being. Variable selection is a common issue that arises frequently across many fields, and we believe that predictive projection is a very useful method that has key advantages over other methods currently in use and will hopefully become more widely adopted.

Abbreviations

BMI: Body mass index

CI: Credible interval

CES-D: Center for Epidemiological Studies - Depression scale

DSSS: Depression and somatic symptoms scale

ELPD: Expected log-predictive density

PSIS-LOO cv: Pareto-smoothed importance sampling leave-one-out cross-validation

RMSE: Root-mean-squared error

SE: Standard error

SES: Socioeconomic status

Declarations

Ethics approval and consent to participate

Ethical approval was required for data collection in this study and was reviewed and granted by the University of Otago Human Ethics Committee (approval #10/177). All participants in the study signed written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The Daily Life Study dataset contains sensitive demographic information that could potentially be used to identify individual participants. De-identified dataset will be made available from the corresponding

author upon reasonable request. All R code used for the analysis is available as part of the Supplementary Materials.

Competing interests

The authors do not have any competing interests.

Funding

The Daily Life Study was funded by the Health Research Council of New Zealand (grant 12/709 to TSC), a University of Otago Research Grant to TSC, and the Department of Psychology, University of Otago. Neither the Health Research Council of New Zealand nor the University of Otago had any influence on the study design, collection, analysis, interpretation, or writing of the manuscript.

Author information

Department of Psychology, University of Otago

William James Building, 275 Leith Walk, Dunedin 9016

Adam Bartonicek, Shay-Ruby Wickham, Narun Pat (Pornpattananangkul), Tamlin Conner

Authors' contributions

AB and TC conceived the study. AB performed data cleaning and analysis, SW contributed to interpretation of the data, and TC and NP provided consultations on theory and statistical aspects of the analysis. All authors (AB, SW, NP, and TC) contributed to drafting, reading, writing, and revising the manuscript.

Acknowledgements

We would like to thank Dr. Jayde Flett, Kathryn Bees, and Samantha Jones for their outstanding work coordinating the Daily Life Study, members of The Daily Experiences Lab in 2011-2014 for assistance with data collection, and Hadyn Youens for assistance with study programming.

References

Adams, T. B., & Colner, W. (2008). The association of multiple risk factors with fruit and vegetable intake among a nationwide sample of college students. *Journal of American College Health, 56*(4), 455–461. <https://doi.org/10.3200/JACH.56.44.455-464>

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

- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>
- Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6), 708–721. <https://doi.org/10.1002/bimj.200900299>
- Blanchflower, D. G., Oswald, A. J., & Stewart-Brown, S. (2013). Is Psychological Well-Being Linked to the Consumption of Fruit and Vegetables? *Social Indicators Research*, 114(3), 785–801. <https://doi.org/10.1007/s11205-012-0173-y>
- Brookie, K. L., Best, G. I., & Conner, T. S. (2018). Intake of raw fruits and vegetables is associated with better mental health than intake of processed fruits and vegetables. *Frontiers in Psychology*, 9(APR), 1–14. <https://doi.org/10.3389/fpsyg.2018.00487>
- Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1). <https://doi.org/10.18637/jss.v080.i01>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Conner, T. S., Brookie, K. L., Carr, A. C., Mainvil, L. A., & Vissers, M. C. M. (2017). Let them eat fruit! the effect of fruit and vegetable consumption on psychological well-being in young adults: A randomized controlled trial. *PLoS ONE*, 12(2), 1–19. <https://doi.org/10.1371/journal.pone.0171206>
- Conner, T. S., Brookie, K. L., Richardson, A. C., & Polak, M. A. (2015). On carrots and curiosity: Eating fruit and vegetables is associated with greater flourishing in daily life. *British Journal of Health Psychology*, 20(2), 413–427. <https://doi.org/10.1111/bjhp.12113>
- Daróczi, G., & Roman, T. (2018). *pander: An R "Pandoc" Writer* (0.6.3). <https://cran.r-project.org/package=pander>
- Diener, E., & Chan, M. Y. (2011). Happy People Live Longer: Subjective Well-Being Contributes to Health and Longevity. *Applied Psychology: Health and Well-Being*, 3(1), 1–43. <https://doi.org/10.1111/j.1758-0854.2010.01045.x>
- Diener, E., Seligman, M. E. P., Choi, H., & Oishi, S. (2018). Happiest People Revisited. *Perspectives on Psychological Science*, 13(2), 176–184. <https://doi.org/10.1177/1745691617697077>
- Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, 97(2), 143–156.

- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals*, *32*(2), 407–499.
- Efroymson, M. A. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, 191–203. <https://ci.nii.ac.jp/naid/10007639144>
- Flom, P. L., & Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. *Northeast SAS User Group (NESUG) Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland*, 1–7.
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *182*(2), 389–402. <https://doi.org/10.1111/rssa.12378>
- Gelman, Andrew, & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511. <https://doi.org/10.1214/ss/1177013437>
- Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian Regression Models. *American Statistician*, *73*(3), 307–309. <https://doi.org/10.1080/00031305.2018.1549100>
- Goodwin, R. D. (2003). Association between physical activity and mental disorders among adults in the United States. *Preventive Medicine*, *36*(6), 698–703. [https://doi.org/10.1016/S0091-7435\(03\)00042-2](https://doi.org/10.1016/S0091-7435(03)00042-2)
- Griskevicius, V., Delton, A. W., Robertson, T. E., & Tybur, J. M. (2011). Environmental contingency in life history strategies: The influence of mortality and socioeconomic status on reproductive timing. *Journal of Personality and Social Psychology*, *100*(2), 241–254. <https://doi.org/10.1037/a0021082>. Environmental
- Hassmén, P., Koivula, N., & Uutela, A. (2000). Physical exercise and psychological well-being: A population study in Finland. *Preventive Medicine*, *30*(1), 17–25. <https://doi.org/10.1006/pmed.1999.0597>
- Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. <https://doi.org/10.1021/ci0342472>
- Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579–586. <https://doi.org/10.1021/ci025626i>
- Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. <https://doi.org/10.1002/bimj.201700067>
- Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology*, *13*(1), 1–19. <https://doi.org/10.20982/tqmp.13.1.p001>

- Höge, M., Wöhling, T., & Nowak, W. (2018). A Primer for Model Selection: The Decisive Role of Model Complexity. *Water Resources Research*, *54*(3), 1688–1715. <https://doi.org/10.1002/2017WR021902>
- Hong, S. A., & Peltzer, K. (2017). Dietary behaviour, psychological well-being and mental distress among adolescents in Korea. *Child and Adolescent Psychiatry and Mental Health*, *11*(1), 1–12. <https://doi.org/10.1186/s13034-017-0194-z>
- Hung, C. I., Weng, L. J., Su, Y. J., & Liu, C. Y. (2006). Depression and somatic symptoms scale: A new scale with both depression and somatic symptoms emphasized. *Psychiatry and Clinical Neurosciences*, *60*(6), 700–708. <https://doi.org/10.1111/j.1440-1819.2006.01585.x>
- Huppert, F. A., & So, T. T. C. (2013). Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-Being. *Social Indicators Research*, *110*(3), 837–861. <https://doi.org/10.1007/s11205-011-9966-7>
- Jacka, F. N., Kremer, P. J., Berk, M., de Silva-Sanigorski, A. M., Moodie, M., Leslie, E. R., Pasco, J. A., & Swinburn, B. A. (2011). A prospective study of diet quality and mental health in adolescents. *PLoS ONE*, *6*(9), 1–7. <https://doi.org/10.1371/journal.pone.0024805>
- Jacka, F. N., Kremer, P. J., Leslie, E. R., Berk, M., Patton, G. C., Toumbourou, J. W., & Williams, J. W. (2010). Associations between diet quality and depressed mood in adolescents: Results from the Australian Healthy Neighbourhoods Study. *Australian and New Zealand Journal of Psychiatry*, *44*(5), 435–442. <https://doi.org/10.3109/00048670903571598>
- Jacka, F. N., O’Neil, A., Opie, R., Itsiopoulos, C., Cotton, S., Mohebbi, M., Castle, D., Dash, S., Mihalopoulos, C., Chatterton, M. Lou, Brazionis, L., Dean, O. M., Hodge, A. M., & Berk, M. (2017). A randomised controlled trial of dietary improvement for adults with major depression (the “SMILES” trial). *BMC Medicine*, *15*(1), 1–13. <https://doi.org/10.1186/s12916-017-0791-y>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. <https://doi.org/10.1016/j.peva.2007.06.006>
- Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, *99*(465), 279–290. <https://doi.org/10.1198/016214504000000269>
- Kassambara, A. (2019). *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2' (0.1.3)*. <https://cran.r-project.org/package=ggcorrplot>
- Kawada, T., Kuratomi, Y., & Kanai, T. (2009). Lifestyle determinants of depressive feeling and a feeling of unhappiness among workers: a study in Japan. *Work*, *33*(3), 255–260.
- Kern, M. L., Della Porta, S. S., & Friedman, H. S. (2014). Lifelong pathways to longevity: Personality, relationships, flourishing, and health. *Journal of Personality*, *82*(6), 472–484. <https://doi.org/10.1111/jopy.12062>

- Keyes, C. L. M. (2007). Promoting and Protecting Mental Health as Flourishing: A Complementary Strategy for Improving National Mental Health. *American Psychologist*, *62*(2), 95–108. <https://doi.org/10.1037/0003-066X.62.2.95>
- Keyes, C. L. M., Dhingra, S. S., & Simoes, E. J. (2010). Change in level of positive mental health as a predictor of future risk of mental illness. *American Journal of Public Health*, *100*(12), 2366–2371. <https://doi.org/10.2105/AJPH.2010.192245>
- Keyes, C. L. M., & Simoes, E. J. (2012). To flourish or not: Positive mental health and all-cause mortality. *American Journal of Public Health*, *102*(11), 2164–2172. <https://doi.org/10.2105/AJPH.2012.300918>
- Kohout, F. J., Berkman, L. F., Evans, D. A., & Cornoni-Huntley, J. (1993). Two shorter forms of the CES-D depression symptoms index. *Journal of Aging and Health*, *5*(2), 179–193.
- Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, *9*(JUL), 1–4. <https://doi.org/10.3389/fpsyg.2018.01117>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Larmarange, J. (2020). *labelled: Manipulating labelled data* (2.2.2). <https://cran.r-project.org/package=labelled>
- Li, N. P., & Kanazawa, S. (2016). Country roads, take me home... to my friends: How intelligence, population density, and friendship affect modern happiness. *British Journal of Psychology*, *107*(4), 675–697. http://ink.library.smu.edu.sg/soss_research
- Liao, Y., Schembre, S. M., O'Connor, S. G. O., Belcher, B. R., Maher, J. P., Dzibur, E., & Dunton, G. F. (2018). An electronic ecological momentary assessment study to examine the consumption of high-fat/high-sugar foods, fruits/vegetables and affective states among women. *Journal of Nutrition Education and Behavior*, *50*(6), 626–631. <https://doi.org/doi:10.1016/j.jneb.2018.02.003>. An
- Lund, H. G., Reider, B. D., Whiting, A. B., & Prichard, J. R. (2010). Sleep Patterns and Predictors of Disturbed Sleep in a Large Population of College Students. *Journal of Adolescent Health*, *46*(2), 124–132. <https://doi.org/10.1016/j.jadohealth.2009.06.016>
- McNeish, D. M. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. <https://doi.org/10.1080/00273171.2015.1036965>
- Meys, J. (2009). *Automated model selection*. <https://stackoverflow.com/questions/3701170/stepwise-regression-using-p-values-to-drop-variables-with-nonsignificant-p-value>

Moreno-Agostino, D., Caballero, F. F., Martín-María, N., Tyrovolas, S., López-García, P., Rodríguez-Artalejo, F., Haro, J. M., Ayuso-Mateos, J. L., & Miret, M. (2019). Mediterranean diet and wellbeing: evidence from a nationwide survey. *Psychology and Health, 34*(3), 321–335. <https://doi.org/10.1080/08870446.2018.1525492>

Mujcic, R., & Oswald, A. J. (2016). Evolution of well-being and happiness after increases in consumption of fruit and vegetables. *American Journal of Public Health, 106*(8), 1504–1510. <https://doi.org/10.2105/AJPH.2016.303260>

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science, 28*(2), 135–167. <https://doi.org/10.1214/12-STS410>

Oishi, S., Kesebir, S., & Diener, E. (2011). Income inequality and happiness. *Psychological Science, 22*(9), 1095–1100. <https://doi.org/10.1177/0956797611417262>

Piironen, J., Paasiniemi, M., & Vehtari, A. (2018). *Projective Inference in High-dimensional Problems: Prediction and Feature Selection. 2015*, 1–42. <http://arxiv.org/abs/1810.02406>

Pilcher, J. J., Ginter, D. R., & Sadowsky, B. (1997). Sleep quality versus sleep quantity: Relationships between sleep and measures of health, well-being and sleepiness in college students. *Journal of Psychosomatic Research, 42*(6), 583–596. [https://doi.org/10.1016/S0022-3999\(97\)00004-4](https://doi.org/10.1016/S0022-3999(97)00004-4)

Pinquart, M., & Sörensen, S. (2000). Influences of socioeconomic status, social network, and competence on subjective well-being in later life: A meta-analysis. *Psychology and Aging, 15*(2), 187–224. <https://doi.org/10.1037/0882-7974.15.2.187>

Piqueras, J. A., Kuhne, W., Vera-Villaruel, P., Van Straten, A., & Cuijpers, P. (2011). Happiness and health behaviours in Chilean college students: A cross-sectional survey. *BMC Public Health, 11*. <https://doi.org/10.1186/1471-2458-11-443>

Prendergast, K. B., Mackay, L. M., & Schofield, G. M. (2016). The Clustering of Lifestyle Behaviours in New Zealand and their Relationship with Optimal Wellbeing. *International Journal of Behavioral Medicine, 23*(5), 571–579. <https://doi.org/10.1007/s12529-016-9552-0>

Prendergast, K. B., Schofield, G. M., & Mackay, L. M. (2016). Associations between lifestyle behaviours and optimal wellbeing in a diverse sample of New Zealand adults. *BMC Public Health, 16*(1), 1–11. <https://doi.org/10.1186/s12889-016-2755-0>

Pressman, S. D., & Cohen, S. (2005). Does positive affect influence health? *Psychological Bulletin, 131*(6), 925–971. <https://doi.org/10.1037/0033-2909.131.6.925>

R Core Team. (2019). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing, Vienna, Austria* (3.6.0). <https://www.r-project.org/>

- Ridner, S. L., Newton, K. S., Staten, R. R., Crawford, T. N., & Hall, L. A. (2016). Predictors of well-being among college students. *Journal of American College Health, 64*(2), 116–124. <https://doi.org/10.1080/07448481.2015.1085057>
- Shrout, P. E., Stadler, G., Lane, S. P., Joy McClure, M., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences of the United States of America, 115*(1), E15–E23. <https://doi.org/10.1073/pnas.1712277115>
- Smith, G. (2018). Step away from stepwise. *Journal of Big Data, 5*(1). <https://doi.org/10.1186/s40537-018-0143-6>
- Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E., & Habbema, J. D. F. (2001). Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making, 21*(1), 45–56. <https://doi.org/10.1177/0272989X0102100106>
- Tay, L., & Kuykendall, L. (2013). Promoting happiness: The malleability of individual and societal subjective wellbeing. *International Journal of Psychology, 48*(3), 159–176. <https://doi.org/10.1080/00207594.2013.779379>
- Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences of the United States of America, 112*(25), 7629–7634. <https://doi.org/10.1073/pnas.1507583112>
- Thompson, B. (1995). Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply here: A Guidelines Editorial. In *Educational and Psychological Measurement* (Vol. 55, Issue 4, pp. 525–534). <https://doi.org/10.1177/0013164495055004001>
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tibshirani, R. (2019). *Statistical learning and sparsity*. <https://www.stat.auckland.ac.nz/en/about/news-and-events-5/events/events-2019/04/ihaka-lecture-series-2019-statistical-learning-and-sparsity.html>
- Tierney, N., Cook, D., McBain, M., & Fay, C. (2020). *naniar: Data Structures, Summaries, and Visualisations for Missing Data* (0.5.0). <https://cran.r-project.org/package=naniar>
- van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology, 89*, 31–50. <https://doi.org/10.1016/j.jmp.2018.12.004>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>

- Wagner, C. H. (1982). Simpson ' s Paradox in Real Life Author. *The American Statistician*1, 36(1), 46–48.
- Wahl, D. R., Villinger, K., König, L. M., Ziesemer, K., Schupp, H. T., & Renner, B. (2017). Healthy food choices are happy food choices: Evidence from a real life sample using smartphone based assessments. *Scientific Reports*, 7(1), 1–8. <https://doi.org/10.1038/s41598-017-17262-9>
- Wallace, D. D., Boynton, M. H., & Lytle, L. A. (2017). Multilevel analysis exploring the links between stress, depression, and sleep problems among two-year college students. *Journal of American College Health*, 65(3), 187–196. <https://doi.org/10.1080/07448481.2016.1269111>
- White, B. A., Horwath, C. C., & Conner, T. S. (2013). Many apples a day keep the blues away - Daily experiences of negative and positive affect and food consumption in young adults. *British Journal of Health Psychology*, 18(4), 782–798. <https://doi.org/10.1111/bjhp.12021>
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, 75(5), 1182–1189. <https://doi.org/10.1111/j.1365-2656.2006.01141.x>
- Wickham, H. (2017). *tidyverse: Easily Install and Load the "Tidyverse"* (R package version 1.2.1). <https://cran.r-project.org/web/packages/tidyverse/index.html>
- Wickham, S.-R., Amarasekara, N. A., Bartonicek, A., & Conner, T. S. (2020). The Big Three Health Behaviors and Mental Health and Well-Being Among Young Adults: A Cross-Sectional Investigation of Sleep, Exercise, and Diet. *Frontiers in Psychology*, 11, 3339. <https://doi.org/10.3389/fpsyg.2020.579205>
- Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2"*. (R package version 0.9.4). <https://cran.r-project.org/package=cowplot%0A>
- Wilson, K. T., Bohnert, A. E., Ambrose, A., Davis, D. Y., Jones, D. M., & Magee, M. J. (2014). Social, behavioral, and sleep characteristics associated with depression symptoms among undergraduate students at a women's college: A cross-sectional depression survey, 2012. *BMC Women's Health*, 14(1). <https://doi.org/10.1186/1472-6874-14-8>
- Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

Tables

Table 1. Comparison of variable selection methods				
Model	R ²	RMSE	# of selected predictors	Selected predictors
Reference model	0.331	0.858	(28)	-
Freq. multiple regression	0.332	0.858	(28)	-
Projected submodel (1 SE)	0.253	0.883	3	Felt refreshed after waking up today, had trouble concentrating today, servings of fruit today
Projected submodel (matched)	0.284	0.864	6	Felt refreshed after waking up today, had trouble concentrating today, servings of fruit today, servings of softdrink last night, servings of vegetables today, gender: female
Stepwise selection (AIC)	0.315	0.872	10	Felt refreshed after waking up today, ethnicity: asian, had trouble concentrating today, gender: female, servings of soft drink last night, servings of sweets today, servings of sweets last night, felt tired today, servings of fruit today, bmi
Stepwise selection (p-values)	0.275	0.871	8	Felt refreshed after waking up today, had trouble concentrating today, gender: female, servings of sweets today, felt tired today, servings of sweets last night, servings of fruit today, servings of soft drink last night
LASSO (1 SE)	0.139	0.897	4	Felt refreshed after waking up today, had trouble concentrating today, servings of fruit today, servings of softdrink last night
LASSO (min.)	0.283	0.857	23	-

Table 1 legend: Summary statistics of model selection strategies, showing test data RMSE and Bayesian R², number of selected predictors, and the names of the significant predictors (where 10 or fewer predictors were selected, ranked by absolute slope size).

Figures

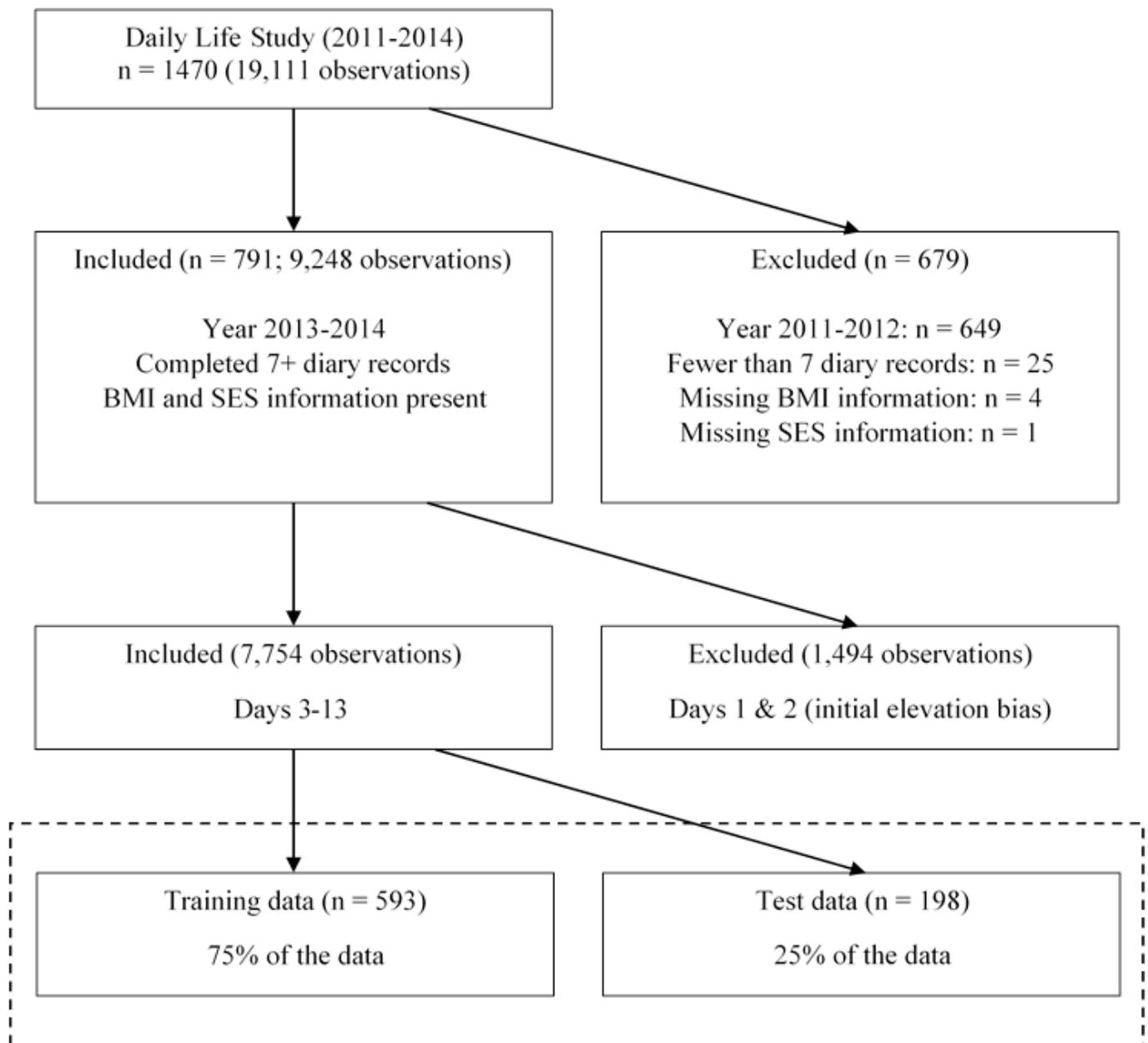


Figure 1

Flow diagram of the data selection procedure

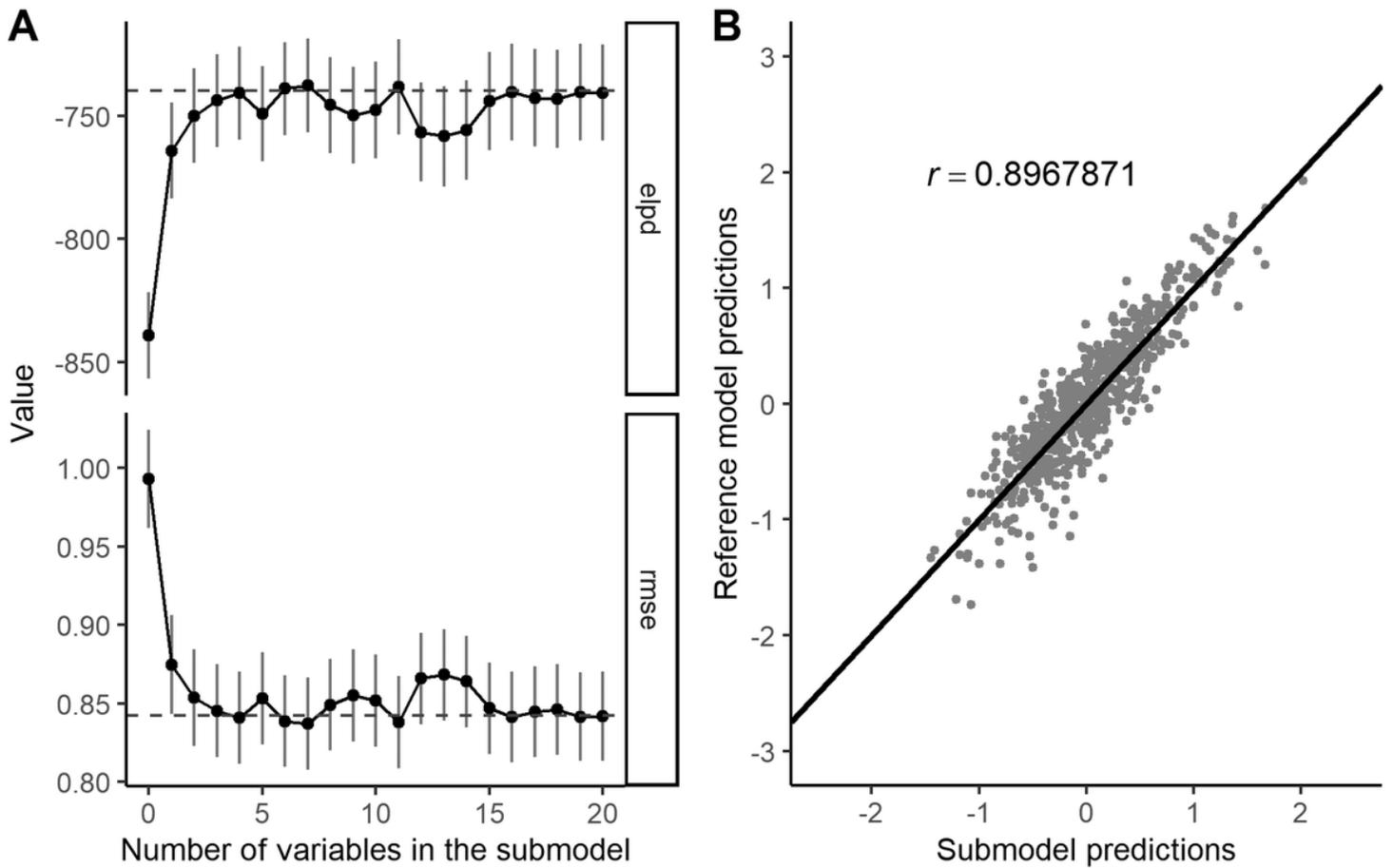


Figure 2

Predictive projection feature selection trajectory and scatterplot of reference model's vs. submodel's prediction Legend: A) Change in ELPD/decrease in RMSE as more predictors entered the submodel. B) Average daily flourishing predicted by the submodel (3 predictors) vs. the average daily flourishing predicted by the reference model (28 predictors; both predicting training data).

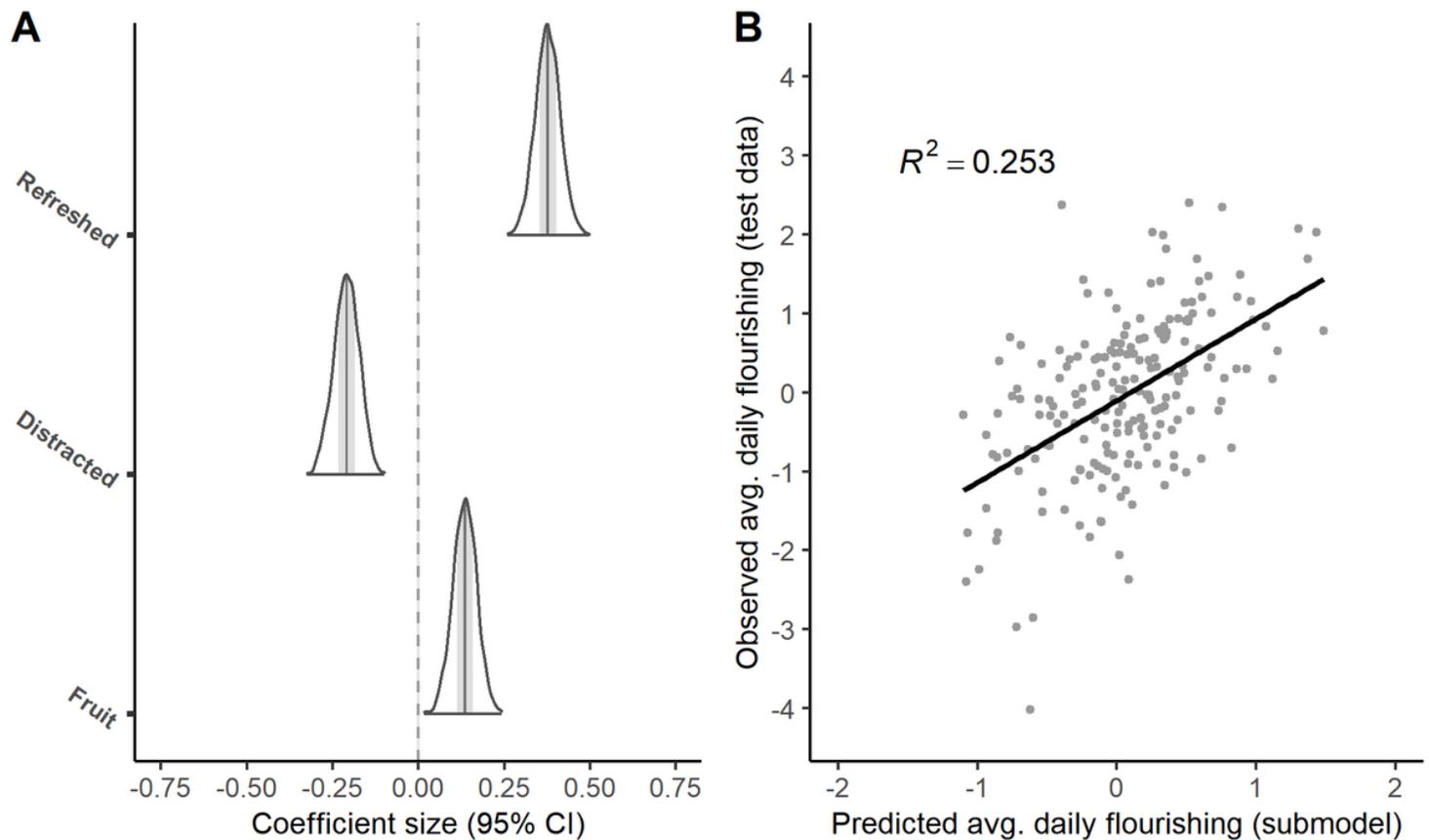


Figure 3

Credible intervals for predictors in the submodel and scatterplot of submodel's predictions vs. observed values Legend: Marginal posterior distributions of predictors selected for the submodel (in order: felt refreshed after waking up today, had trouble concentrating today, servings of fruit today). B) Average daily flourishing predicted by the submodel vs. observed daily flourishing (unseen test data), with overlaid least squares fit.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [dailylife5.pdf](#)
- [AppendixA.docx](#)