Sleep Efficiency inference and prediction with a Statistical Learning approach

Nicolas Anselmi, Andrea Arici, Francesco Corrini

University of Bergamo

Keywords: Statistical Learning, Sleep Efficiency, Lasso, Ridge Regression, Stepwise, Random Forest

Abstract

It's well known that insufficient sleep leads to different problems such as cardiovascular disease, obesity or decrease in productivity[1]. Although some people spend too little time sleeping, others don't get as much sleep as they would like, having a negative impact on their life. Sleep efficiency is the percentage of time spent asleep compared to time spent in bed and it's useful to understand if a person is able to sleep the right amount of time. The aim of this work is to study how different factors affect sleep efficiency using a statistical learning approach and find a model to make predictions.

1 Introduction

One of the most time consuming activity in people life is sleeping. It is common for many people to give up hours of sleep to spend more time working, studying or having fun. However, this practice can lead to many problems, such cardiovascular problems, obesity, diabetes and decrease of in productivity[1]. There are also other factors that affect how much time a person sleeps. A research team from the University of Oxfordshire did a study aimed to investigate the impact of lifestyle factors on sleep quality. A part of this study was collecting data from real person, using a combination of self-reported surveys, actigraphy and polysomnography which is a sleep monitoring technique. This data are avibile on Kaggle[2]. The aim of this work is, using the provided dataset, to make inference on the "sleep efficiency" variable, which is a measure of the proportion of time in bed spent asleep, obteined dividing the time spent asleep by the total time spent in bed. Another goal is to estimate a model that can make prediction on sleep efficiency. To achieve this statistical learning approach are used.

Every feature in this dataset is from a different subject and there aren't multiple measurements on the same person. Age and gender are also recorded. The raw dataset contain 15 attributes and 468 observations. The "Bedtime" and "Wakeup time" features indicate when each subject goes to bed and wakes up each day and the "Sleep duration" feature records the total amount of time each subject slept in hours. The "REM sleep percentage", "Deep sleep percentage", and "Light sleep percentage" features indicate the amount of time each subject spent in each stage of sleep. The "Awakenings" feature records the number of times each subject wakes up

during the night. Additionally, the dataset includes information about each subject's caffeine and alcohol consumption in the 24 hours prior to bedtime, their smoking status, and their exercise frequency. Features description is summarized in Table 1.
"Bedtime" and "Wakeup time" features has been removed from the initial dataset: the difference between them is used ("Sleep duration feature").
The observations that included NA values has been removed and the two categorical variables "Gender" and "Smoking status" are treated as follows: 'zero' in "Gender" means the female sex and 'one' means male sex and, where the "Smoking status" is 'zero', it means that the subject doesn't smoke, 'one' in the opposite case.
In the end the edited dataset has 388 data points and ten features (eleven considering sleep efficiency).

2 Methodology

As shown in the introduction, the first aim of this work is to understand which features affect sleep efficiency the most. In order to do this, methods for selection and regularization for linear models are used, in particular stepwise regression and lasso regularization.
The first method taken into account is stepwise regression. This is a subset selection approach that iteratively estimate a model with a different group of features and choose the best using a statistical criteria*. In this work both forward and backward stepwise are considered: the first, at every iteration, estimate p† models, then it chooses the best among them and the associated feature will be always used in the next iteration, where the same approach is repeated. Backward stepwise uses the same approach, however instead of adding the best regressor at every iteration, it remove the worst one. In the end the best model is chosen. Stepwise is then combined with a bootstrap simulation approach: at every iteration, a stepwise is performed on a resampled dataset, counting which regressors the model picks. In the end, if a feature were included in less than half models it will be discard. Remaining feature are used to estimate a linear model to make inference on. The model is validated using a bootstrap approach computing test mean squared error (MSE). Also the distribution of coefficients is measured to see which of them are significant, providing a confidence interval.
The second method used is the lasso regularization which is a regression method that performs both regularization and variable selection, based on a penalization factor λ in the objective function that penalize the coefficients of the model in order to make some of them null. In particular the optimal λ is selected using k-fold cross-validation. Like in stepwise approach, a bootstrap simulation is performed to understand which features are relevant, discarding those which has been included in less than half of models. Bootstrap simulation is also used to vali-

*in this work Akaike Information Criterion (AIC) is used.
†where p is the number of regressors that have not been chosen by the algorithm.

date the model.
In the end stepwise and lasso results are compared to understand which is best to make inference. In particular, the best model is chosen analyzing the trade-off between model's performance (less MSE) and complexity (lower number of features).
Second goal of this study is to make prediction on sleep efficiency. For this purpose more complex methods like ridge regression and tree-based methods (random forest and bagging) are considered.
The first approach taken into account is ridge regression, an approach similar to lasso that use a different penalization term in the objective function. In particular, while lasso tends to make some coefficients null, ridge regression tends only to make them small. Because of that, the resultant model comprehends all the initial variables. Like in the Lasso approach, the best λ is chosen with k-fold cross-validation and bootstrap is used to compute MSE.
The second approach considered is random forest. Random forest is a tree-based method which consist in using bootstrap simulation to estimate different trees over different (resampled) dataset and take as prediction the average of predictions of all the trees in the model. This approach tends to reduce the variance respect to classical single tree estimation approach. Random forest also use only a subset of features in every iteration to avoid estimating similar trees. A common dimension for predictors subset that is used in this study is the square root of the total number of predictors. In this work is also considered bagging approach, which is like random forest but using all predictors for every tree.
All these approaches will be evaluated using MSE. For prediction is useful to choose the model that minimize MSE, even if it's too complex to understand. However, results are compared with models used for inference to understand if it's necessary and useful to use more complex approaches. In the end predictions are made on a test dataset. A bootstrap approach is used to provide a confidence interval on predictions.

3 Results Analysis

3.1 Inference results

Stepwise results are in Table 2. First column shows the eight features that are selected by counting how much times they are included in models during bootstrap simulation. All the related coefficients are significant as the confidence interval (columns 3 and 4) does not contain zero and so the hypothesis that their null is rejected. The test mean squared error is 0.003884.
Lasso regularization results are shown in Table 3. The optimal λ value obtained by cross-validation is 0.001120. The second column shows the coefficients estimated: two of them are zero because lasso shrinks them. The last column represents how much time the predictor is selected during bootstrap simulation. Only one of them (sleep duration) is not selected. Test mean squared error is 0.003924. Both models give similar results so stepwise selection is

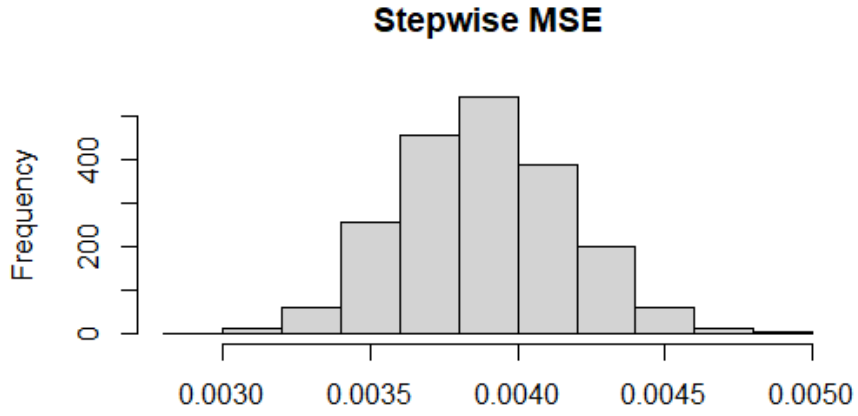

Figure 1: Test MSE distribution with stepwise approach

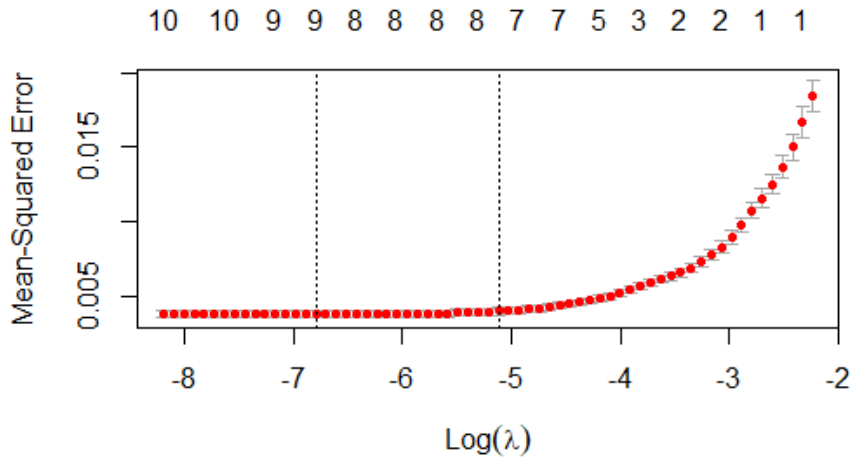

Figure 2: MSE value in function of $lambda during lasso cross validation

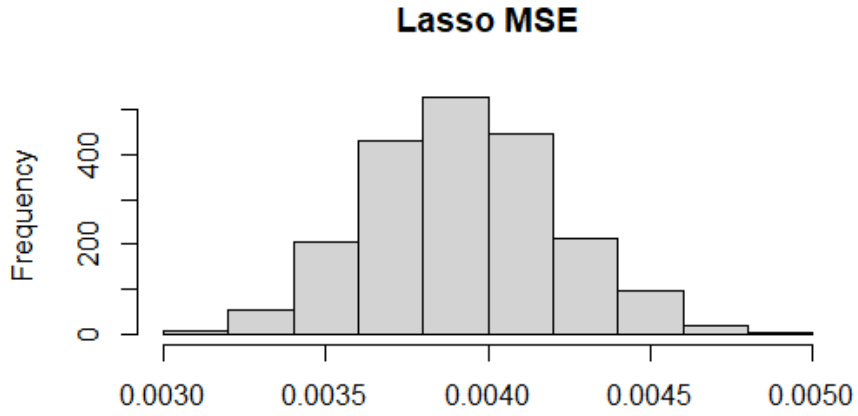

Figure 3: Test MSE distribution with lasso approach

preferred because it has a bit less MSE. Analyzing coefficients, it appears that older people has a little bit better sleep efficiency. Also doing physical exercising increase the response variable. Smoking and drinking alcohol have a negative impact on sleep efficiency. As expected, the number of awakenings during bedtime decrease significantly sleep efficiency while high value of REM and deep sleep increase it. It's strange that caffeine consumption has a positive coefficient (also with lasso approach): it can be that subject who answered the survey assume the same amount of caffeine every day and so their metabolism is used to those doses and so consuming them doesn't affect their sleep efficiency.

3.2 Prediction results

The first approach considered to build a prediction model is ridge regression. The optimal value of λ obtained by cross-validation is 0.010695. Test mean squared error computed by the bootstrap simulation is 0.003924 (quite similar to lasso one). Random forest approach gives a significant decrease in MSE, about 0.002141. The number of trees that gives it is between 80 and 100, as shown in Figure 6. Bagging method leads to similar results, with an MSE a little bit less (0.002128). So this last model is chosen to make predictions. In this work, tree-based methods perform way better than regression model: mean squared error is about an half compared to them. In Table 4 prediction made with a test sample are shown with their confidence interval computed using bootstrap simulation on a bagging model.

4 Conclusions

In conclusion, stepwise and lasso both performed well in feature selection allowing analyzing the resultant model to make inference. It's clear that smoking and drinking are bad habits that has negative impact on sleep efficiency, while regular physical exercise increase it. Tree-based methods performed well predicting sleep efficiency on test data, while in this work ridge regression wasn't good either for inference (no features are shrunken) and predictions (MSE was double respect to random forest and bagging).

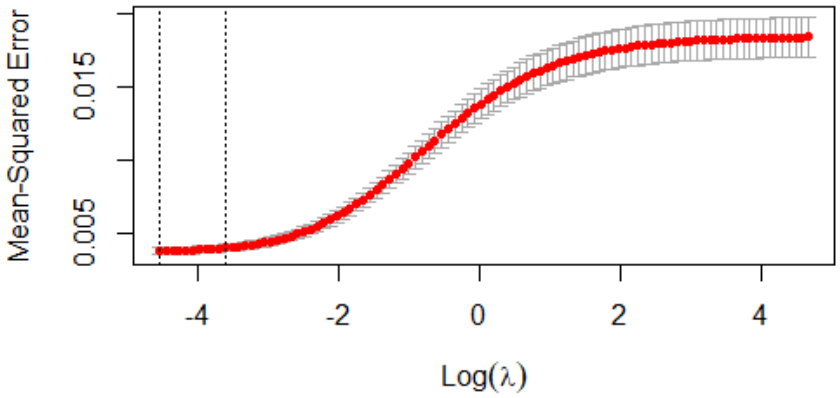

Figure 4: MSE value in function of λ during ridge regression cross validation

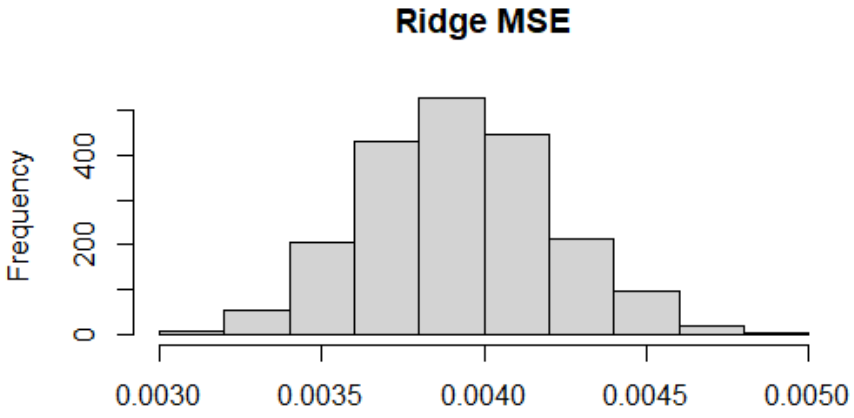

Figure 5: Test MSE distribution with ridge regression approach

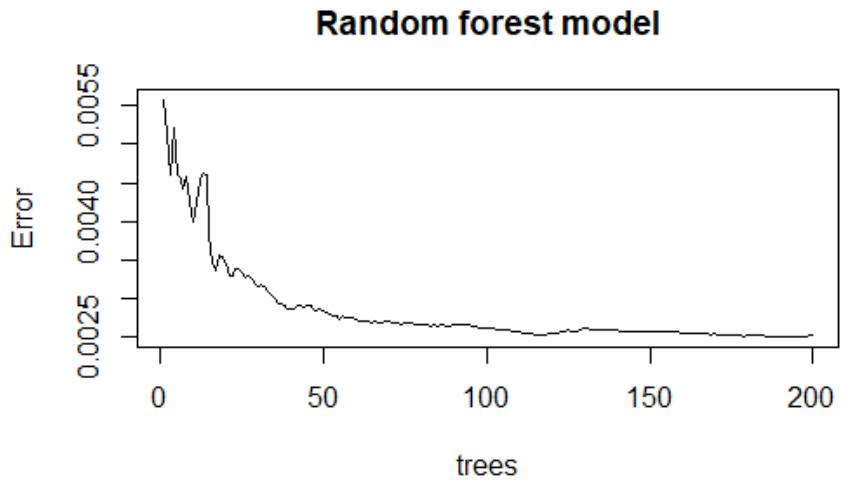

Figure 6: Random forest MSE in function of the number of tree estimated

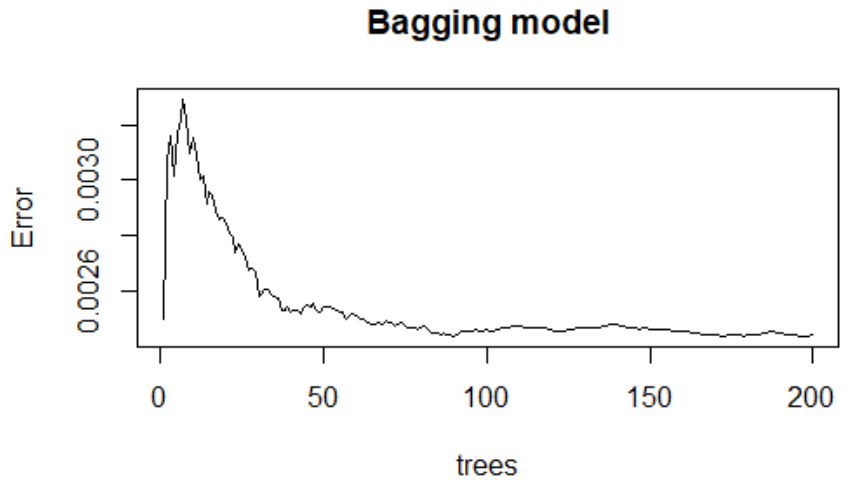

Figure 7: Random forest MSE in function of the number of tree estimated

Tables

Feauture	Description
Sleep efficiency	Percentage
Age	Number of years
Gender	"Male" or "Female"
Bedtime	"Date"
Wake-up time	"Date"
Sleep duration	Number of hours
REM sleep percentage	Percentage
Deep sleep percentage	Percentage
Light sleep percentage	Percentage
Awakenings	Number of awakenings
Caffeine consumption	$[mg/day]$
Caffeine consumption	$[mg/day]$
Alcohol consumption	$[oz/day]$
Smoking status	"Yes" or "No"

Table 1: Summary of features in the dataset

	Coefficients	Lower CI.	Upper CI	Significance	Appearance
Intercept	3.642E-01	2.988E-01	4.332E-01	1.0	2001
Age	9.639E-04	3.920E-04	1.517E-03	1.0	1940
REM.sleep.percentage	6.629E-03	4.575E-03	8.596E-03	1.0	2001
Deep.sleep.percentage	5.550E-03	4.973E-03	6.131E-03	1.0	2001
Awakenings	-3.201E-02	-3.752E-02	-2.638E-02	1.0	2001
Caffeine.consumption	2.327E-04	1.154E-05	4.795E-04	1.0	1391
Alcohol.consumption	-6.306E-03	-1.179E-02	-8.906E-04	1.0	1678
Smoking.status	-4.560E-02	-6.116E-02	-3.046E-02	1.0	2001
Exercise.frequency	6.538E-03	1.770E-03	1.152E-02	1.0	1700

Table 2: Stepwise regression results: first column shows the predictor selected, second, third and fourth the coefficients values and their confidence interval, fifth if the coefficient is relevant and sixth the number of appearance during bootstrap simulation

	Coefficients	Appearance	Significance
(Intercept)	0.00000E+00	0.0000	No
Age	1.06096E-03	1995.0000	Yes
Gender	1.80296E-03	1429.0000	Yes
Sleep.duration	2.57907E-03	0.0000	No
REM.sleep.percentage	5.93846E-03	2001.0000	Yes
Deep.sleep.percentage	5.63776E-03	2001.0000	Yes
Awakenings	-3.07414E-02	2001.0000	Yes
Caffeine.consumption	2.94978E-04	1939.0000	Yes
Alcohol.consumption	-2.70418E-03	1967.0000	Yes
Smoking.status	-4.34432E-02	2001.0000	Yes
Exercise.frequency	4.74424E-03	1984.0000	Yes

Table 3: Lasso regression results: first column shows the predictor selected, second the coefficients value, third the number of appearances during bootstrap simulation and fourth if the coefficient is relevant (not null)

	Prediction	CI_down_predictions	CI_up_predictions
2	7.66307E-01	7.58547E-01	7.73905E-01
3	5.58153E-01	5.52576E-01	5.64462E-01
4	9.03385E-01	8.96440E-01	9.09721E-01
5	5.55855E-01	5.45302E-01	5.67034E-01
6	7.58634E-01	7.50195E-01	7.66603E-01
7	8.90375E-01	8.80690E-01	8.99792E-01
8	8.04695E-01	7.98221E-01	8.11014E-01
9	5.36679E-01	5.30987E-01	5.42852E-01
10	9.42272E-01	9.38933E-01	9.45295E-01
11	8.98390E-01	8.90582E-01	9.06307E-01
12	7.83242E-01	7.74709E-01	7.92018E-01
13	9.01890E-01	8.97513E-01	9.06376E-01
14	7.91452E-01	7.84048E-01	7.98584E-01
15	8.79383E-01	8.66324E-01	8.91783E-01
16	8.06288E-01	7.97813E-01	8.14140E-01
17	8.69718E-01	8.61220E-01	8.78112E-01
18	8.06514E-01	7.97748E-01	8.14644E-01
19	9.04541E-01	8.97378E-01	9.11671E-01
20	7.89760E-01	7.79676E-01	7.99187E-01
21	7.90836E-01	7.85776E-01	7.95860E-01
22	5.66771E-01	5.56609E-01	5.77800E-01
23	9.37062E-01	9.31641E-01	9.41800E-01
24	8.71878E-01	8.55614E-01	8.87149E-01
25	8.05789E-01	7.97368E-01	8.13972E-01
26	5.37287E-01	5.31321E-01	5.44046E-01
27	9.13627E-01	9.08768E-01	9.18067E-01
28	8.61113E-01	8.55510E-01	8.66720E-01
29	7.70743E-01	7.62165E-01	7.78724E-01

Table 4: Prediction made with bagging using bootstrap simulation to provide confidence interval

References

[1] Sleep and Health, Division of Sleep Medicine at Harvard Medical School, https://sleep.hms.harvard.edu/education-training/public-education/sleep-and-health-education-program/sleep-health-education-86

[2] Sleep Efficiency Dataset, University of Oxfordshire, https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency