Bayesian Statistics Assignment

Gaja Nenadović g.nenadovic@students.uu.nl

15 junij, 2022

Data

The data was collected in an online survey on a German sample in 2017, within the study (Volkmer and Lermer, 2018) on the relationship between well-being, satisfaction with life, mindfulness and mobile phone use. I find this study particularly appropriate for this assignment, since it (1) deals with psychological phenomena, where the interest lies in uncovering relationships on the level of *individual and not aggregate*, (2) first introduces 15 prior studies on the topic as a result of authors' meta-analysis and (3) highlights how the previous evidence on the relationship between well-being/satisfaction with life and mobile phone use is contradictory across different studies and there is only one previous study investigating the relationship between mindfulness and mobile phone use.

The variables of interest (with the range of possible values in the parantheses) are:

- well being (0-25): an unweighted total test score on the German version of the WHO-Five well-being index (WHO-5). This questionnaire consists of five items referring to the past two weeks using a 6-point Likert scale (1-6). Higher total score indicates higher well-being;
- satisfaction with life (5-35): an unweighted total score on the German version of the Satisfaction with Life Scale (SWLS). This questionnaire also contains 5 items pertaining to general satisfaction with life on the scale from 1 to 7. Higher score indicates higher satsifaction.
- mindfulness (14-56): an unweighted total score on the German version of the short version of the Freiburg Mindfulness Inventory (FMI). It contains 14 items on a scale from 1 to 4. Higher score indicates higher levels.
- mobile phone use (0-44): an unweighted total score on a German translation of the brief version of the Test of Mobile Phone Dependence (TMDbrief). It refers to abstinence syndrome, lack of control, tolerance development and interference with other activities. It contains 12 items on a scale from 1 to 5. Higher scores indicate higher mobile phone use.

The analysis of choice is a multiple linear regression, with intercept, regression coefficients and standard deviation as unknown parameters to be estimated.

The research questions posed are:

- Do mindfulness, satisfaction with life and well-being predict mobile phone use?
- Is well-being negatively or positively associated with mobile phone use?
- Is mindfulness negatively associated with mobile phone use?
- Which prior results in a better fitting model, uninformative or informative based on the prior study?

Descriptives are presented in Table 1.

estimation, MH, convergence, interpretation of estimates and intervals

Table 1: Descriptive statistics of the data.

	mean	sd	median	min	max	skew	kurtosis
WHO_all	2.398e-16	4.358	0.1866	-11.81	10.19	-0.2587	-0.3837
SWL_all	-1.163e-15	5.822	0.8655	-17.13	10.87	-0.4905	-0.4469
FFA_all	3.435e-15	6.043	0.006508	-14.99	15.01	-0.1687	-0.4861
TMD_all	16.29	8.686	16	0	42	0.1539	-0.3678

Gibbs sampler with a random walk Metropolis step for the regression coefficient of well-being has been used for sampling the parameters from the posterior. Prior for the intercept and regression coefficients is $\mathcal{N}(\mu, \sigma^2)$ and for $1/\sigma^2$ it is $\Gamma(\alpha, \beta)$. Proposal for the random walk step is $\mathcal{N}(\mu, \sigma^2)$, where μ is the previously sampled value, and since it is symmetric, acceptance ratio in each iteration is calculated as posterior density of candidate sampled from the proposal divided with posterior density of the previously sampled value. It is accepted with a probability of min(1,r).

Since there is conflicting prior evidence on the relationships between mobile phone use and well-being and satisfaction with life, with reported effects of well-being and satisfaction with life positive, negative and zero, uninformative priors $\mathcal{N}(0, 1000)$ have been used for these two coefficients and the intercept. There is one prior study suggesting the effect of mindfulness on mobile phone use is negative, so informative prior $\mathcal{N}(-0.39, 0.01)$ with hyperparameters representing the effect size and its squared SE from the study was used for mindfulness. Partly informative prior for $1/\sigma^2$ was used, $\Gamma(\frac{N}{2}, \sigma^2_{resid} \times \alpha)$, where N is the number of participants in the study (302). As σ^2_{resid} I chose 0.001 to express some degree of uncertainty in the previously obtained results of the study, since residual variance in the study was not reported and this is the only study investigating effect of mindfulness on mobile phone use.

All three predictors were highly correlated so they were centered beforehand to aid convergence. The output using these priors was compared with the output obtained using only uninformative priors. The estimates for the informative prior model are presented in Table 2. The estimates for each parameter are across 2 chains, but the means and standard deviations did not differ between the chains by more than 0.01-0.02, which indicates stable results. For the model with the informative prior, we can conclude that well-being and mindfulness are negatively associated with mobile phone. Their posterior means (-0.3282 and -0.2669) and 95% of the posterior values are negative, which means and there is 95% probability that true values lie within the credible intervals. Satisfaction with life has a close to 0 mean coefficient (centered around 0.08956) and its credible interval ranges from negative to positive values, which indicates that it is quite likely that its value is 0 and therefore is not predictive of mobile phone use. For the model with uninformative priors, estimates are slightly different, but the main conclusions stay the same. However, as expected, posterior mean for residual σ^2 is larger (8.428, as a result of using the prior with lower rate and so more spread out prior values) and the evidence for mindfulness is less strong (smaller absolute posterior mean, -0.2232), reflecting prior uncertainty and lack of subjective beliefs about the relationships. This is also reflected in wider credible intervals (and so a more spread out posterior) for all parameters, where a larger range of values is probable.

To assess convergence, in total 10000 iterations of the algorithm with 2000 burn-in iterations and 2 chains were ran and Gelman-Rubin statistic for each parameter was calculated (R in Table 2). For the coefficient obtained by MH algorithm, autocorrelations per chain up to lag 30 were obtained. I will elaborate on it only for the model with informative priors, but it is very similar for the other. G-R statistics are very close to 1 (smaller than 1.1) and MC errors are small for all coefficients, smaller than 5% of SD. Inspections of plots showed good mixing of chains (fat caterpillars), and coefficients seemed to converged after 1000-200 iterations, the fastest being $\frac{1}{\sigma^2}$. except the one estimated with MH step, which needed more iterations. Acceptance rate was 60% overall, 59 in the first chain and 60 in the second, and autocorrelation at the first lag was relatively high, 0.83 for the first chain and 0.78 for the second, but fell under 0.2 at lag 9 and 7 for the first and second chain respectively. This result is not desired and to further explore the effect of the type

of MH sampler and proposal I ran the independent MH sampler keeping all other variables of the sampler constant and with informative prior:

- $\mathcal{N}(\beta, SE^2)$: where parameters are ML estimate and its SE. Autocorrelation for the first chain at lag 1 is 0.80 and 0.79 for the second chain, while acceptance rate is 55% for both chains. This proposal approximates posterior well, but is moving around the posterior very slowly, since the variance is very small (0.11821).
- $\mathcal{N}(0, 1)$: an autocorrelation at lag 1 is 0.92 and acceptance rate is 9% for both chains. This proposal does not approximate the posterior well (the posterior mean obtained previously is around -0.3) and the variance is too large, so many candidates are rejected.

Trying out different values of σ^2 for the first proposal described did not yield much different results, either it was too high and acceptance rate increased or it was too low and autocorrelation increased.

Table 2: Estimates, CI and convergence for the model with informative prior. (continued below)

	Mean	95% CI lwrbd	95% CI uppbd	R
intercept	16.28	15.68	16.87	0.9999
well-being	-0.3282	-0.5045	-0.1532	0.9999
satisfaction with life	0.08956	-0.03871	0.2178	0.9999
mindfulness	-0.2669	-0.3667	-0.1651	0.9999
sigma	6.534	6.206	6.872	0.9999

	MC error
intercept	0.002407
well-being	0.0007074
satisfaction with life	0.0005163
mindfulness	0.0004087
sigma	0.001334

posterior predictive check

The proposed discrepancy measure is calculated as follows: for each sample of parameters, a data set of the same size as the observed is simulated from a normal distribution. For each sample, predicted outcome is calculated for all observations and residuals are calculated for the simulated and original data set. Then, autocorrelations of residuals up to a specified lag are computed for residuals of simulated and observed data separately. Then, the proportion of correlations (in absolute value) higher than 0.2 is computed for the observed and simulated data set and whether the proportion is at least as high in the simulated as for the observed data set. This measure checks the assumption of independence of observations. When the assumption is violated, their residuals are correlated, even after accounting for the predictors' effect. One way to go about this would be to look at the maximum or mean/median autocorrelation present as a discrepancy measure. However, it is possible that the maximum is not much higher than what would be observed if the assumption would hold, but that it remains higher at more lags than would be expected. Similarly, the mean could be not much higher than what would be obtained under the null model, but there would be some extreme autocorrelations present at certain lags. This measure instead looks at the area under the curve for the distribution of autcorrelation values. It is still possible that p value using this discrepancy measure is high, as in the observed data there are not many correlations above the specified threshold, but there are many slightly below it, whereas we would expect only very small (or in ideal situation 0) ones, but

this is not as problematic in terms of analysis. It is also possible that p changes with changing the number of lags we look at, e.g. with higher lags correlations might be a smaller even when the assumption is violated - the proportion of higher than 0.1 correlations would then decrease. This is also not as problematic, since we can check at which lag the autocorrelations stabilize and choose lags accordingly.

Predictive p value for the discrepancy measure of percentage of autocorrelations higher than 0.2 up to lag 40 is 0.35. This means that 35% of future samples would have at least as high proportion of autcorrrelations higher than 0.2 up to and including lag 40 if the assumption of independent observations would hold and we cannot reject the hypothesis that the assumptions holds. By comparison, at lag 10 p is 0.22. Comparing these results with the results of Durbin-Watson test on the observed data gives the same conclusions: the assumption is not violated.

Model selection

As the evidence on the investigated relationships is conflicting, it was hard to formulate good hypotheses. A negative effect of mindfulness is expected, but evidence of effects of well-being and satisfaction with life varied. Looking into 15 prior studies, the number of previous studies investigating the effect of SWL was much smaller and they included, in similar shares, positive, negative and zero effects. On the other hand, all the studies except 1 did find a significant association of well-being and mobile phone use either positive or negative). Therefore, there were no hypotheses about SWL, a hypothesis of negative effect of mindfulness and both hypothesized negative and positive effect of well-being.

$$H_u : \beta_0, \beta_{wb}, \beta_{swl}, \beta_{mind}, \sigma^2$$

.

$$H_1 : \beta_0, \beta_{wb} < 0, \beta_{swl}, \beta_{mind} < 0, \sigma^2$$

.

$$H_2 : \beta_0, \beta_{wb} > 0, \beta_{swl}, \beta_{mind} < 0, \sigma^2$$

.

Furthermore, we wanted to asses which prior results in a better fitting model.

DIC

DIC is useful for models with informative priors, where determination of number of parameters may be more difficult, but can be unstable. It is based on the deviance of the model (given the posterior means and mean deviance across the whole posterior, where deviance is based on the likelihood of data).

Comparing the deviance of the two models with informative and uninformative prior, we see that the model with informative prior actually has a worse fit than the one with the uninformative prior. Observed effect of mindfulness is weaker than what we specified in the informative prior and estimated in the posterior, and since this informative prior also changes estimates of other predictors, the overall model fits the data worse. In the uninformative prior model, data has a large influence in drawing posterior samples and consequently it can overfit, since DIC is based on the likelihood of data given sampled parameters. Using the informative prior, the observed effects in this sample are smaller than the ones we estimated, but it is possible that this model would in fact fit better to future samples.

Choosing only the samples of posterior for the unconstrained model in line with H1 and H2, DIC was calculated for H1 and H2. DIC for the first model is lower than DIC for the second model. Out of all four models, the model proposing negative coefficients for well-being and mindfulness while estimating other parameters freely has the lowest DIC. The results are presented in Table 3. This model thus fits best.

Table 4: DIC values for 4 evaluated models.

Hu informative prior	6192.23287737055
Hu uninformative	4977.93413472652
H1	4975.26776307432
H2	4976.26401591081

Bayes factor

Table 5: Bayes factor for unconstrained H, H1 and H2.

BF1u	BF2u	BF12
6.045	0.0005959	10143
6.106	0.001166	5239
6.33	0.001739	3640
6.252	0.001176	5316
6.457	0.0002923	22090
6.336	0.001767	3585
6.227	0.001751	3557
6.321	0.001795	3521
6.304	0.0005944	10606
6.196	0.0002946	21029

Approximate Bayes Factor for H1 and H2 vs unconstrained was computed as a ratio of the percentage of prior and posterior conforming to the hypothesis. Priors are bivariate normal distributions centered on the boundaries of the hypotheses (0). Covariance matrix of the prior is the same as of posterior, divided by $\frac{J}{N}$. Posterior is a bivariate normal with ML estimates of coefficients and their covariance matrix as parameters. Choosing J as the number of constraints imposed, we see that there is approximately 6 times more evidence for the first hypothesis than for the unconstrained hypothesis, which is moderat support and the effects of mindfulness and well-being are likely negative. The second hypothesis is then not supported well and the corresponding BF is close to 0, that is there is little support for the positive effect of well-being on mobile phone use (also compared to unconstrained hypothesis). Table 4 shows the results obtained by varying J from 1 to 10, where $b = J/N$ and Σ_{post} is multiplied by b, that choosing variance of prior has an effect on BF. We see that BF slightly changes with increasing J, but it remains relatively stable and the conclusions stay the same. Comparing H1 and H2, there is very strong support for the first one (500x stronger for J=2).

Overall, both BF and DIC lead to the conclusion that there is strong support for negative effects of well-being and mindfulness. This has already been seen by looking at the mean and CI obtained in the previous question, where at least 95% of the samples were negative. We can also conclude that for this specific sample, the uninformative prior model fits worse.

comparison of Bayesian and frequentist approaches

There are several differences between the two approaches. I chose this data, because it includes psychological phenomena, where the interest lies in uncovering relationships on the level of an individual. In frequentist approach, probability is defined over frequencies of events or in the limit, which means that we can interpret the results on average, on the level of the aggregate, as expected to be seen if we observed many times. However, if we define probability as a subjective belief, influenced both by observations and prior beliefs, we can interpret the results in terms of probability. In this case, the obtained effects of mindfulness, satisfaction

with life and well-being can be interpreted as the effects we believe are likely to be seen in an individual, taking into account the data and prior evidence and/or belief. In the frequentist approach, we would only say we expect these effects on average, but we cannot say how likely it is, for example, that a person with highly expressed mindfulness would have low levels of problematic phone use.

Moreover, in frequentist approach, information is reduced on a point estimate (with an interval), while in Bayesian, the uncertainty can be captured and taken into account both in the prior and posterior. For example, when specifying the informative prior for $\frac{1}{\sigma^2}$, the rate parameter was set to a low value, indicating lack of certainty about the values of residual variance. Or, we can say that, for example, the effect of satisfaction with life close to 0 is very likely and that the positive effect of mindfulness is highly unlikely, but the actual values are still uncertain. Relating to this and first point, even using confidence intervals does not tell us about the likelihood of our estimate (or how likely a person will exhibit a certain behaviour). whereas in Bayesian models we can conclude how likely it is to observe a certain value of parameter.

Bayesian model allows for more flexibility and specificity, making use of the prior evidence and our prior expectations. The hypotheses I specified and the informative prior were in line with prior studies and thus could be evaluated, providing much more information than what would be obtained by solely testing if a certain effect is significantly different (higher or lower) than 0. In addition, the informative hypotheses could be compared with Bayes factor and we could quantify the support of one hypothesis over other, which again is not possible in frequentist approach, where we could only not reject the null hypothesis (which in my case was not even of interest) but not accept any hypothesis. In frequentist approach, we could assess model fit or calculate p-value but the latter would not give information about magnitude or importance of the effects and would get smaller with increasing sample size.

This described flexibility can result in "bad science", since researchers can tweak the models and selectively use prior information not in line with their previous hypotheses to obtain desired results and not based on strong substantive grounds, which would not aid in theory development and could thus hamper scientific progress. Again, this point applies to much psychological research, including my dataset.

A downside compared with frequentist approach are some computational problems and instability, for example choosing proposal distribution or problems with evaluating convergence of sampled parameters. Such problems would not be encountered with a simple linear regression in frequentist approach.

Literature

Volkmer S.A. & Lermer E., (2018). Unhappy and addicted to your phone? – Higher mobile phone use is associated with lower well-being, Computers in Human Behavior, https://doi.org/10.1016/j.chb.2018.12.015