Winter miles make summer smiles: The effect of frequent bike rides on power output

Thom Volker - t.b.volker@uu.nl

Utrecht University

07-06-2020

1 Introduction

As Lance Armstrong noted early on, success comes from training harder and digging deeper than others (Maxwell, 2007). Although it eventually appeared that there were some other factors that came into play, it cannot be denied that training harder is of fundamental importance if one wants to be faster on his or her bike. To investigate whether this also applies to myself, an ordinary sports enthusiast, with a huge lack of time and without any exceptional talents (and without access to performance enhancing drugs), my own cycling data will be analyzed by means of Bayesian linear regression. The dataset is extracted from my Strava[1] account, where I recently upload my bikerides.

The dataset contains per ride an estimate of my average power output (in watts), the date of the activity, the duration (in hours) and the distance (in kilometers). From the duration and the distance, the average speed can be computed, and based on the dates, the number of bikerides during the four weeks before the actual ride can be computed. The main question of interest is whether the number of activities in four weeks prior to a training can predict differences in my power output, controlling for speed (to make sure that the estimated power output is not biased for rides during which I have had the wind with me) and distance (because plausibly, I cannot both ride fast and ride far). Of secondary interest is whether this model provides an adequate description of the data, or whether other models might fit better. The data is analysed in a Bayesian manner, that is, Gibbs sampling with a Metropolis-Hastings step is used to compute the posterior distribution of the regression parameters. Convergence is assessed by means of trace plots, autocorrelations, the Gelman-Rubin statistic and the MCMC error. To assess normality of the residuals, a posterior predictive check is executed. Additionally, the Deviance Information Criterion DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) is calculated for model comparison, and the individual contribution of regression coefficients is assessed by means of Bayes Factors (BF; Kass & Raftery, 1995) and 95%-Credibility Intervals (CI).

In addition to the initial analyses, the original dataset is extended with some cycling activities from the period in which I started the report, which is easily done within a Bayesian framework. Incorporating prior data can be done by using the posterior from the initial analysis as prior of the second analysis as well as by combining all data into a single analysis. Due to computational simplicity, I adopt the latter approach. That is, initially only a subset of the final dataset is used, namely the part that was present before starting the assignment, and this is extended with some additional data. For the sake of brevity, after combining the results, only the parameter estimates are reported and interpreted.

[1] www.strava.com; a platform where sporters can upload their their activity, either a training or a match, and look into their performances afterwards.

2 Methods

To get an idea of the initially used data, some summary statistics are presented below (Table 1). It can be seen that most of my bike rides are relatively short, with an average distance of 46 kilometers, and an average duration of one hour and 32 minutes. Additionally, my average speed is about 30 km/h and the average number of activities in four consecutive weeks is 6 (like I said, I have a lack of time). Distance and duration are slightly skewed, but nothing too severe, so for now, any deviations from normality are not suspected. Note that differences in sample sizes over the variables exist due to the fact that Strava is sometimes unable to estimate every aspect of the ride. Since missing data is dealt with using listwise deletion, the used data consists of 57 observations.

Table 1: Descriptive statistics of the data used in the analyses

	n	mean	sd	median	min	max
Duration	76	1.530	0.581	1.401	0.254	4.101
Distance	88	46.129	16.398	43.694	7.423	110.165
Power output (watts)	60	155.540	18.193	154.985	113.570	201.782
Speed (km/h)	76	30.379	2.314	31.013	24.019	34.542
Activities (4 weeks before activity)	88	6.170	4.259	6.000	0.000	17.000

Initially, the analysis is set up with uninformative normal conjugate priors for the regression coefficients with a mean of $\mu_{0j} = 0$ and a variance of $\tau_{0j}^2 = 10000$. The prior distribution for the variance is the inverse of a Gamma distribution with shape and rate parameters $\alpha_0 = .0001$ and $\beta_0 = .0001$, respectively. Using normal priors for the regression coefficients yields a normal posterior distribution, and using an Inverse-Gamma prior for the variance leads to a posterior Inverse-Gamma distribution. Initial values are randomly drawn from a normal distribution for the regression coefficients, and from a log-normal distribution for the variance. Additionally, the effect of distance on average power output is estimated by means of an independent Metropolis-Hastings step within the Gibbs sampling procedure. To account for the uncertainty with regard to the effect of distance, a t-distribution is used as prior since it has somewhat fatter tails than a normal distribution, with mean $\mu_{0_{distance}} = 0$, variance $\sigma^2_{0_{distance}} = 10000$ and degrees of freedom $\nu_{0_{distance}} = 2$, which yields a posterior distribution without a closed form solution. However, since the conditional posterior of $\beta_{distance}$ is known up to a proportionality constant, the Metropolis-Hastings sampler can be used. The proposal distribution is normal with a mean equal to the maximum likelihood estimate of $\beta_{distance}$ and its corresponding variance, which are both equal to the ordinary least squares estimates, under the assumption that the errors are distributed normally, which will be tested later on by means of the posterior predictive check.

3 Results

3.1 Convergence

Looking at the traceplots, it seems to be the case that both the Gibbs sampler and the Metropolis-Hastings sampler move freely through the parameter space. All three chains are overlapping, which seems to indicate that all three chains sample throughout the same distribution. Additionally, the autocorrelations remain rather low. That is, for all parameters that are sampled by means of Gibbs sampling, the autocorrelations are about zero from the first lag onwards, and the regression coefficient of distance has an autocorrelation

of about zero from the second lag onwards. Furthermore, the acceptance rate of the sampled values in the Metropolis-Hastings step is 0.756, and obviously, the acceptance rates of the parameters sampled by means of Gibbs sampling are all 1. Additionally, the Monte Carlo error (Naive SE) is sufficiently low for all variables ($< 5\%$ of the standard deviation of the sampled values), and the Gelman-Rubin diagnostic is equal to one, rounded after three decimals (Table 2). Although it is not possible to prove that the sampling algorithm converged, all indicators of convergence do not raise suspicions.

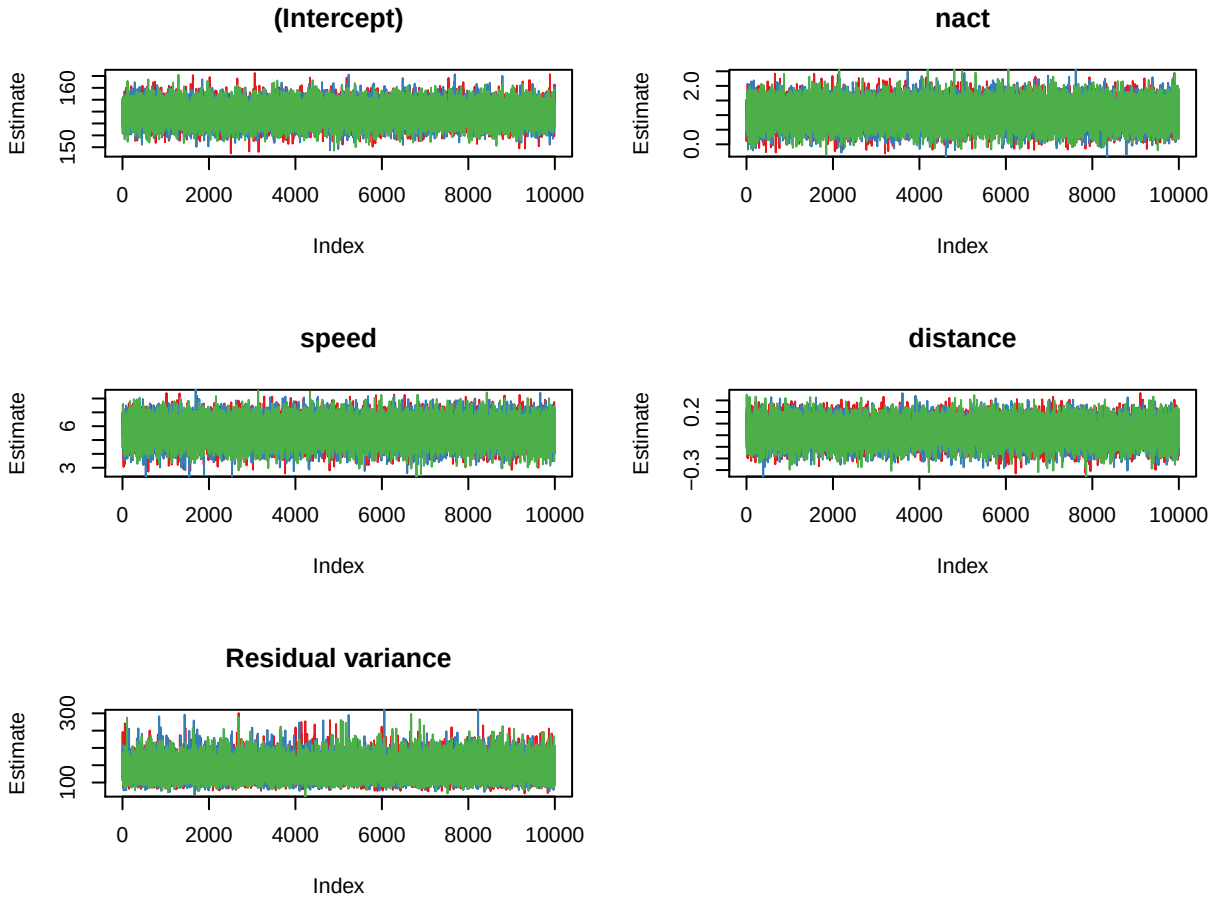

Figure 1: Traceplots from the parameter estimates, using three chains, for the intercept and residual variance, and the variables *number of activities (nact), speed* and *distance*.

3.2 Parameter estimates

Table 2: Bayesian linear regression results of the effect of the distance of the training, the average speed and the number of activities four weeks beforehand on the average power output during a training.

	Mean	SD	2.5%	97.5%	Naive SE	Gelman-Rubin	Acceptance rate
Intercept	155.680	1.578	152.590	158.792	0.009	1	1.000
Activities	1.043	0.381	0.292	1.792	0.002	1	1.000
Speed	5.603	0.801	4.006	7.184	0.005	1	1.000
Distance	0.027	0.092	-0.154	0.207	0.001	1	0.756
Residual Variance	135.873	27.453	92.445	199.178	0.158	1	1.000

As can be seen in Table 2, the number of activities four weeks prior to a training is related to the power output. Every additional activity is related to 1.04 added watts to my power output. The credibility interval indicates that there is a 95% probability that the true parameter estimate is within the boundary values of 0.29 and 1.79. Additionally, there is a relationship between speed and power output ($\beta_{speed} = 5.6$). That is, for every additional kilometer per hour that I go faster, the estimated power output increases with 5.6 watts (CI: 4.01; 7.18). A relationship between power output and distance is not present ($\beta_{distance} = 0.03$; CI: -0.15; 0.21), indicating that after controlling for speed and the number of activities, I do not tend to use more power during shorter rides.

3.3 Model comparison

Additionally, it is of interest whether the number of activities improves the model fit, as evaluated by the DIC. The DIC balances the fit of the model with the complexity of the model, and aims at providing the simplest best model. The model described above has a DIC of 437.38, but this value is meaningless on itself and has to be interpreted relative to the DIC of other models. Therefore, the uninformative predictor distance is removed from the first model, (i.e., in the new model, the average power output is regressed on speed and number of activities), which yields a DIC of 436.33,[2] which is slightly lower than the previous DIC, indicating that this model is slightly, but for from conclusive, better than the previous one, mainly due to it's simplicity. Additionally, a model in which speed serves as the only predictor of the average power output is considered, which yields a DIC of 442.00,[3] indicating no improvement as compared to the previous model. Therefore, considering all evidence collected thus far, the number of activities prior to a ride seems to be related to my average power output.

3.4 Bayes Factors

Before the Bayes Factors can actually be computed, the data has to be standardized to get meaningful results. Then, based on the model containing all three parameters, the following informative hypotheses were considered, whether or not the effects of speed number of activities were both positive, and whether or not the effect of the number of activities was larger than the effect of speed:

$$H_1 : \beta_{speed} > 0, \beta_{activities} > 0 \qquad H_2 : \beta_{activities} > \beta_{speed}.$$

Note that these hypotheses were constructed a priori, and are thus not dependent on the previous analyses.

The Bayes Factor of the first hypothesis H_1 was equal to 5.29, indicating more than 5 times more support as compared to the unconstrained hypothesis. The Bayes Factor of the second hypothesis H_2 was equal to 0.01, indicating a lack of support. The support of the first hypothesis H_1 in favor of the second hypothesis H_2 then is 577.14. Additionally, posterior model probabilities can be calculated. However, to do this, a "fool proof" unconstrained hypothesis (i.e., the classical alternative hypothesis H_A), is added to the set of hypotheses, to prevent placing to much confidence in a hypothesis that is inappropriate, but the best of a set of inappropriate hypotheses:

$$H_A : \beta_{speed}, \beta_{activities}, \beta_{distance},$$

where no constraints are placed on the hypotheses. Then, the posterior model probabilities can be computed as the Bayes Factor of the hypothesis of interest, divided by the sum of the set of Bayes Factors, which yields a posterior model probability of 0.84 for H_1. This can be interpreted as the relative support for the

[2]Model diagnostics have been assessed, but are equivalent to the diagnostics discussed for the first model.
[3]For this model as well, diagnostics discussed thus far have been inspected, and did not yield any suspicions of non-convergene.

first hypothesis on a 0-1 scale. Thus, this hypothesis clearly receives the most support of all models under consideration, which does not exclude the possibility that a different hypothesis fits the data even better.

3.5 Posterior predictive check

To check whether the model was in accordance with one of the most important assumptions of linear regression, normality of residuals, a posterior predictive check was executed. This was done by means of comparing the skewness of the observed residuals after each iteration (S^t_{obs}) with the skewness of residuals simulated under a $\mathcal{N}(0, \sigma^2_t)$ (S^t_{sim}), where σ^2_t refers to the modeled variance in each iteration. The proportion of datasets of simulated values that have a skewness larger than the skewness of the corresponding observed datasets,

$$P(abs(S^t_{sim} \mid [\mathbf{Y}^t, \mathbf{X}, \beta^t, \sigma^{2,t}]) > abs(S^t_{obs} \mid [\mathbf{Y}, \mathbf{X}, \beta^t, \sigma^{2,t}], H_0)),$$

can be regarded as a Bayesian measure of evidence against the null hypothesis of normally distributed residuals. Namely, if none of the absolute skewnesses of the residuals simulated under the assumption of normality is larger than the skewnesses of the corresponding observed residuals under the model, it is very unlikely that the observed residuals are normally distributed. However, no deviations from normality were present, and the posterior predictive p-value equals 0.524. That is, it is quite as likely to observe a larger skewness for the simulated residuals as compared to the observed residuals, as to observe a smaller skewness for the simulated residuals.

Additionally, the correlations between the ordered simulated residuals in every (simulated) dataset and the theoretical quantiles of a standard normal distribution (ρ_{sim}) are compared with the correlations between the observed residuals over all sampled values and the theoretical quantiles from a standard normal distribution (ρ_{obs}). If the correlation between the ordered residuals and the theoretical standard normal quantiles is equal to one, the residuals exactly follow the pattern that is expected under normality. Thus, if $-\rho_{obs}$ is not systematically larger than $-\rho_{sim}$, the deviations of the observed residuals from the expected residuals under normality can be regarded as random fluctuations. This setup yields

$$P(-\rho_{sim} \mid [\mathbf{Y}^t, \mathbf{X}, \beta^t, \sigma^{2,t}] > -\rho_{obs} \mid [\mathbf{Y}, \mathbf{X}, \beta^t, \sigma^{2,t}], H_0),$$

which is the proportion of simulated correlations times minus 1 that is larger than minus 1 times the corresponding observed correlation between the ordered residuals and the theoretical standard normal quantiles. However, there are no suspicions of deviations from normality, as indicated by the Bayesian p-value of 0.479. Thus, it can be concluded that the residuals are not skewed, and follow the quantiles of a normal distribution.

3.6 Sequential updating

One of the advantages of the Bayesian approach is that it is very easy to incorporate new data. Since the start of this project, I have of course ridden the bike a couple more times, so it is interesting whether including this data strengthens the results. Incorporating the previously analysed data can be done by specifying the information in the priors, but it can also be done by simply adding the data into one dataset, and analyzing this complete dataset. The latter approach is chosen due to it's computational simplicity. Apart from not using the Metropolis-Hastings step for the effect of distance, the prior specifications for all parameters remained the same (i.e., normal, uninformative priors were used).

First, the analysis is performed on the newly gathered data, which yields a very small but positive effect of $\beta_{activities} = 0.02$, that is indistinguishable from zero ($CI : -2.82, 2.68$), so based on this data, the presence

of an effect cannot be concluded. However, if all collected data is analyzed, it can be seen that the effect after the initial analyses shrunk somewhat, from 1.04 to 0.94 ($CI : 0.19, 1.69$), after controlling for speed and distance. Additionally, the residual variance increases somewhat, from 135.87 to 146.61, after adding more data, indicating that the new data is not completely in line with the previous data. The reason hereof might be that I was involved in a crash, resulting in some minor injuries, after which I needed a couple of rather slow rides to recover that might have been hard to predict.

4 Conclusion

Overall, it can be concluded that training is related to an increase in my average power output, since the credibility intervals, the DIC and the Bayes Factors all indicate that there is an effect of the number of activities on my average power. However, although this might hold for multiple people besides me, it might not hold for everyone. Additionally, people with already a very high training frequency might not benefit from further increasing this frequency, or it might even be harmful in such instances.

Although a frequentist analysis would have yielded the same conclusion, it would lack flexibility. Instead of obtaining a single point estimate (with confidence intervals), Bayesian analyses provide the complete posterior distribution, and additionally, it is even possible to create new test statistics that serve the purpose of the test at hand. In my opinion, the flexibility of Bayesian statistics is why it stands out, especially since it provides the opportunity (although some would say the necessity) to use prior information and to update (i.e., adding new data to the same analysis). From a practical viewpoint, one does not always know beforehand how much certainty the data can provide, and having the opportunity to update your model in order to get more confidence about your estimates yields more accurate results in the long run. This is not to say that new data may not increase the uncertainty about the estimates, as is also shown in the evidence aggregation method displayed in the last section. However, such discrepancies are interesting on itself, and warrant that researchers investigate the causes of discrepancies in further depth.

Additionally, I would say that easily updating the current state of knowledge with the possibility to directly add this information to your estimates (incorporating it in the posterior distribution) is more sensible than having to compare multiple studies indirectly, by looking at different (frequentist) confidence intervals, or by having to perform a new follow up study to combine the results (i.e., perform a meta analysis). That should make it easier to cross-validate and aggregate scientific evidence, which is, in my opinion, very much needed to strengthen, or even regain, the public's trust as well as the trust of the scientific community itself, and I believe that moving from a framework of hypothesis testing to a framework that easily aggregates evidence can help to establish this. Or, to stay in terms in cycling, one cannot win the tour by having a single good day, one has to establish supremacy over the complete three weeks; the rankings are updated after every stage, and eventually, after three weeks of combining evidence, the truely best rider wins the race.

5 References

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Maxwell, J. C. (2007). *Talent is never enough: Discover the choices that will take you beyond your talent* (1st ed.). New York: Harper Collins.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.