Data and research questions

We will be using the NHANES data available in R. These are survey data from the US National Health and Nutrition Examination Study (NHANES), collected by the US National Center for Health Statistics (NCHS). The NCHS has conducted a series of health and nutrition surveys since the early 1960's, and since 1999 approximately 5,000 individuals complete the health examination component of the survey every year. The data available in R consists of 75 variables available for the 2009-2010 and 2011-2012 sample years. We will use a subset that contains only the data of the adult participants for the year 2011-2012, and in which the predictors are mean centred to aid interpretation of the intercepts.

The dataset contains several health variables, such as measurements relating to cholesterol, diabetes, depression, and sleep. High-density lipoprotein (HDL) cholesterol is considered to play an important role in overall health. It is also known as the "good" cholesterol, as it reduces and removes Low-density lipoprotein (LDL) cholesterol from the body (WebMD, 2020). As a result, high levels of HDL cholesterol can lower the risk for heart disease and stroke (CDC, 2020).

As HDL cholesterol is measured directly from serum, blood needs to be taken and analysed in a laboratory in order to determine the amount of HDL cholesterol. It would be useful to be able to predict whether a patient is at risk of having low levels of HDL cholesterol with variables that are more easily measured or reported. Therefore, we will evaluate for some of the other variables available in the dataset whether they are predictors of HDL cholesterol levels. We will specify and compare two different models: one with health variables that are measured externally, and one with lifestyle variables that are self-reported. The variables used are elaborated on below.

Outcome:

DirectChol	Direct HDL cholesterol in mmol/L

Predictors model 1:

Pulse	60 second pulse rate
BMI	Body mass index (weight/height2 in kg/m2).

Predictors model 2:

PhysActiveDays	Number of days in a typical week that participant does moderate or vigorous-intensity activity
SleepHrsNight	Self-reported number of hours study participant usually gets at night on weekdays or workdays

Estimation

To estimate the parameters, prior distributions need to be specified for each of them. For Model 1, we specify an uninformative normal prior for the intercept with mean = 0 and variance = 1000. For the coefficient of the first predictor Pulse, the prior specified is normal with mean = 0.01 and variance = 100. This is based off previous research on pulse and HDL cholesterol levels, which shows a positive relationship between the two (Toprak, Reddy, Chen, Srinivasan, & Berenson, 2009). As there is not much research done on this specific relationship, however, we are not very certain and the variance is therefore chosen to be quite large. For the second predictor, BMI, we have more certainty. More research has been done on the relationship between BMI and HDL cholesterol and it is well known that a higher BMI is associated with lower levels of HDL cholesterol (Bertiere et al., 1988; Arai et al., 1994; Stadler et al., 2021). Based on this information, we specify a normal prior with mean = -0.05 and variance = 10. For the residual variance, an uninformative inverse-gamma prior is specified with the shape and scale parameter both set to 0.001.

For Model 2, we specify an uninformative normal prior for the intercept with mean = 0 and variance = 1000. There is a relatively large amount of research done on the relationship between HDL cholesterol and physical activity, which points towards a positive association (Fonong et al., 1996;

Marrugat et al., 1996; Kokkinos & Fernhall, 1999; LeCheminant, Tucker, Bailey, & Peterson, 2005). For the coefficient of the first predictor PhysActiveDays, we therefore specify a normal prior with mean = 0.01 and variance = 10. Likewise, there is generally a positive association between the amount of sleep and HDL cholesterol levels (Nadeem et al., 2014; Fobian, Elliott, & Louie, 2018). For the coefficient of the second predictor SleepHrsNight, we therefore use a normal prior with mean = 0.02 and variance = 10. For the residual variance, again an uninformative inverse-gamma prior is specified with the shape and scale parameter both set to 0.001.

Metropolis-Hastings

Through Gibbs sampling, we approximate the posterior of the parameters. First, we use the initial values specified before to derive the conditional posterior mean and the conditional posterior variance for the intercept b0. Using these values, we sample a value from the conditional posterior of b0. In the second step, we use this new value b0 in deriving the conditional posterior mean and variance for the slope of the first predictor b1. Then, we sample a value from the conditional posterior of b1 using this mean and variance. For the slope of the second predictor b2, a random walk Metropolis step is implemented as an alternative approach. In this step, we use the new values we have for b0 and b1 to again derive the conditional posterior mean and variance. We sample a candidate value for b2 from a proposal distribution. Then, we sample a random value u from the uniform distribution on the interval from 0 to 1. With the conditional posterior mean and variance, we can compute the target densities of the current and candidate values of b2. The target densities are used to compute an acceptance ratio r. In the final part of the Metropolis step, the candidate value is accepted if $u \leq r$, and the current value is retained if $u > r$. The last step of the Gibbs sampler is for the posterior of the residual variance. Here, we use the new values for b0, b1, and b2 to derive the conditional posterior shape parameter alpha and rate parameter beta. With alpha and beta, we sample a value from the conditional posterior of the residual variance. This process is repeated over 10000 iterations, through which the newly sampled values are used in sampling the subsequent values in each new loop.

Convergence

There are multiple techniques to assess convergence. We will be looking at both the history or trace plots of the parameters and the MC error. The trace plots for the parameters of Model 1 are displayed in Figure 1, and those for Model 2 are displayed in Figure 2. The trace plots show the two chains of the sampled parameters over the iterations, excluding the burn-in of a 1000 iterations. It is noticeable that in both models, the trace plot for b2 does not converge as well as the ones for the other parameters. This is because this parameter is sampled using the Metropolis step. The Metropolis algorithm causes the sampled value to get 'stuck' at times, when the candidate value is continuously rejected. Especially for b2 of Model 1, the acceptance rate seems to be quite low. In Model 2, however, it converges better. For the other parameters, the development of the plots is stable and the two chains mix well. The chains are not stuck at a local maximum for any of the parameters. The output gives us the MC error, which is the standard deviation divided by the square root of the number of iterations. The MC error is well below the threshold of 5% of the sample standard deviation for all the parameters. Overall, we will conclude that the models seem to converge reasonably well.

Figure 1. Trace plots for the first model.

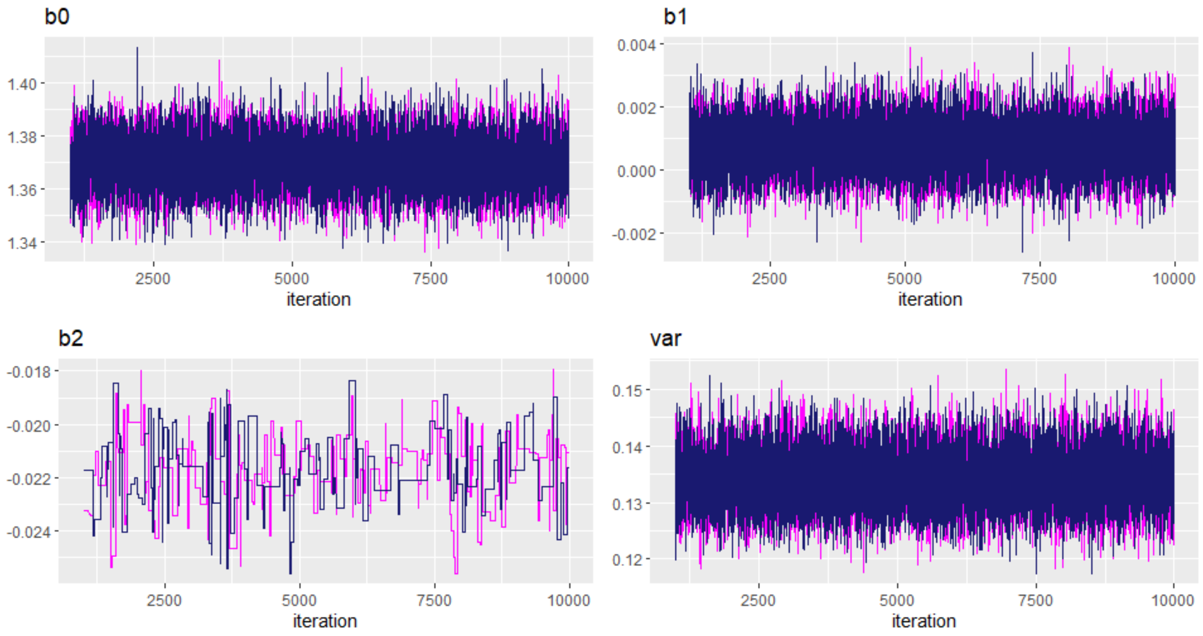
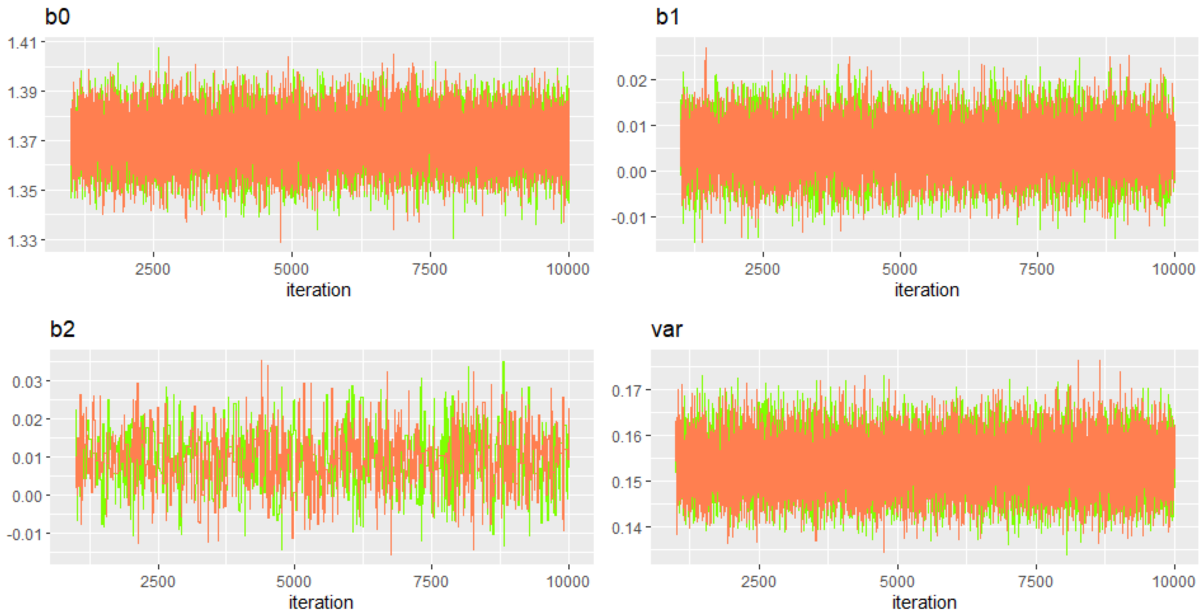


Figure 2. Trace plots for the second model.



Interpretation of estimates and intervals

The output gives us the posterior means, standard deviations, lower (2.5%) quantiles, and upper (97.5%) quantiles, from which we can derive the 95% central credible intervals (CCIs). The means, standard deviations, and 95% CCI for both Model 1 and Model 2 are displayed in Table 1. For both models, the posterior mean direct HDL cholesterol is approximately 1.37 mmol/L for a person with an average pulse and an average BMI. There is a 95% probability that the population value for a parameter will lie in the 95% CCI, given the data. Therefore, we are less sure about the variables with a 95% CCI that contains zero being predictors of direct HDL cholesterol. In this case, we are only quite sure about BMI in Model 1 being a predictor. The parameter estimate indicates that for every unit increase in BMI, direct HDL cholesterol decreases with -0.02168. This implies that people with a higher BMI would be at a higher risk of having low HDL levels.

Table 1. Posterior parameter estimates obtained with the Gibbs sampler.

	Model 1		Model 2	
	M (SD)	95% CCI	M (SD)	95% CCI
Intercept	1.37023 (0.00918)	[1.35217, 1.38821]	1.37026 (0.00982)	[1.35090, 1.38946]
Pulse	0.00075 (0.00077)	[-0.00079, 0.00227]		
BMI	-0.02168 (0.00134)	[-0.02440, -0.01900]		
PhysActiveDays			0.00506 (0.00540)	[-0.00566, 0.01563]
SleepHrsNight			0.00927 (0.00715)	[-0.00543, 0.02311]
σ^2	0.13414	[0.12512, 0.14389]	0.15342	[0.14312, 0.16456]

Posterior predictive check

We will check the assumption of normality for Model 1. This is the assumption that for any fixed value of pulse and BMI respectively, the dependent variable HDL cholesterol is normally distributed. To check this assumption, we can compute a posterior predictive p-value. In order to do this, we first establish what is the null model. In this case, it is the assumed model. We need the posterior distribution of the parameters. We obtained these before with the Gibbs sampler. Then, we use the sampled parameter values to simulate datasets. The resulting collection of datasets is the posterior predictive distribution. The simulated data sets are the ones we would expect if our null model were true. We want to know if the observed data actually resembles these simulated datasets. For this, we compute a test statistic for both the observed dataset, and for each of the simulated datasets. Then, we calculate the proportion of times that the test statistic of a simulated data set is larger than that of the observed data set. This proportion is the posterior predictive (or Bayesian) p-value. If the value is around 0.5, we can say that the assumption is not violated: we conclude that the data come from our null model. In this case, it is 0.001, which is not close to 0.5 at all. We conclude that our data do not come from the null model and that the assumption of normality is violated for Model 1.

Model selection using the DIC

As the models are not nested, it will be useful to compare them with an information criterion. The DIC is a criterion that quantifies the performance of a model in terms of misfit and complexity. As a model is ‘better’ when it fits the data well and it is simple, the model with the smallest DIC would best predict a replicate dataset of the same structure as the observed dataset. In this case, Model 1 has a DIC of 1317 and Model 2 has a DIC of 1525, and therefore Model 1 would be preferred. The rule of thumb is that a difference larger than 10 indicates that the model with the higher DIC is to be ruled out. The difference is 208, and therefore much larger than 10. Based on the DIC, we select Model 1.

Model selection using the Bayes Factor

Another approach for selection is that using the Bayes Factor (BF). The BF quantifies the relative support in the data for a set of hypotheses. We will compare hypotheses for Model 1. The informative hypothesis is that the slope of Pulse is larger than 0 and that the slope of BMI is smaller than 0. We will compare this to the hypothesis where they are both equal to 0. Therefore:

H_1 = b1 < 0 & b2 > 0
H_2 = b1 = 0 & b2 = 0

The result is:

BF_{12} = (fit H_1 / complexity H_1) / (fit H_2 / complexity H_2) = 3.697

This means that H_1 is 3.697 times as likely as H_2. As interpreting the size of the BF by means of a threshold would lead to the same problems as in NHST, we will not do this. However, we can say that there is more support for Model 1 with the slope for Pulse being positive and the slope for BMI being negative, than for the model where these slopes are equal to 0, and the variables are not predictors.

Comparison of Bayesian and frequentist approaches

In order to compare the Bayesian approach that has been used thus far, a frequentist analysis of the same data has been conducted. The means, standard errors, and 95% confidence intervals are displayed in Table 2. We can see that the estimates are quite similar to those obtained with the Gibbs sampler. We would also come to the same conclusion about which variables are predictors of direct HDL cholesterol and which are not: BMI is the only predictor that is significant in the frequentist results, and the only predictor of which the 95% CCI does not contain zero in the Bayesian results. There are, however, noteworthy differences between the two approaches. First of all, there is a difference in interpretation. Even though we would come to the same conclusion in terms of which variables are predictors, the frequentist results are more definitive and lead to dichotomous decision-making on which predictors are significant. With the Bayesian results, it is not as black-and-white. For example, the 95% CCI of Pulse is [-0.00079, 0.00227] and does not contain zero, but it does 'lean' more towards the direction of a positive value. Therefore, this result does not completely rule out that Pulse is a predictor of HDL. In addition, the Bayesian approach offers the possibility to evaluate the model by computing the BF, which does show support for the model in this case. Second of all, prior information is incorporated in the Bayesian results. Especially with topics in health research, it can be valuable to use prior results so as not to 'waste' the often large amount of information provided by studies done on similar subjects.

Table 2. Parameter estimates obtained with a frequentist approach.

	Model 1		Model 2	
	M (SE)	95% CI	M (SE)	95% CI
Intercept	1.37023 (0.00920)*	[1.35219, 1.38827]	1.37023 (0.00984)*	[1.35093, 1.38952]
Pulse	0.00128 (0.00078)	[-0.00025, 0.00280]		
BMI	-0.02187 (0.00144)*	[-0.02470, -0.01904]		
PhysActiveDays			0.00500 (0.00541)	[-0.00561, 0.01560]
SleepHrsNight			0.00929 (0.00739)	[-0.00521, 0.02379]
Residual SE	0.366		0.3915	

* Significant at the p < .001 level

References

Arai, T., Yamashita, S., Hirano, K. I., Sakai, N., Kotani, K., Fujioka, S., ... & Shinohara, E. (1994). Increased plasma cholesteryl ester transfer protein in obese subjects. A possible mechanism for the reduction of serum HDL cholesterol levels in obesity. *Arteriosclerosis and thrombosis: a journal of vascular biology, 14*(7), 1129-1136.

Bertiere, M. C., Fumeron, F., Rigaud, D., Malon, D., Apfelbaum, M., & Girard-Globa, A. (1988). Low high density lipoprotein− 2 concentrations in obese male subjects. *Atherosclerosis, 73*(1), 57-61.

CDC. (2020). *LDL and HDL Cholesterol: "Bad" and "Good" Cholesterol.* Retrieved from: https://www.cdc.gov/cholesterol/ldl_hdl.htm

Fobian, A. D., Elliott, L., & Louie, T. (2018). A systematic review of sleep, hypertension, and cardiovascular risk in children and adolescents. *Current hypertension reports, 20*(5), 1-11.

Fonong, T., Toth, M. J., Ades, P. A., Katzel, L. I., Calles-Escandon, J., & Poehlman, E. T. (1996). Relationship between physical activity and HDL-cholesterol in healthy older men and women: a cross-sectional and exercise intervention study. *Atherosclerosis, 127*(2), 177-183.

Kokkinos, P. F., & Fernhall, B. (1999). Physical activity and high density lipoprotein cholesterol levels. *Sports Medicine, 28*(5), 307-314.

LeCheminant, J. D., Tucker, L. A., Bailey, B. W., & Peterson, T. (2005). The relationship between intensity of physical activity and HDL cholesterol in 272 women. *Journal of Physical Activity and Health, 2*(3), 333-344.

Marrugat, J., Elosua, R., Covas, M. I., Molina, L., Rubies-Prat, J., & Marathom Investigators. (1996). Amount and intensity of physical activity, physical fitness, and serum lipids in men. *American journal of epidemiology, 143*(6), 562-569.

Nadeem, R., Singh, M., Nida, M., Waheed, I., Khan, A., Ahmed, S., ... & Champeau, D. (2014). Effect of obstructive sleep apnea hypopnea syndrome on lipid profile: a meta-regression analysis. *Journal of Clinical Sleep Medicine, 10*(5), 475-489.

Stadler, J. T., Lackner, S., Mörkl, S., Trakaki, A., Scharnagl, H., Borenich, A., ... & Marsche, G. (2021). Obesity Affects HDL Metabolism, Composition and Subclass Distribution. *Biomedicines, 9*(3), 242.

Toprak, A., Reddy, J., Chen, W., Srinivasan, S., & Berenson, G. (2009). Relation of pulse pressure and arterial stiffness to concentric left ventricular hypertrophy in young men (from the Bogalusa Heart Study). *The American journal of cardiology, 103*(7), 978-984.

WebMD. (2020). *HDL Cholesterol: The Good Cholesterol.* Retrieved from: https://www.webmd.com/cholesterol-management/guide/hdl-cholesterol-the-good-cholesterol