The effect of maternal age and smoking on birth weight of babies

Kyuri Park 5439043 (k.park@uu.nl)

Bayesain Statistics - Utrecht University

13-06-2022

1 Introduction

Welcoming a new life into the world is one of the most blissful and surreal moments in our lives. Recently, I have become an auntie and met a beautiful baby girl, who is officially my first niece. Everything was perfect with her, except that she came out a few weeks before the supposed due date. She weighed much less than the other full-term babies, which was initially concerning, but the baby has been doing well so far. Yet, my sister kept blaming herself that the preterm birth was her fault. It got me thinking whether the maternal factors are indeed associated with the baby's premature birth. To investigate the relationship between the preterm birth and maternal factors, the birthweight dataset (observational data) contributed by Ellen Marshall from University of Sheffield[1] is analyzed employing Bayesian linear regression model. The dataset contains 14 variables in total (i.e., id, baby's length, baby's weight, baby's head circumference, mother's smoking habit, mother's height, mother's age, father's smoking habit, father's height, father's age, father's education years, etc.), but in this analysis I look at the baby's birth weight, mother's age, and the number of cigarettes smoked by mothers per day.

The dependent variable is the baby's weight measured at birth (kilogram), which is highly correlated with the gestation period; the premature baby has a lower birth weight (Stanford Children's Health, 2022). The independent variables are mother's age (year) and the number of cigarettes smoked by mothers (per day), which are of interest as relevant maternal factors. Table 1 below shows the summary statistics of the data that are used in this analysis. It can be seen that the data is relatively small (n = 47). The baby's weight (birthweight) seems to be symmetrically distributed while centered around 3. Mother's age is slightly skewed and the number of cigarettes smoked is highly skewed to the right with a fairly large dispersion (sd).

Table 1: Descriptive statistics of the data

	n	mean	sd	median	min	max	skew
Birthweight (kg)	47	3.09	0.62	3.1	1.92	4.55	0.22
Mother age (yr)	47	27.98	7.43	27.0	18.00	47.00	0.66
Number of cigarettes smoked (per day)	47	8.55	12.09	1.0	0.00	50.00	1.46

The main research question is whether there is a significant relationship between the birth weight of the babies and chosen maternal factors (i.e. mother's age and the number of cigarettes smoked by mothers). The secondary research questions of interest are what is the direction of the relationship (i.e., positive, negative) if there is, and which of the maternal factors has more influence on the baby's birth weight. Lastly, in addition to the main analyses, the average causal effects (ACE) of the maternal factors are explored in Bayesian framework. Bayesian causal inference is a fairly new area of research and what could be shown here is rather limited. Regardless, I add this part, as the causal inference is a very relevant and crucial aspect of this type of analyses, where the research questions encompass causality (e.g., the *effect* of maternal factors on birth weight of babies) but only equipped with the observational data. Hence, this last additional part is deemed to be a basic illustration for a potential extension to this analysis, which can be studied further in the future.

[1] https://www.sheffield.ac.uk/mash/ statistics/datasets

2 Methods

In order to investigate the main research question, first the posterior distributions of the regression parameters are estimated using the Gibbs sampler. Here I used the normal conjugate priors for the intercept (β_0) and regression coefficient for the mother's age (β_2), which leads to the normal conditional posterior distributions. For the precision ($\tau = \frac{1}{\sigma^2}$), gamma conjugate prior is used, which leads to the gamma conditional posterior distribution. To accommodate the relatively large dispersion present in the number of cigarettes smoked by mothers, the t-distribution that has heavier tails than a normal distribution is used as a prior for the regression coefficient for the number of smokes (β_1). As there is no closed-form expression available for the normalizing constant of the conditional posterior distribution in this case, the random walk Metropolis-Hastings (MH) step is incorporated. MH sampler may lack efficiency compared to the Gibbs sampler, but it allows more flexibility with regard to choosing a prior, as it can be used when the conditional posterior is known only up to the proportionality constant. The proposal density here is chosen to be normal with the mean equal to the previous sampled value ($\beta_{1,t-1}$) and the variance of 0.01, which is tuned based on trial and error. Given that I barely have any prior knowledge, all the aforementioned prior distributions are specified to be uninformative (i.e., mean of zero and large variance (σ^2) for normal priors, mean of zero, large variance (σ^2) and small degrees of freedom (ν) for t prior, small shape (α) and rate (β) parameter for the gamma prior). The resulting model is thus as follows:

$$Birthweight_i = \beta_0 + \beta_1 \cdot Number\ of\ Smokes_i + \beta_2 \cdot Mother's\ age_i + \epsilon_i, where\ \epsilon_i \sim N(0, \tau^{-1})$$

$$\beta_0, \beta_2 \sim N(\mu = 0, \sigma^2 = 100^2), \qquad \beta_1 \sim t(\mu = 0, \sigma^2 = 100^2; \nu = 3), \qquad \tau \sim Gamma(\alpha = 0.001, \beta = 0.001)$$

In total, two chains run in parallel with different starting values that are randomly chosen. Out of 100000 iterations, the first 1000 samples are discarded (burn-in period). Subsequently, the convergence is assessed by checking the trace plots, autocorrelations, Gelman-Rubin statistic, and MC error. It is followed by assessing a couple of the model assumptions by means of the posterior predictive check. Upon checking the assumptions, the obtained parameter estimates are reported along with the (highest density) credible intervals. Additionally, in order to ensure that the current model describes the data the best, the model is compared with several other models based on WAIC (Watanabe Akaike information criterion), which is considered to be the most Bayesian extension of AIC (Gelman et al., 2014). The secondary research question is investigated by evaluating the hypotheses concerning the direction and absolute value of regression coefficients using the Bayes Factors. Lastly, average causal effect (ACE) of maternal smoking is estimated using the inverse probability weighting (IPW) method based on Bayesian propensity scores.

3 Results

3.1 Convergence assessment

Figure 1 shows the trace plots (left), autocorrelation plots (center), and density plots (right). Trace plots for all parameters show a stable 'fat hairy caterpillar' shape, meaning that they traverse the whole posterior space freely, while exploring the area where the density is low as well. Two chains overlap nicely, that is, the location and width of both chains match (i.e., relatively constant mean and variance). This implies that they are not stuck at a local maximum and it is likely that they have reached a stationary posterior distribution. The posterior density plots of all parameters also look nice and smooth without any bumps or wiggly lines. The autocorrelation plots look fine as well. b_1 has a relatively higher autocorrelation compared to the others, indicating that it mixes rather slower. Yet, it reaches zero at about $lag13$ and this is thus considered to be not problematic given that our sample is large enough (n = 100,000). Gelman-Rubin diagnostics for all parameters are equal to 1 when rounded up to three decimal places, which implies that there is hardly any between-chain variance and the samples can be considered to have arisen from the same stationary distribution. In addition, Monte-Carlo errors (MC error) for all parameters are zero when rounded up to three decimal places and sufficiently smaller than the SDs (see Table 2). This indicates that the standard error of MC sampler is very small and there is almost no uncertainty originated from the iterative sampling procedure. Overall, it is concluded that both chains for each parameter are likely to to have reached convergence based on all the diagnostic results discussed. Note that however it is never guaranteed that the convergence is reached and in fact, it is not possible to prove convergence. As in this case, we can only say that there is no evidence against the convergence, and therefore they seem to have reached the target (posterior) distribution.

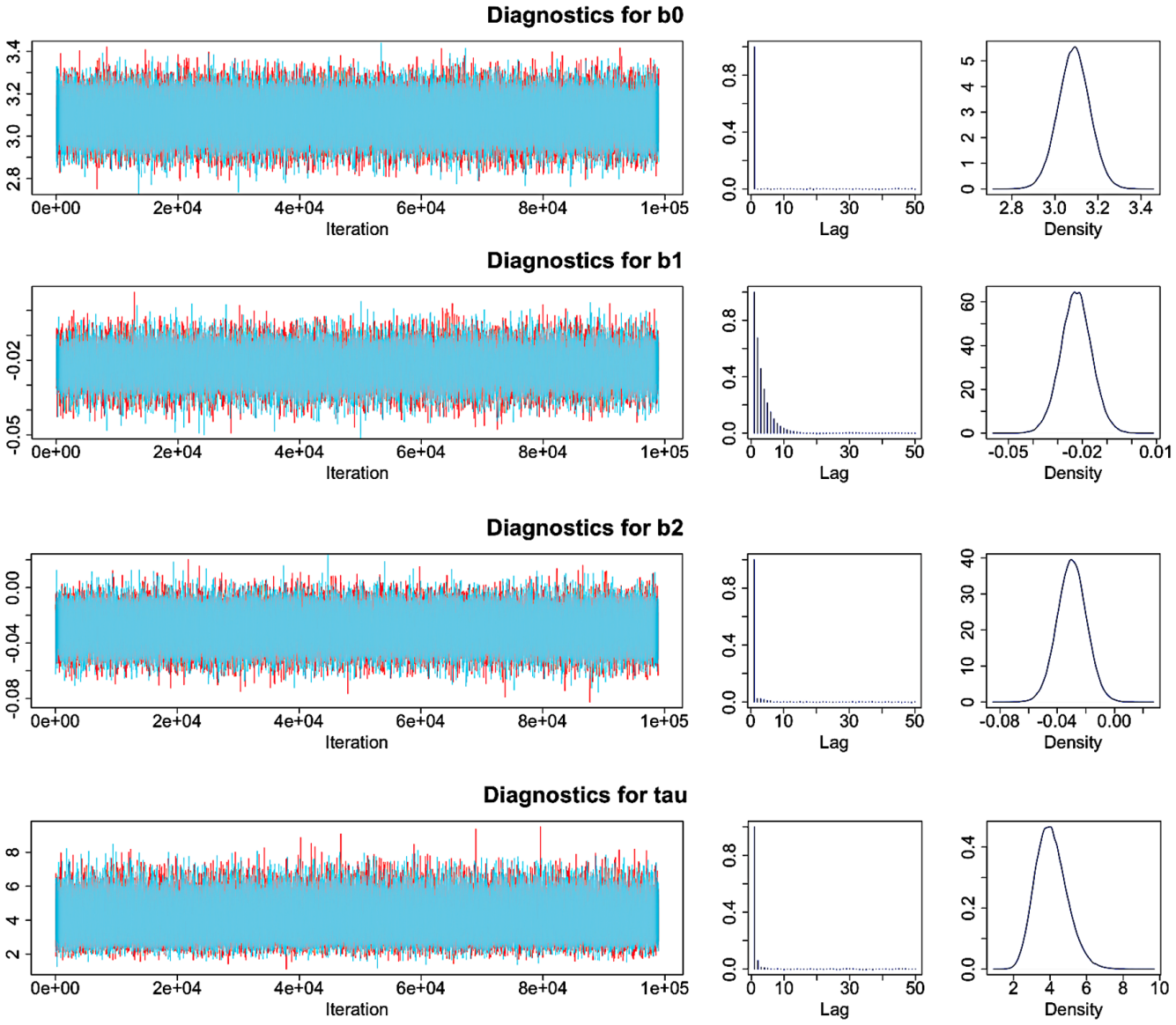

Figure 1: Trace plots, autocorrelation plots, and density plots of the generated samples for each parameter

3.2 Model assumption check: homoscedasticity & normality of residuals

Two of the central assumptions of the linear regression model are assessed using posterior predictive check: homoscedasticity and normality of residuals. To test homoscedasticity, 10000 sets of residuals based on the simulated data and observed data are ordered by the simulated y and observed y values respectively. The ordered residuals are then divided by two: first half (e_{h1}) and second half (e_{h2}). The discrepancy measure used for checking homoscedasticity is the absolute difference between the variance of the first half of ordered residuals and the second half of ordered residuals ($D = |Var(e_{h1}) - Var(e_{h2})|$). It is expected that the absolute difference in variance of two sets of residuals would be small if homoscedasticity is satisfied (i.e., residuals are equally distributed across the fitted value $\rightarrow$ variance of e_{h1} and e_{h2} would be similar), and the absolute difference would be large if homoscedasticity is violated (e.g., residuals are not equally distributed across the fitted value $\rightarrow$ variance of e_{h1} and e_{h2} would differ $\rightarrow$ the more they differ, the larger D becomes). The proportion that D (absolute difference in variance) computed from the simulated data is larger than D computed from the observed data over 10,000 iterations is the posterior predictive p-value (Bayesian p-value), which turns out to be 0.504, in this case.

$$\mathbf{p_{Bayesian}} = P(D^t_{sim} > D^t_{obs} \mid [Y, X], H_0) = \mathbf{0.504}, \ \textit{where } t = 1, ..., 10000$$

Since Bayesian p-value has an uni-modal distribution concentrated around 0.50, p-value that is close to 0.50 is regarded as an evidence for the null hypothesis. That is, if the data violates the model assumption, then values of the discrepancy measure for the observed data should be substantially larger than most of the discrepancy measure values for the simulated data, which would lead to a p-value close to 0. Therefore, when the value of the observed discrepancy measure is more or less the same as the value of the simulated discrepancy measure, namely p-value is close to 0.5, like in this case, the null hypothesis is not rejected (i.e., homoscedasticity assumption is satisfied).

The assumption of normality of residual is tested again comparing the simulated and observed residuals. The discrepancy measure used to test normality is the absolute difference between 0.95 and the area under the probability density function (PDF) curve within 1.96 standard deviation from its mean: $AUC = |0.95 - P(\mu_e \pm 1.96\sigma_e)|$, where μ_e = mean of residuals, and σ_e = SD of residuals. As 95% of the area under a normal curve lies within 1.96 SD from the mean, it is expected that $P(\mu_e \pm 1.96\sigma_e)$ would be close to 0.95, if the residuals indeed follow a normal distribution. As the residuals deviate more from normality, the AUC (absolute difference in the area under curve) is expected to be larger. The proportion that the AUC value computed from the data simulated under normality is larger than AUC values computed from the observed data over 10,000 iterations turns out to be 0.470.

$$\mathbf{p_{Bayesian}} = P(AUC^t_{sim} > AUC^t_{obs} \mid [Y, X], H_0) = \mathbf{0.470}, where\ t = 1, ..., 10000$$

Since the Bayesian p-value is close to 0.5 in this case, meaning that the deviation from 0.95 is almost equally likely in the simulated and observed residuals, it is concluded that the normality assumption is met.

3.3 Parameter estimates

As seen in Table 2, mean and median values are almost identical when rounded up to three decimal places for all parameters. Similarly, the 95% credible intervals are almost equivalent to the 95% highest density intervals (i.e., the shortest interval containing 95% probability density) for all parameters, except for the residual SD (yet, only minor difference is observed). It makes sense that they are almost all identical, because the posteriors of all parameters seem to follow a symmetrical distribution as seen in Figure 1. 95% credible intervals indicate that there is 95% probability that the true parameter estimate lies within the boundary values. Given that the credible intervals do not include zero for both coefficient estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$), the number of cigarettes smoked by mothers as well as mother's age are considered as reasonable (significant) predictors. $\hat{\beta}_1$ is -0.023, meaning that every extra cigarette smoked by mothers is expected to decrease the birth weight of babies on average by 0.023 kg, when mother's age remains constant. $\hat{\beta}_2$ is -0.03, meaning that for every additional year in mother's age, the baby's birth weight is expected to decrease by 0.03 kg, when the number of cigarettes smoked by mothers remains constant.

Table 2: Parameter estimates and corresponding credible intervals

Parameter	Mean	Median	SD	95% CI [2.5%, 97.5%]	95% HDI [2.5%, 97.5%]	MC error
Intercept (β_0)	3.088	3.089	0.074	[2.943, 3.234]	[2.943, 3.235]	0.000
Number of smokes (β_1)	-0.023	-0.023	0.006	[-0.035, -0.010]	[-0.035, -0.010]	0.000
Mother's age (β_2)	-0.030	-0.030	0.010	[-0.050, -0.010]	[-0.050, -0.010]	0.000
Residual SD (σ)	0.504	0.499	0.055	[0.410, 0.626]	[0.402, 0.614]	0.000

Note. CI: Credible Interval, HDI: Highest Density Interval, SD: Standard Deviation, MC error is the same thing as Naive SE.

3.4 Model comparison: WAIC (Watanabe–Akaike information criterion)

To explore any possible improvement of our model, several other models are compared (including our current model) by the predictive accuracy and model complexity by means of WAIC. WAIC is often preferred to DIC, since not only that DIC assumes the posterior is approximately normal, but also it is seen as not fully Bayesian, given that it is based on the point estimate (posterior mean). WAIC, on the other hand, is deemed to be fully Bayesian because it uses the entire posterior distribution, which makes it more accurate than DIC and it does not assume the normality on posterior distributions (Gelman et al., 2014). WAIC is defined as follows:

$$WAIC = -2lppd + 2p_{WAIC}$$

,where $lppd$(log-pointwise predictive density)= $\sum_{i=1}^{N} log(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta_s))$ is the sum of log-averaged likelihoods across N observations over the S posterior samples, and p_{WAIC}(effective parameters)= $\sum_{i=1}^{N} Var(log\, p(y_i|\theta_1), ..., log\, p(y_i|\theta_s))$ is the sum of variance of loglikelihood across S posterior samples over N data points. The interpretation is the same as DIC; the lower the WAIC, the better the model is in terms of the balance between goodness of fit against the number of parameters used in the model. The following four models are compared as shown in Table 3, and our current model, $Model4$ (with the number of cigarettes smoked by mothers and mother's age as predictors) is turned out to be the best model among the models considered here, given that it has the lowest WAIC value.

Table 3: Model comparisons based on WAIC

	Model1: $y = b_0$	Model2: $y = b_0 + b_1 \cdot x_1$	Model3: $y = b_0 + b_2 \cdot x_2$	Model4: $y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$
WAIC	90.71	78.83	82.96	72.44

Note. y: Baby's birth weight, x_1: Mother's smoking (number of cigarettes smoked), x_2: Mother's age

3.5 Hypotheses evaluation: Bayes Factor

Here, the following informative hypotheses (H_1, H_2), which were formulated prior to conducting the analysis are evaluated using Bayes Factor to investigate the secondary research question.

$$H_1 : \beta_{smoke} < 0,\ \beta_{age} < 0 \qquad H_2 : |\beta_{smoke}| > |\beta_{age}| \qquad H_u : \beta_{smoke},\ \beta_{age}$$

H_1 specifies that the effect of mother's smoking and mother's age on baby's birth weight are both negative, and H_2 specifies that the influence of mother's smoking is larger than that of mother's age. H_u is the unconstrained hypothesis (i.e., no constraints placed on parameters), which is included in order to prevent from selecting the best among a set of bad hypotheses (hence called the *fail-safe hypothesis*). Note that the data are standardized so that the hypotheses are evaluated in terms of the standardized coefficients whose values are directly comparable.

Table 4: Bayesian informative hypothesis testing results

	Fit	Complexity	BF.u	BF.c	Posterior Model Probability (PMP)
H_1	0.998	0.224	4.457	1970.544	0.655
H_2	0.667	0.494	1.350	2.052	0.198
H_u					0.147

Note. BF.u denotes the Bayes factor of the hypothesis versus the unconstrained hypothesis H_u. BF.c denotes the Bayes factor of the hypothesis versus its complement.

As shown in Table 4, the Bayes factor of H_1 comparing to H_u is 4.457, indicating that H_1 receives 4.5 more support from the data than H_u. The Bayes factor of H_2 comparing to H_u is equal to 1.350, implying that H_2 is almost as equally supported by the data as H_u (i.e., only 1.35 times more). Bayes factors comparing to their respective complement (BF.c) also indicate that there is substantially more support for H_1. The posterior model probability (PMP) helps determining which of the two hypotheses is the best. Since H_1 has the highest PMP of 0.655, H_1 is considered to be the best in this case. Yet, choosing H_1 comes with the Bayesian error probability of $0.198 + 0.147 = 0.345$, implying that the probability of choosing H_1 being incorrect is about 35%. That is, H_1 receives the most support out of all hypotheses under consideration, but H_2 and H_u cannot be excluded. Also, there is clearly a possibility that another hypothesis could be an even better candidate than H_1.

3.6 Bayesian causal inference

Trying to truly talk about the *effect* of maternal factors, I attempted to estimate the average causal effect (ACE) of mother's smoking on baby's birth weight using inverse propensity weight (IPW) based on Bayesian propensity scores. Currently, an active discussion is going on how to compute Bayesian propensity scores legitimately, and here I followed the suggestion by Liao and Zigler (2020). Simplifying the suggested procedure is: 1) Estimate the

likelihood of cause using Bayesian model to generate the posterior of propensity scores. 2) Compute the posterior predicted propensity scores for each of N draws from the posterior. 3) Compute IPW using the predicted propensity scores, and run the outcome model N times to obtain ACE estimates. Note that I used extra other variables from the original data set (i.e., father's smoking, father's education) besides mother's age as covariates when computing the propensity scores, and the dichotomized version of mother's smoking (0: number of cigarettes smoked is zero, 1: otherwise) is used as the cause variable[2]. The estimated ACE is -0.754 (95% $CI = -0.965, -0.612$) and it is significant as the credible interval does not include zero. It could be interpreted as the weight of babies from the mothers who smoke is expected to be 0.76 kg lower than the weight of babies from the mothers who do not smoke.

4 Conclusion

Overall, the analysis shows that the baby's birth weight is negatively associated with mother's age and the number of cigarettes smoked by mothers per day. The negative coefficient estimates along with the credible intervals (of both coefficients excluding zero), WAIC (lowest in the model including both IVs), and Bayes factor (of H_1: $\beta_{smoke} < 0, \beta_{age} < 0$ being supported the most) together indicate that both mother's age and mother's smoking have a negative effect on baby's birth weight. Given that the Bayes factor of $H_2 : (|\beta_{smoke}| > |\beta_{age}|)$ is rather small (BF = 1.3), it is concluded that mother's smoking has only a slightly more influence on the baby's birth weight compared to mother's age. Additionally, it is shown that the average causal effect (ACE) of mother's smoking on birth weight is negative and significant, which is in agreement with the rest of the results.

Running a frequentist regression analysis would actually provide more or less the same estimates in this case, as I used the uninformative priors. Even though the value of estimates themselves would not differ much, using Bayesian approach provides more extensive information. In frequentist approach, I would obtain classical p-value and reject the H_0 ($\beta = 0$) concluding that there exist some relationships. With Bayesian, I was able to construct the specific hypotheses that I was interested in and Bayes factors showed, to what degree that each hypothesis was actually supported by the data as well as how likely each hypothesis was, compared to the other. In addition, having the whole posterior distribution rather than a single point estimate gave me a lot of flexibility to compute many cool statistics/measures here, such as median (would've not been this easy to obtain in the frequentist approach), highest density interval, WAIC, and my very own test statistics to check model assumptions. This also applies to the causal inference; using frequentist approach, I would have only obtained a single ACE estimate, whereas applying Bayesian I was able to get the whole distribution of ACE, which enables me to compute numerous statistics relatively easily. Lastly, here I had a small data (n=47), which naturally comes with more uncertainty when it comes to generalizing the results. Hence, it is logical to have another study to follow up, and it can be easily implemented in Bayesian by specifying a prior based on the results I obtained in this current analysis. I think that Bayesian way of aggregating the evidence is more straightforward than performing a classical meta analysis where you try to compare multiple studies indirectly. However, I also want to point out that this incredibly high flexibility allowed in Bayesian is a double-edged sword. It lets a lot of my personal opinions into the analysis, which makes it more subjective. It was my statistics to test homoscedasticity that I thought as an appropriate measure, and it was my opinion that the obtained p-value seemed to be close enough to 0.5. Every step of analysis in a way was infused with my idea/opinion that might not be agreed by others. Thus, applying Bayesian allows more flexibility but it also requires much of cautious thinking and reasoning in every step, which makes it challenging. But I believe that this challenge is what makes Bayesian more appealing: you really need to think in order to to perform Bayesian analysis properly.

5 References

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6), 997–1016.

Liao, S. X., & Zigler, C. M. (2020). Uncertainty in the design stage of two-stage bayesian propensity score analysis. *Statistics in Medicine*, 39(17), 2265–2290.

Stanford Children's Health (2022). *Low birth weight*. Retrieved from https://www.stanfordchildrens.org/en/topic/default?id=low-birth-weight-90-P02382

[2] I chose to show the maternal smoking variable here, because not only that it is easier to dichotomize (in *smokers* and *non-smokers*), but also there aren't any proper variables available in the dataset to predict mother's age. For mother's smoking, some of the variables could be used as covariates, although it is not sufficient/ideal in my opinion. So please consider this as more of illustration purpose to show how the ACE could be estimated rather than focusing too much on the substantive interpretation of the results.