Posterior Prediction of Mask Mandate Importance under Bayesian GLM Framework

Oliver Hill and Roman Kouznetsov

Introduction

The COVID-19 pandemic has raised many questions about how effective policies have been in limiting the spread of the virus. Luckily, there is a plethora of data available from reliable sources that can be used to investigate these effects. Mask mandates are a very political and publicized intervention that has been used throughout the pandemic. These policies force citizens to wear cloth masks covering their mouth and nose in public in an attempt to block aerosol passage between citizens.

The politicization of masks has led states to implement mandates along political party lines, with more Republican states not implementing mask mandates, not enforcing them, or having more lenient policies than their Democratic counterparts. There are costs and benefits to mask mandates and we are interested in determining if the benefits are truly benefits. Specifically, did early adopters of masks, states that had mask mandates in the early months of the pandemic (March - July 2020), have lower death rates than late adopters? Is a mask mandate a significant predictor of the death rate for a state? Here we are not concerned about the specific implementations of these mandates, only that they existed in a state during a given month.

Starting from the belief that earlier and longer mask mandates will decrease the number of total deaths from COVID-19 in a state, we use Bayesian analysis to measure the effect of the policy. This is enabled by a very thorough dataset published by the U.S. Centers for Disease Control and Prevention (CDC). The data includes COVID-19 death counts for every state for every month from January 2020 to present day (April 2021 at the time of writing).

Data

Published weekly by the CDC, the COVID-19 deaths by state is dataset is pleasantly clean and complete. The data is sourced from the National Center for Health Statistics (NCHS), which collects and collates death records from each state through the Vital Statistics Cooperative Program. We are able to embark on this work because the CDC is committed to not merely taking a random sample from the population, but to attempting to understand the complete distribution of the population. This enables some interesting exploratory data analysis involving the distribution of the deaths.

One of the main arguments for the importance of large-scale quarantines was the impact of COVID-19 on the older members of society. In figure 1, we see that indeed older ages tend to experience more deaths than younger ages. Ideally we would have controlled for the populations of these age groups, but we did not have data on that granularity for this project.

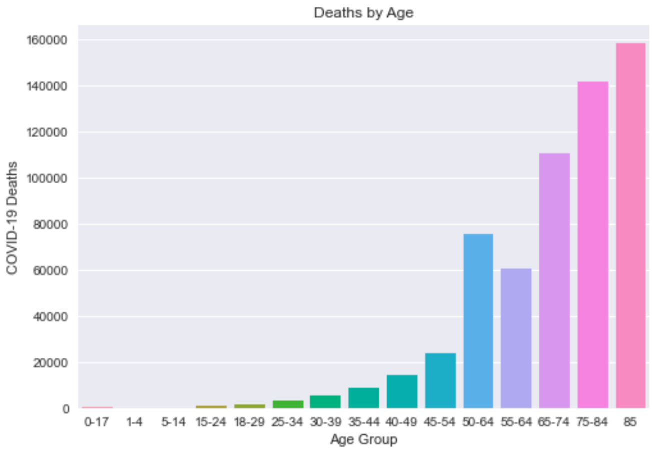


FIGURE 1. *COVID-19 deaths by age across the country. We can see there is a clear trend showing that older populations are more vulnerable to death by the virus.*

Figure 2 is the gender identity breakdown of the death rates in our dataset. We saw that more men died than women in our dataset, which follows observations by science and media.

Population Interpolation

The goal of the model presented in this paper is to identify the impact that mask mandates had on reducing monthly death rates. Doing so requires accurate population estimates by month so that the rate of death is specific to the month in question. The U.S. Census Bureau releases estimates of populations by state annually, providing 11 data points since the 2010 Census of populations by state every July. Several interpolation methods were used to accurately estimate the population by month. The interpolation method influences the value of the population estimates - especially at the ends due to Runge's phenomenon. For this reason, linear, cubic, natural cubic, polynomial, and fakenodes [De Marchi et al. 2020] interpolations were conducted. Figure 3 showcases the curves fit for interpolation for all the aforementioned methods.

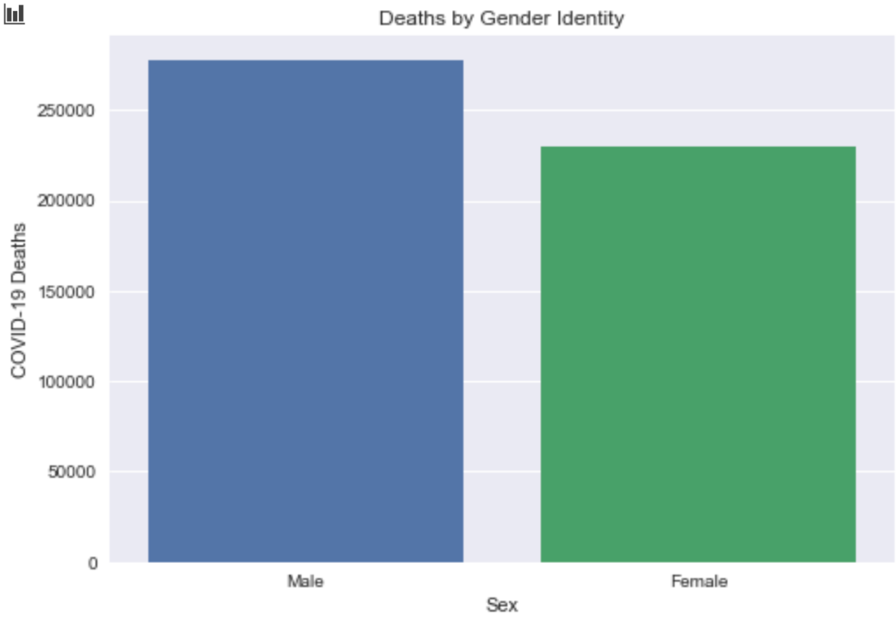
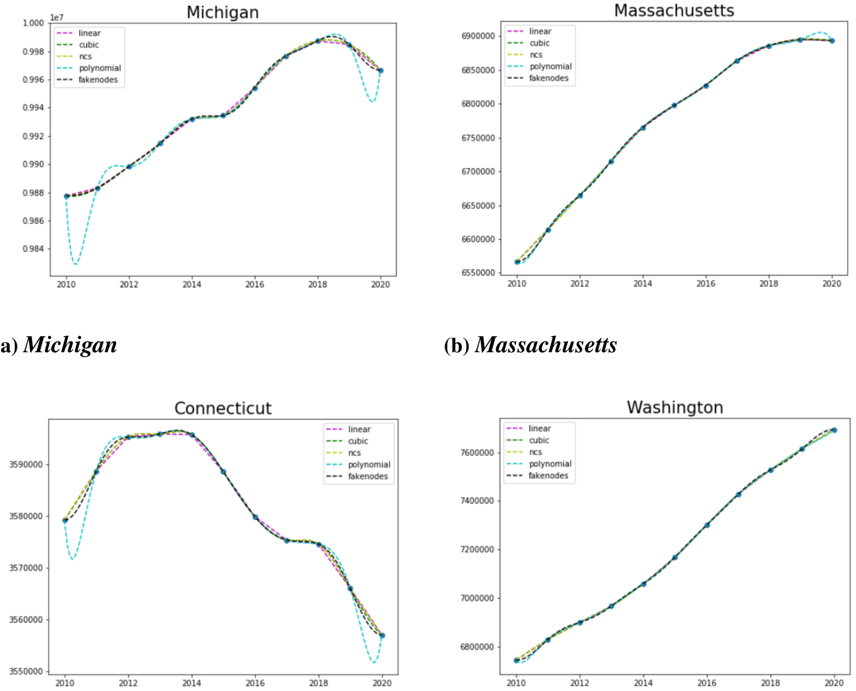


FIGURE 2. *COVID-19 deaths by gender identity across the country. There is clear indication that more men die than women to COVID-19.*



(a) ***Michigan***

(b) ***Massachusetts***

(c) ***Connecticut***

(d) ***Washington***

FIGURE 3. *Population interpolation graphs for selected states from July 2010 to July 2020. Many of the interpolation methods yield similar results barring Runge's phenomenon. Table 3 showcases how each interpolation method impacts the posterior coverage.*

The remaining figures throughout this report will assume the population estimates are generated via natural cubic splining, but the final results include the posterior coverage for all interpolation methods.

Model

The purpose of this model is to assess the impact that mask mandates at all previous months have on the current death rate in a Bayesian setting. Our dataset allows us to control for the demographic features of age group and sex. This means that the covariates used in this model are age (a set of binary variables), sex (binary), and mask mandates for all months from March to July (a set of binary variables).

The choice to make the response variable the death rate is imperative. Modelling death counts would lead to the unhelpful conclusion that states with higher populations (regardless of mask mandate implementation) would have higher COVID-19 death counts. Therefore, the response variable should be modelling death counts, but should be offset by the state's population, allowing for all state death count data to be compared on the same scale. This was the reason for the emphasis on the population data interpolation described in the Data section.

Moreover, a vanilla count model would not be appropriate given the observed data. Figure 4 showcases that the observed death rates are highly concentrated at 0, suggesting that a zero-inflated likelihood is better tailored to the observed data.

All these observations lead to a zero-inflated Poisson likelihood as a statistically sound choice for our model. As for the prior information we have at our disposal, Figure 1 shows that older age groups are more likely to die of COVID-19. We decided to assume that sex did not play a role in determining the death rate. This does not represent what we saw in figure 2, but we wanted to see if the posterior would represent this difference or not, as it is not clear why COVID-19 would affect men more than women. The posterior does end up representing this difference. As for the mandates, we impose a spike and slab prior over all monthly mandate regression coefficients. This is to ensure that posterior distributions that do not include zero imply that the likelihood is highly concentrated away from zero. For this model, the sex and intercept coefficients were given normal priors, the age coefficient was given a multivariate normal prior with a diagonal covariance matrix, and the mandate coefficients were given a spike-and-slabe prior with a "spike" mass at 0 and a "slab" multivariate normal distribution with a diagonal covariance matrix.

All of this put together yields the model outline below. Figure 5 showcases the graphical representation of this model.

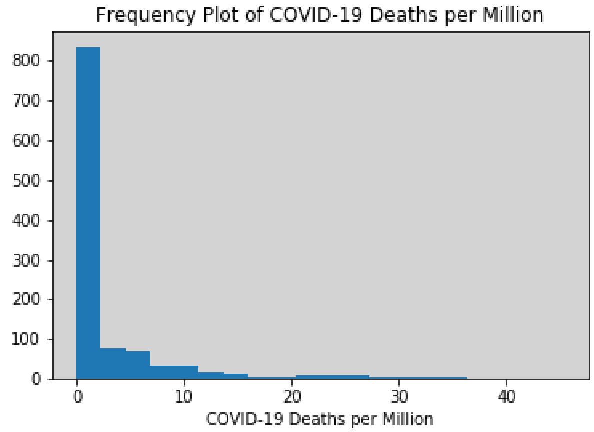


FIGURE 4. *This frequency plot of COVID-19 deaths per million demonstrates the necessity of a zero-inflated count model. Roughly 74.3% of the observations are classified as 0, suggesting that a vanilla count model would wildly underestimate the posterior mass at 0.*

$$\mu_0 \sim \mathcal{N}(0,1)$$

$$\sigma_0 \sim |\mathcal{N}(0,4)|$$

$$\beta_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$\tau \sim InvGamma(0.05, 0.05)$$

$$\Sigma = I$$

$$\Xi \sim Bernoulli(p = 0.9)$$

$$\beta_{mandate} \sim \Xi * \mathcal{N}_4(\mathbf{0}, \tau\Sigma)$$

$$\mu_{age,i} = [-5 + i * \frac{10}{9}] \text{ for } i = 0, ..., 9$$

$$\sigma_{age} \sim |\mathcal{N}(0,4)|$$

$$\beta_{age} \sim \mathcal{N}_{10}(\mu_{age}, \sigma_{age}^2 \mathbf{I})$$

$$\mu_{sex} \sim \mathcal{N}(0,1)$$

$$\sigma_{sex} \sim |\mathcal{N}(0,4)|$$

$$\beta_{sex} \sim \mathcal{N}(\mu_{sex}, \sigma_{sex}^2)$$

$$\frac{y_i}{pop_i} \sim ZIP(\psi = 0.7, \theta = \beta_0 + \text{age} * \beta_{age} + \text{sex} * beta_{sex} + \text{mandates} * \beta_{mandates})$$

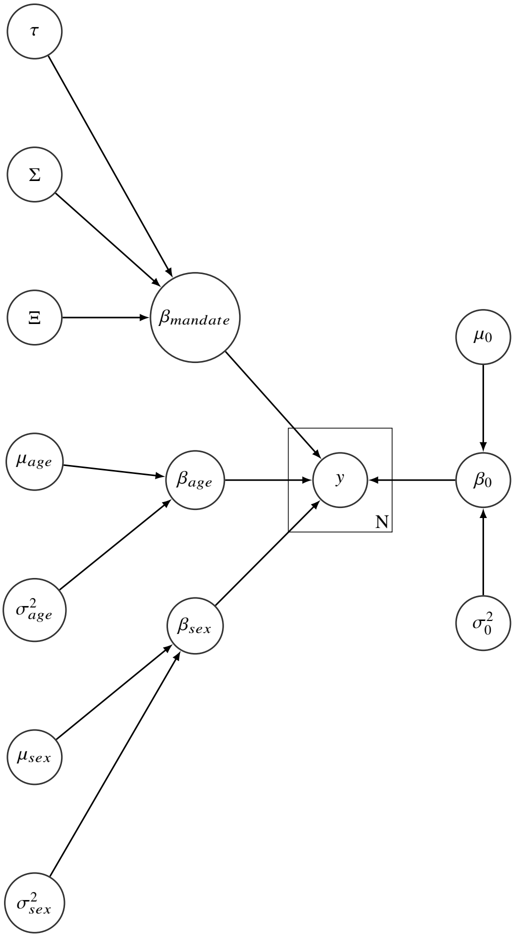


FIGURE 5. *Graphical Model of COVID-19 Offset Poisson Bayesian GLM. The structure of the model will follow an offset-Poisson GLM, where the main posterior inference of interest will be on the coefficients of the mandate dummy variables.*

Cost of Null Hypothesis

It is worthy to point out that the spike-and-slab prior imposed on the mandate coefficients in the GLM represents a high-cost null hypothesis. This is purely for the sake of argument and not a reflection of what we believe should be empirically used as the null hypothesis for all mask mandate decision making.

Tables 1 and 2 showcase the consequences of making errors when the null is that masks do not ($\beta_{\text{mandate}} = \mathbf{0}$) and do help ($\beta_{\text{mandate}} < \mathbf{0}$) with lowering COVID-19 death rates, respectively.

H_0: **Mandates don't work.**	Null Accepted	Alternative Accepted
Null True	Mandates don't work. Mandates not implemented.	**Type I Error** Mandate don't work. Cost: Wear masks when it's not needed.
Null False	**Type II Error** Mandates work but we conclude they don't. Cost: Major loss of life.	Mandates help and we act on that conclusion by implementing masks.

TABLE 1. *An interpretation and cost analysis of decisions made under the null hypothesis that mandates do NOT lower the death rate.*

H_0: **Mandates work.**	Null Accepted	Alternative Accepted
Null True	Mandates work. Mandates implemented.	**Type I Error** Mandates work, but we conclude that they don't. Cost: Major loss of life.
Null False	**Type II Error** Mandates don't work but we conclude they do. Cost: Wear masks when we don't have to.	Mandates don't work. Mandates not implemented

TABLE 2. *An interpretation and cost analysis of decisions made under the null hypothesis that mandates do lower the death rate.*

This suggests that the null we use is costly because it would require sufficient evidence to enforce life-saving mask mandates. A more cost-efficient null would be the one represented in Table 2 because it would take significant evidence to suggest that makes do not save lives, erring on the side of caution rather than comfort.

This does not directly impact the results of our model, but we believe it is important to address given the severity of the subject.

Results

In this project we rejected the null hypothesis - that mask mandates had no effect on COVID-19 death rates - for all months (April, May. June, July 2020). This indicates that states that implemented statewide mask mandates were effective in lowering the COVID-19 death rate. Moreover, this result is compounded the longer the mandate was in place.

This result parallels intuition, as well as professional scientific advice, that mask mandates are effective in reducing the spread of the virus. While we were not able to investigate whether the spread of the virus was limited, death rate seems to be a reasonable proxy. We have visualized the posterior distribution of $\beta_{mandate}$ in Figure 6, and the remaining posterior distributions in the Appendix. These posterior distributions heavily confirm our suspicion that mask mandate are helpful in lowering COVID-19 death rates.

Table 3 displays the 95% posterior coverage (HDI) for our models parameters. We can see that the choice of interpolation algorithm plays a minimal role in the intervals. We also see that none of the intervals include zero, which helps us conclude that mask mandates played a significant role in lowering the death rates. This result provides ample evidence to suggest that states should have implemented statewide mask mandates near the start of the pandemic to curb the COVID-19 death rate as much as feasibly possible.

Interpolation Choice	$\beta_{\text{mandate[April]}}$	$\beta_{\text{mandate[May]}}$	$\beta_{\text{mandate[June]}}$	$\beta_{\text{mandate[July]}}$
Linear	(-0.609, -0.399)	(-0.820, -0.686)	(-0.364, -0.297)	(-0.172, -0.122)
Cubic	(-0.607, -0.402)	(-0.823, -0.693)	(-0.365, -0.297)	(-0.173, -0.123)
Natural Cubic Splining	(-0.614, -0.402)	(-0.823, -0.690)	(-0.365, -0.299)	(-0.173, -0.122)
Polynomial	(-0.605, -0.404)	(-0.823, -0.694)	(-0.362, -0.295)	(-0.175, -0.123)
FakeNodes	(-0.606, -0.401)	(-0.820, -0.693)	(-0.363, -0.299)	(-0.173, -0.121)

TABLE 3. *Posterior 95% highest density interval coverage of the posterior distribution for mandate regression coefficients.*

From May to July, we see a trend towards the mask mandates being less significant. This could be due to several real-world phenomena. One possible explanation is that Summer provided a time for people to be outside and the virus, being airborne, is less easily spread outside. Another is that it takes 2-3 months for the state mandate to fully realize its potency. It is likely the case that in the future (July 2021), if we were to consider all of the months of the pandemic, the spike and slab prior would more properly perform its role in feature selection by zeroing out the months for which the effect of the mandate has weened away.

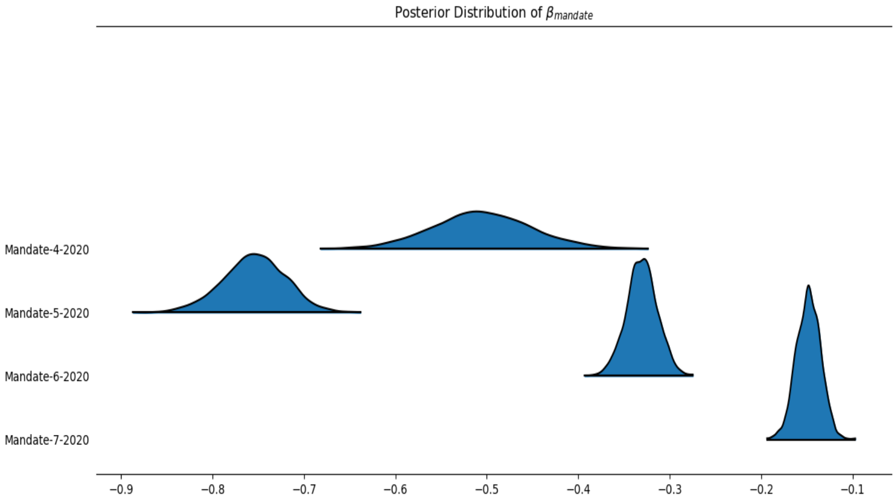


FIGURE 6. *Posterior Distribution of Mandate Coefficients*

Limitations

There are several notable limitations with our analysis. Firstly, and perhaps most importantly, it is not clear what the causal relationship is between mask mandates and death rates. While we hypothesized that mask mandates would cause lower death rates, it may be the case that higher death rates were the reason that many states implemented mask mandates. Figure 7 displays COVID-19 death rates (the blue line) against when mask mandates came into effect (the green line) to help us think about which of these hypotheses make more sense. When the green line is low (at zero), there is no mask mandate in the state. The green line jumps to one the month that the state implemented a mask mandate.

Some states implemented mask mandates during the same month that the state began experiencing high COVID-19 death rates. However, 21 of the 30 states that implemented mask mandates during this time frame implemented them after seeing drastic growth in their death rates. This dataset also does not track infection rates, and there is a two to four week lag between higher infection rates and higher death rates. Mask mandates might be implemented after higher infection rates but before higher death rates occur, which creates another limitation in our dataset, and, therefore, our analysis.

Another limitation in our analysis is that mask mandates are implemented at many different levels of government - sometimes states implement them, but city and county governments also implement mask mandates which confuses the data and our reasoning. If the county that contains Seattle implements a mask mandate a month before the state of Washington, the death rates at the state level will likely react to Seattle's mandate

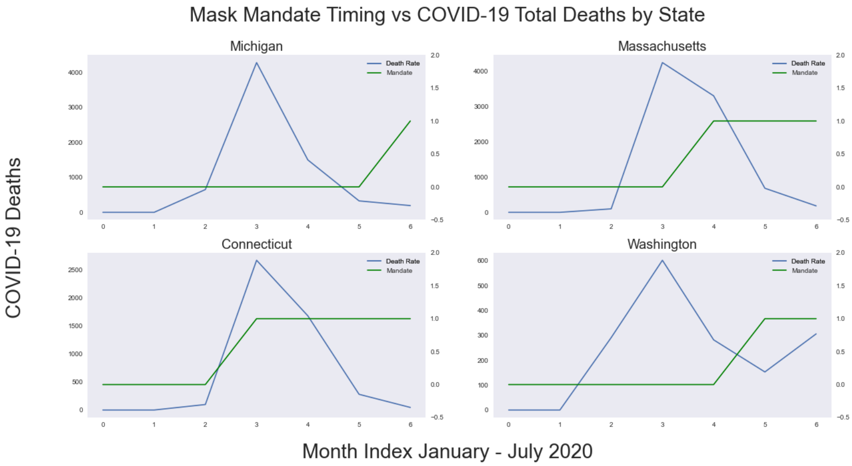


FIGURE 7. *COVID-19 death rates in four states against when mask mandates were implemented in each state.*

as well as Washington's. This is difficult to model but important in understanding the relationship between mask mandates and death rates, so it is a limitation of our analysis.

A third limitation lies in the data we have for the population of each state. The data that we used was sourced from the U.S. Census which means it is quite reliable, but we don't have data after July 2020. This severely limits the scope of our analysis, as the pandemic was only four months old at that time. We considered extrapolating the data we had for population, but since extrapolation creates a large source of error we decided to limit our analysis to January - July 2020 for this project. While this time frame provided substantial data for our hypothesis, it created a major limitation in the realm of hypotheses we could have.

Additionally, we are treating mask mandates as a binary policy. However, mask mandates vary qualitatively across states and across time. Each policy is different and encourages citizen compliance in different ways. We don't capture any of this information which is a serious limitation in our analysis.

Future Work

There are many directions for future work stemming from this project. As mentioned previously in this report, we are very interested in the converse relationship - do higher death rates lead to states instituting mask mandates? This would be possible with the current data, but would take more time than we have for this project. Another potential direction for future work would be to explore the interactions between mandates and the geography of each state - how neighboring states affect death rates in each other and how that in turn affects mask mandates. Due to the availability and abundance of COVID-19 data, we believe that is currently possible to pursue both of the directions for future work mentioned above.

One major flaw in the model is the fact that the zero inflation factor is not incorporated in the Bayesian model. This was due to time constrains of model implementation, but the addition of a parameter ψ to the likelihood in our Bayesian model is fairly easy to implement and can be done as described in Neelon, O'Malley, and Normand 2010. The choice of the priors could be a bit too narrow and would shrink our posterior estimates towards the prior, but our posterior predictive check provided in the supplementary materials, suggest that the posterior is a scaled down version of the raw observed data counts, a behavior that we would expect for a posterior that has an offset. Lastly, due to the limited population data available, only dates prior to July 2020 were considered, but the virus' presence still exists as of April 2021, so this model is easily repeatable once an accurate population estimate comes out for July 2021. We could have considered some population extrapolation techniques, but that would result in noisier estimates, especially due to the fact that the population dynamics are likely to be different than all preceding year to year estimates. With more reliable population estimates in the future, the effects of mandates many months in advance can be investigated easily using the existing model.

Supplementary Material

Code, datasets, graphs, and instructions for recreation can be found at this Github repository. The code there allows for models to be customized for number of samples, binary status of mandate variable, zero-inflation proportion, population interpolation type, the spike parameter, and the posterior coverage HDI value.

Appendix

The figures below showcase the posterior distributions of the regression coefficients for the other covariates that were not the central focus of this report but could be of interest to the reader.

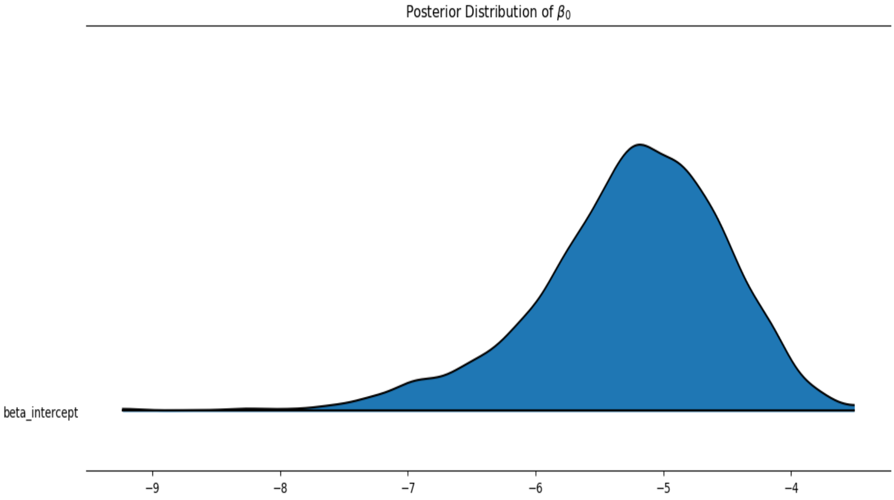


FIGURE A1. *Posterior Distribution of β_0*

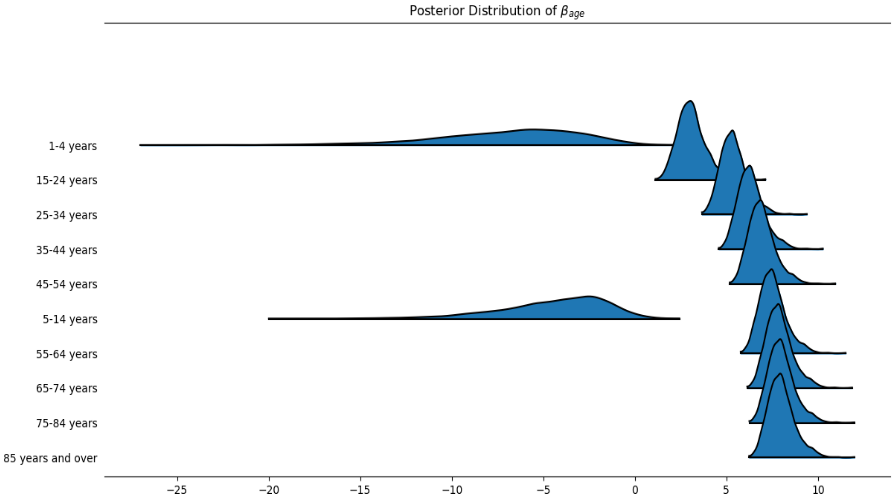


FIGURE A2. *Posterior Distribution of Age Coefficients*

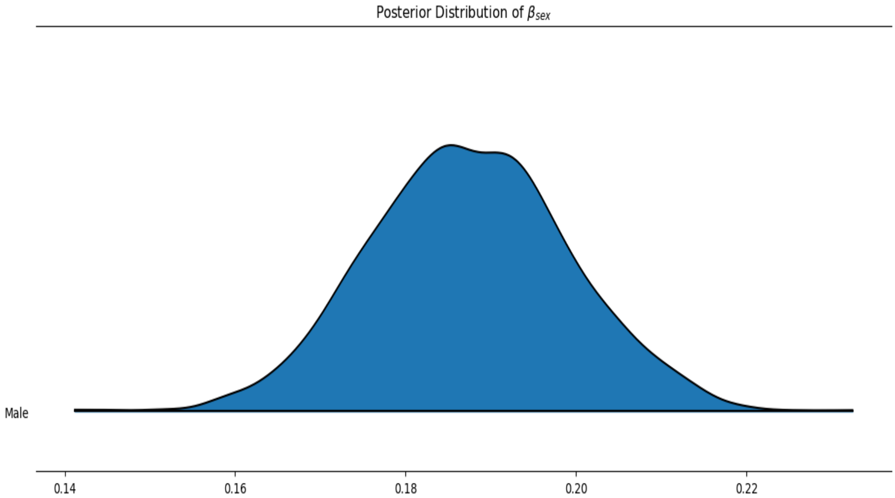


FIGURE A3. *Posterior Distribution of* β_{sex}

References

De Marchi, S., F. Marchetti, E. Perracchione, and D. Poggiali. 2020. Polynomial interpolation via mapped bases without resampling. *Journal of Computational and Applied Mathematics* 364:112347. ISSN: 0377-0427. https://doi.org/https://doi.org/10.1016/j.cam.2019.112347. Available at <https://www.sciencedirect.com/science/article/pii/S0377042719303449>.

Neelon, B.H., A.J. O'Malley, and S.L. Normand. 2010. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use [in en]. *Statistical modelling* 10 (4):421–439. https://doi.org/10.1177/1471082X0901000404. Available at <https://doi.org/10.1177/1471082X0901000404>.