STA2201 Assignment 3

Matthew Scicluna

2016-04-01

Short Answers

Non Parametrics

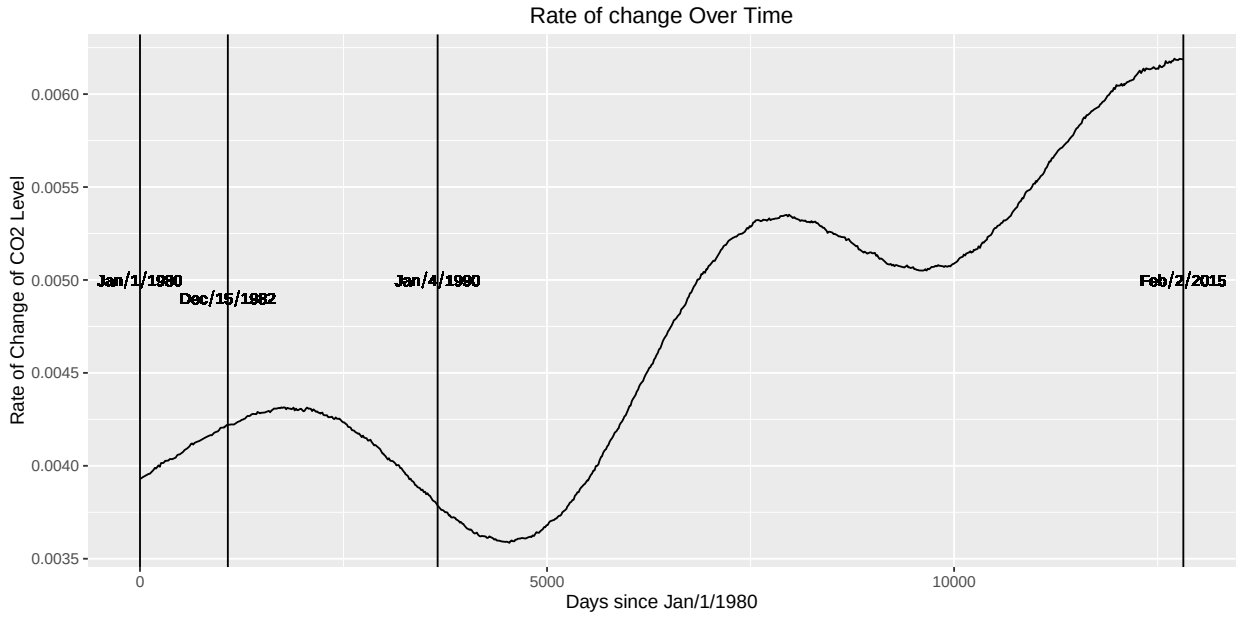

Figure 1: An approximation to the derivative of the GAM Model

We investigated the changes in Carbon Dioxide concentration collected from Antarctica.The data was analyzed using the `gam` function from the mgcv R package. We used a Generalized Additive Model from the Gamma family with a log link function to predict carbon concentration from days since Jan 1, 1980. The model is as follows:

$y_i \sim \Gamma(\theta)$, where $log(E(Y)) = \sum_{j=1}^{4} \phi_j(x_i)\beta_j + f(x_i)$

and $\phi_1(x_i) = cos(2\pi x_i),\ \phi_2(x_i) = sin(2\pi x_i),\ \phi_3(x_i) = cos(4\pi x_i),\ \phi_4(x_i) = sin(4\pi x_i)$, $f(x_i)$ is some B-spline fit to the data. We note that there may be several such B-splines.

Also y_i is concentration of carbon measured on the specific day which occured x_i years since Jan 1st 1980. We note that ϕ_1 and ϕ_2 represent yearly fluctuations and ϕ_3 and ϕ_4 represent biyearly fluctuations. We can see from figure 1 that this model is a good fit from the data.

We approximated the first derivative of the time trend using a finite difference approximation. The approximation is graphed in figure 1. Additionally we labelled the most recent day, the early 1980s and 1990. We noticed that the derivative is at its smallest at the beggining of the 1980s and 1990, which coincides with global economic recessions. Finally while the rate of change is still increasing and at a faster rate than ever, this rate is increasing at a decreasing rate. This may be because of another global recession or from international initiatives aimed at decreasing global carbon emmisions.

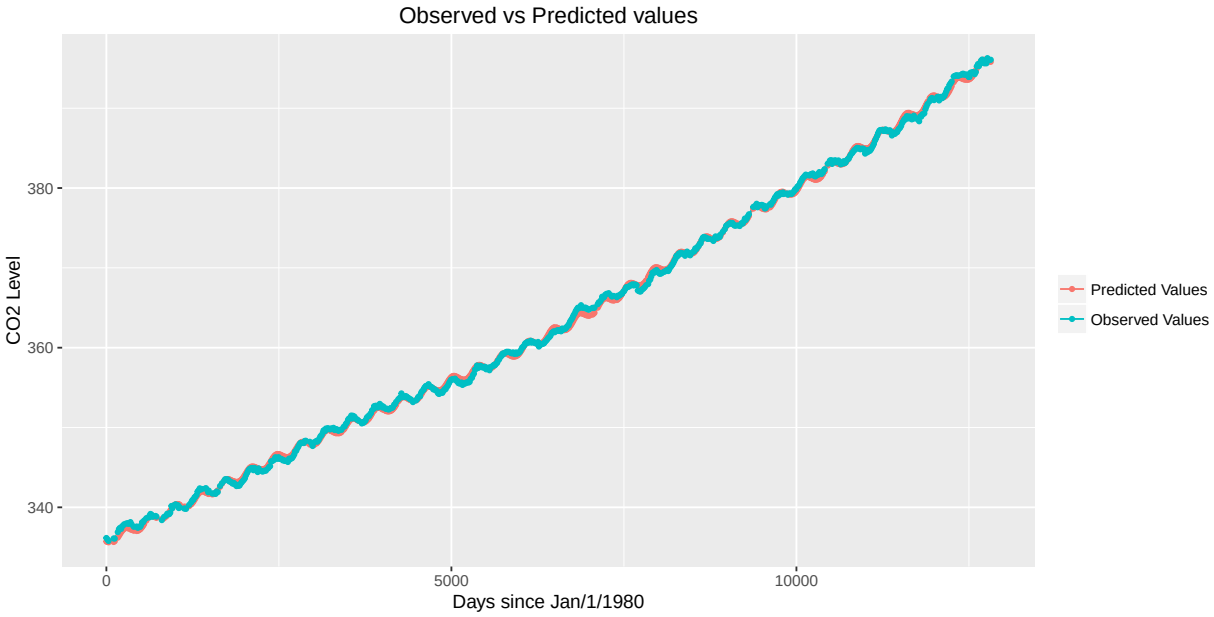

Figure 2: Assessing the fit of the CO2 Emissions Data with our GAM Model

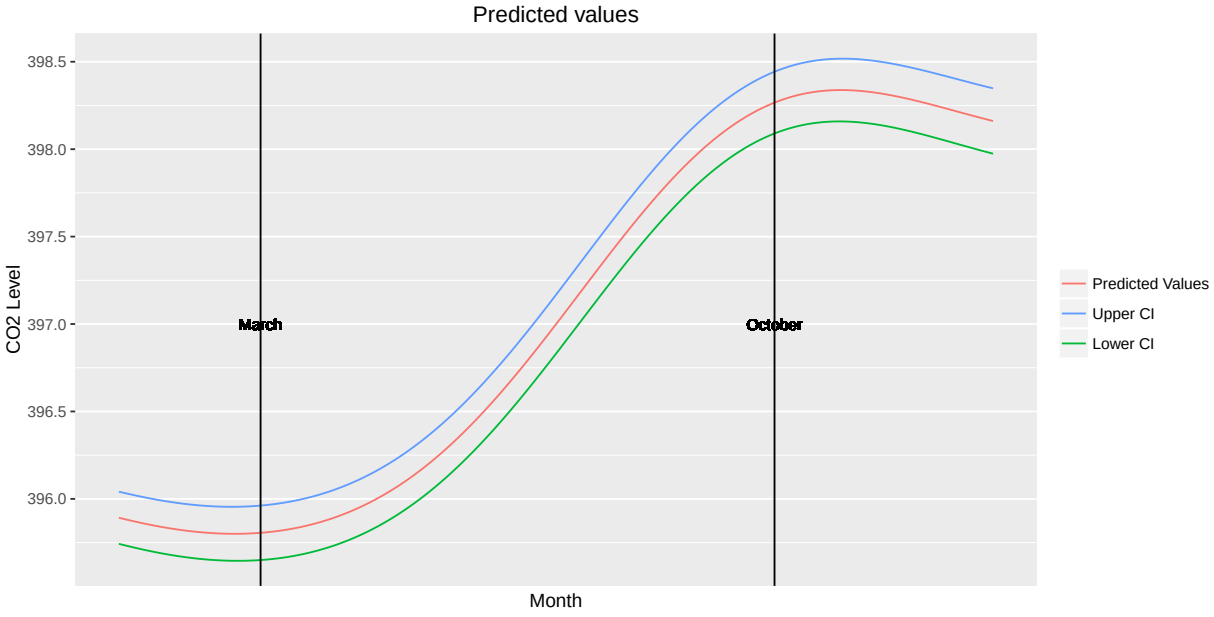

Figure 3: Seasonal Variation in Carbon Levels in 2015

We check the estimates of Carbon levels during the first days of March and October (of 2015) to see if there is a difference. For March the 2 standard deviation CI is (395.65, 395.96) and for October it is (398.09, 398.44). We see that the confidence intervals do not overlap, and so we can conclude that March and October have significantly different CO2 levels. The predicted CO2 level for each day of 2015, along with the confidence intervals are provided in figure 3.

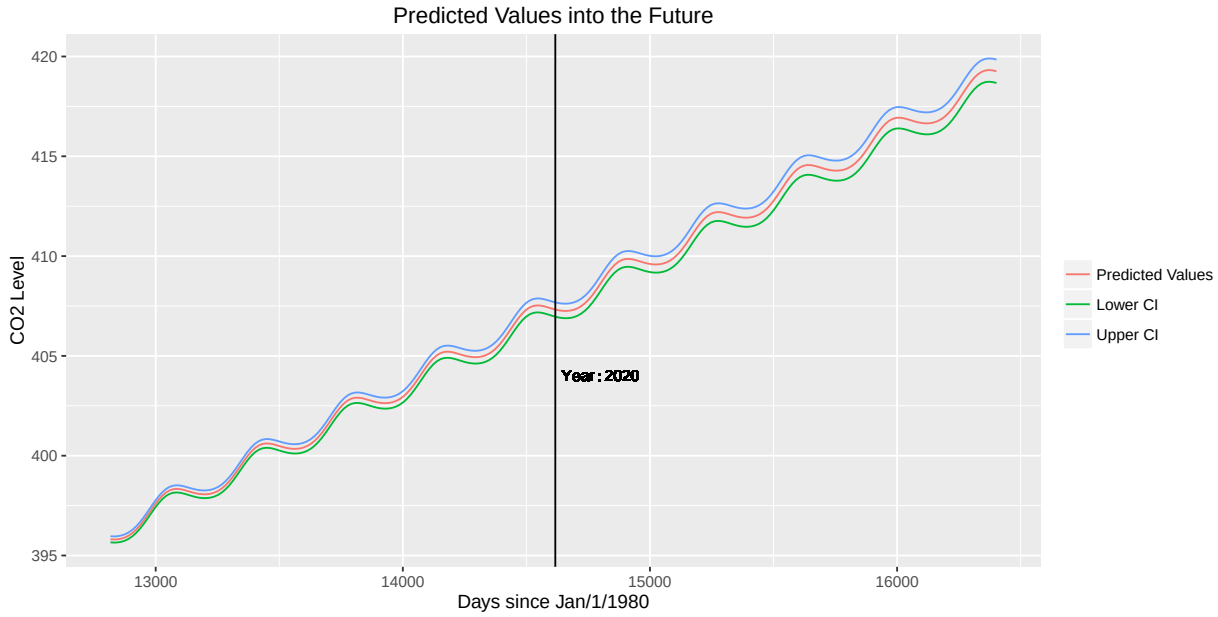

Figure 4: Predicted Carbon levels from Feb 2015 until Dec 2024

We plot predicted CO2 Levels for the next decade in figure 4. We can see that at 2020 Carbon levels are expected to be between 406.96 and 407.32 parts per gallon by 2020 with confidence of 2 standard deviations. From this we can conclude that, by 2020, Carbon levels will indeed exceed 400 parts per gallon.

Math

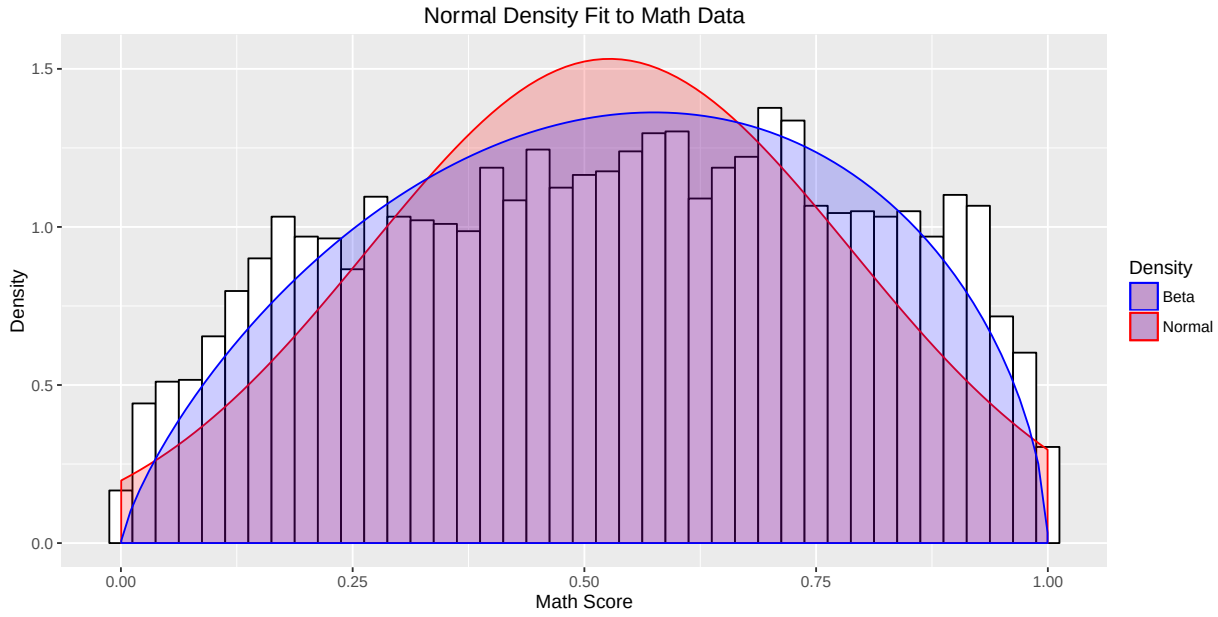

Figure 5: Assessing the fit of a Normal and Beta distribution to the data

We analyzed the `MathAchieve` dataset from the MEMSS package. We see that a normal distribution doesn't capture the heavy tails of this distribution, so we are compelled to try a model with wider tails. We used the `INLA` function from the INLA package to fit a Beta distribution to this data. Note that we first scaled the data so its domain would be the unit interval.

The model is as follows:

$$y_i \sim Beta(a, b)$$

where

$$E(y_i) = \frac{exp(\nu)}{1 + exp(\nu)}, \; Var(y_i) = \frac{\mu(1-\mu)}{1+\phi}$$
$$\nu = X_i\beta + U_i$$
$$a = \mu\phi \; and \; b = -a + \phi$$

Where:

- $X_i\beta$ contains the fixed effects: gender, minority status and SES.
- U_i contains the random effect, school.
- ϕ is the precision parameter, which has a Log-Gamma prior distribution with both hyperparameters set to $\frac{1}{10}$

We chose the Beta model since it had the capacity to fit data with as much variation as the math scores dataset. We also noticed that it had a well defined smallest and largest value (corresponding to the minimal and maximal test score attainable). We used a Log-gamma prior for the precision since it is non-negative and is a standard choice for fitting Beta distributions. The hyperparameters were chosen since they are the defaults that INLA assigns.

We compared the variance in school performance to within school performance. We constructed a 95% confidence interval of a baseline school (one without a random effect) using the above variance equation for the Beta fit. We found that test scores varied between (1.26, 25.24) with 95% confidence. A 95% confidence interval for the differences in average math test score between schools was (7.22, 18.8). From this we can conclude that the differences between schools is clearly smaller that what can be explained by within-school variation.

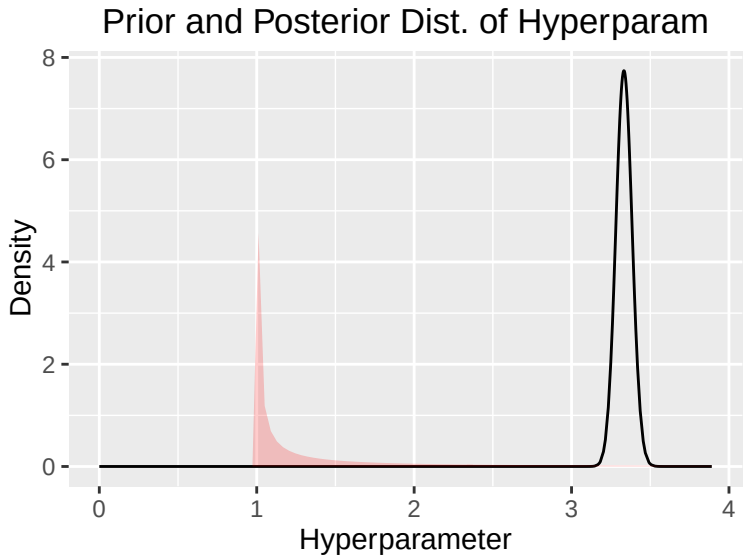

Figure 6: Prior (in red) and Posterior (in black) distribution

Table 1: The hyperparameters of the Beta Model

	mean	sd	mode
precision parameter for the beta observations	3.33	0.05	3.33
Precision for School	12.93	1.94	12.48

Moss in Galacia Redux

We looked at the `glgm` function in the geostatsp package.

The full model is as follows:

$$Y_i \mid U_i \sim \Gamma,\ E(Y_i \mid U_i) = \lambda(s_i)$$
$$\lambda(s_i) = X(s_i)\beta + U(s_i)$$
$$cov(U(s_i), U(s_j)) = \sigma^2 \rho((s_i - s_j)/\phi, \nu)$$

Where:

- Y_i is the lead levels taken from moss growing in or near the province of Galicia.
- $X(s_i)\beta$ is the intercept, the logarithm of the population density, dominant soil types and average annual rainfall.
- $U(s_i)$ is the spacial random effect.

An approximate 95% prior interval for the range, standard deviation and gamma shape parameter are as follows:

- ϕ has prior CI $\approx$ (20000, 100000)
- σ has prior CI $\approx$ (0.05, 0.75)
- ν has prior CI $\approx$ (5, 35)

We notice that based on the 95% CI the environmental variables do not influence the lead content of moss in Galacia. This is since the CIs for Leptosol, Podzol and Vertisol all contain 0 in them. They are (-0.19, 0.069), (-0.26, 0.49) and (-0.71, 0.20) respectively. Rain does not affect population either, as its 95% CI is (-0.0001, 0.0022). Lastly, population has a 95% CI of (0.01 , 0.16) and therefore does significantly affect lead content.

We then assessed the fit of the Gamma and transformed-Normal density functions to the empirical distribution of lead in Galician moss. Notice that we used a Monte Carlo approximation for the transformed-Normal density. The results are presented in figure 6.

We contrast this model with the model from Homework 2. We notice that both models are in agreement that humans have a significant effect on Galacian lead concentration, while the enviornmental variables do not. If we were examining spatial variation in Galician lead, we would prefer the model produced from INLA, since it has more flexibility in the selection of its priors.

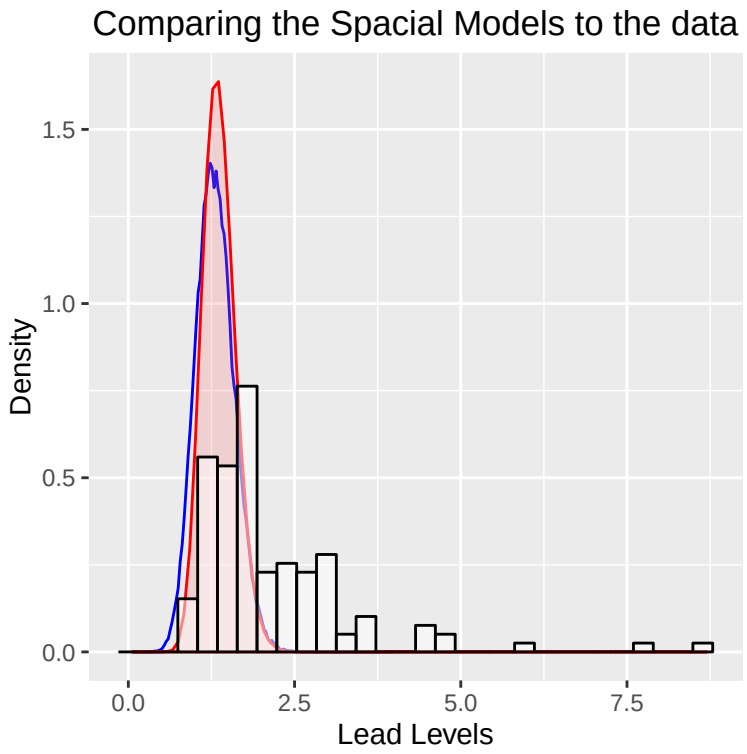

Figure 7: Assessing the fit of the Gamma (in red) and transformed-Normal (in blue) density functions to the empirical distribution of the moss data

Application

Summary

We analyzed the results of the 2014 American National Youth Tobacco Survey to help determine the if resourses should be focused on the state level or school level to combat children trying cigarettes for the first time. We found that the school a child goes to affects when they will smoke their first cigarette less than the state that they live in. We also found that rural males are more likely to start smoking than urban males, all else being equal. As such we recommend that resources be allocated towards statewide initiatives rather than towards schools with a previous history of smoking.

Introduction

We analyzed the 2014 American National Youth Tobacco Survey using an R version of the dataset available at pbrown.ca. The original dataset was released by the Center for Disease Control. The data was collected from a survey administered to 258 Schools across the United States. We wanted to explore whether the age children begin smoking cigarettes depends more on the state the student lives in or the school he or she goes to. We also investigated whether, controling for confounders, any two children are equally as likely to try smoking within a given period of time. Additionally, we investigated the differences between white urban males and white rural males in their smoking uptake habits.

Methods

Since children only start smoking for the first time once, we modelled the data using a Weibull distribution, as is convention for such survival analysis data. The specific model we used is as follows:

$$Y_{ij} \mid U_i, V_i \sim Weibull(\lambda_{ij}, k)$$

where

$$\lambda_{ij} = e^{-\nu_{ij}^k}$$

and

$$\nu_{ij} = X_{ij}\beta + U_i + V_i$$

Where

- $X_{ij}\beta$ is the subjects gender, ethnicity, whether they are from a rural or urban school
- U_i is the school random effect
- V_i is the state random effect
- The variance of U_i and V_i are themselves hyperparameters following a loggamma distribution
- k is the Weibull shape parameter and is normally distributed with its own hyperparameters.

Notice that the above model does not include any interaction terms among the given confounders. We did this for the purposes of model parsimony. For completeness, we included them in a seperate model discussed at the end of this report.

Table 2: Quantiles of Prior Distributions of Parameters

	Mu/Alpha HyperParam	SD/Beta Hyperparam	10th Quantile	90th Quantile
Prior on Weibull shape	0.6	0.60	0.84	3.93
Prior on SD of School	1.0	0.03	0.11	0.49
Prior on SD of State	1.0	0.10	0.21	0.99

We selected the hyperparameters of the above model using information from the collaborating scientists. They believed that some states should have about children starting to smoke about 2-3 times faster than others, but this magnitude is unlikely to exceed 5. By unlikely we interpreted it as with less then 10 percent probability. As such we chose our prior to have the variance exceeding 3 around 10 percent of the time.

Schools were expected to have less variability then states. It was expected that students going to different schools should only start smoking about 10% faster, and not more that 50%. As such we set our priors such that the 90th quantile was 1.6 (roughly 50% age difference when starting smoking).

Finally, we noted that the researchers expected that children were much more likely to start smoking the older they got. We set the 10th quantile to 0.84, corresponding to the children having a decreasing risk of starting smoking as they get older. This was since the researchers couldn't rule out this (improbable) possibility. Again we interpreted improbably to be of 10% probability or less. We set the 90th quantile to around 4- indicating that we believed (as the researchers did) that it was not likely that as a child aged they were 4 or more times more likely to begin smoking at any instant.

Results

Table 3: Posterior estimates of hyperparameters

Weibull Shape Parameter	SD of School	SD of state
5.47	0.24	0.41

Surprisingly, children were found to be over 5 times more likely to begin smoking at any instant as they got older. A less surprising finding was that there was more variability in the smoking rates between states then between schools. We plotted the prior and posterior of each of the parameters in figures 8-10.

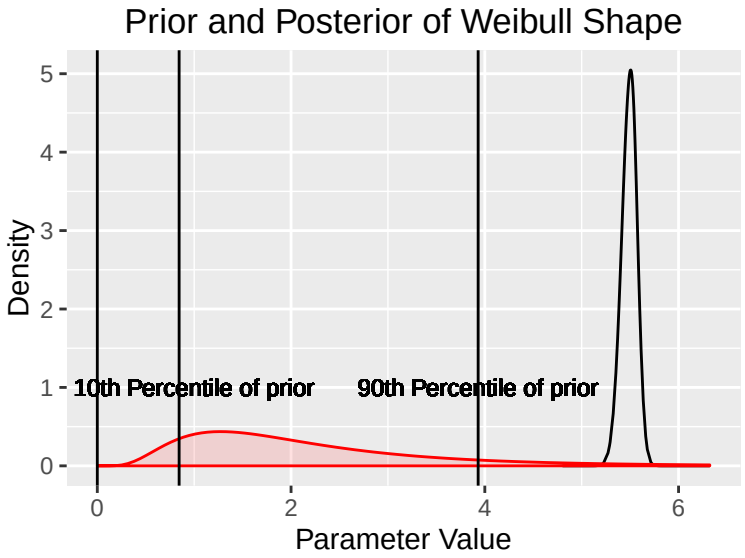

Figure 8: Comparing Prior (red) and Posterior (clear) distribution of Weibull Shape parameter

It is clear from figure 8 that cigarette smoking does not have a flat hazard function. Therefore, two non-smoking children do not have the same probability of trying cigarettes within the next month, even with all covariates fixed. This is since older children are more likely to begin smoking, all else being equal.

According to figures 9 and 10, geographic variation (between states) in the mean age children first try cigarettes is larger than the variation amongst schools. This supports the researchers hypotheses. Therefore, it is recommended for tobacco control programs to target states where smoking is a problem irregardless of any individual schools reputation.

In order to convey the differences between white urban males and white rural males in their smoking uptake habits, we looked at the fixed effects of our model. We found that, relative to urban dwellers, children who lived in rural areas were about 35% more likely to begin smoking. We found that the risk was anywhere between 14% to 60% higher with 95% confidence.

	mean	sd	0.025quant	0.975quant
RuralRural	1.35	1.09	1.14	1.6

Finally, we plot the density of the Weibull distribution of typical male students from urban and rural backgrounds. By typical we mean that we ignore the random effects from states and schools in our computation of typical corresponding scale parameters of each population. The results are presented in figure 11. We can clearly see that rural male youth start smoking on average earlier than their urban counterparts.

Finally, we ran INLA a second time with the interactions to see if there was any significant differences. We

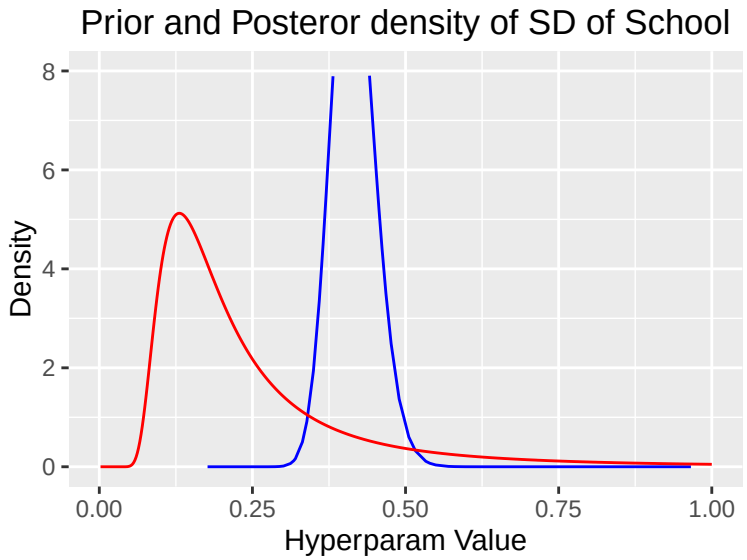

Figure 9: Comparing the prior (red) to the posterior (blue) of the SD of the school random effect

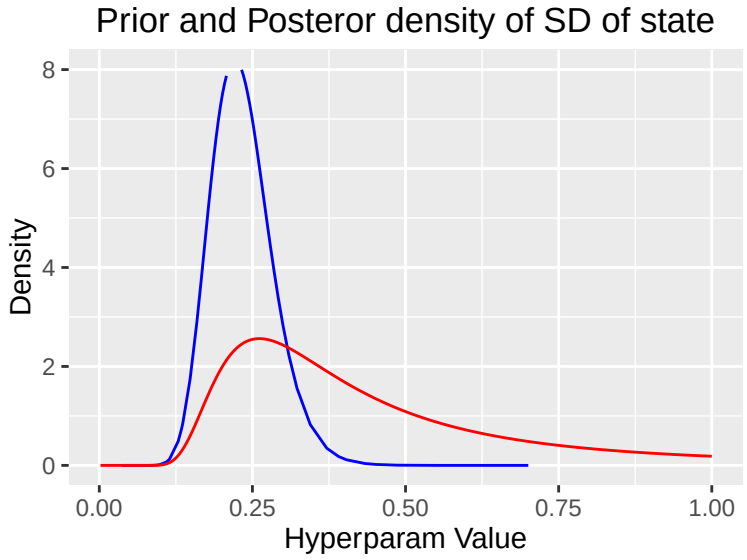

Figure 10: Comparing the prior (red) to the posterior (blue) of the SD of the state random effect

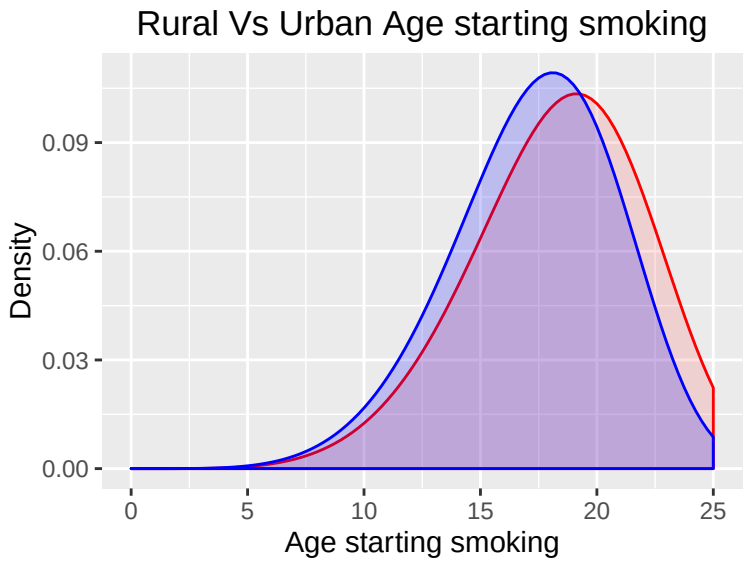

Figure 11: Comparing the ages that Urban children (red) start smoking compared to Rural children (blue)

found that the variance in schools were actually larger than the variance between states. Note that this would result in conclusions that are quite different than the one analyzed thoughout this report, and did not this model did not conform to the researchers expectations. These surprising results imply that more followup studies should be done. We state the values of the hyperparameters of the model with all of the interactions in table 5, and the reader is encouraged to compare this to the results from the first model which are presented in table 3.

Table 5: Posterior estimates of hyperparameters

Weibull Shape Parameter	SD of School	SD of state
5.47	0.42	0.23

Appendix

This file was made using the R markdown package. All code used in this paper can be accessed from within the code blocks of the markdown document.