# Assignment 3

*Steve Wilkins-Reeves*

*2018-04-02*

## Non-Parametrics

The Scripps $CO_2$ Program was initiated by in 1956 by Charles David Keeling. This program involved sampling $CO_2$ from the Mauna Loa Observatory in Hawaii. The four hypotheses under consideration involving atmospheric $CO_2$ include:

1. Although carbon in the atmosphere is still increasing, there are indications that the increase has slowed somewhat recently.

2. The data are consistent with carbon slowing during the global economic recessions around 1980-1982 and 1990

3. Carbon tends to be higher in October than March.

4. Carbon will likely exceed 400 parts per gallon by 2020.

Assessing the concerns of the hypothesis the following model is used.

$$
\begin{aligned}
&Y_i \sim N(\mu_i, \sigma^2) \\
&\mu_i = \beta_0 + \beta_{cos12} cos(2\pi \frac{X_{days}}{365.25}) + \beta_{sin12} sin(2\pi \frac{X_{days}}{365.25}) + \\
&\quad \beta_{cos6} cos(4\pi \frac{X_{days}}{365.25}) + \beta_{sin6} sin(4\pi \frac{X_{days}}{365.25}) + f(X_{days}; \nu)
\end{aligned}
$$

Here we include 12 and 6 month periodic trends captured by the sine and cosine functions in the model, as well as a smoothing thin plate regression spline function over time $X_{days}$ with smoothing parameter $\nu$ degrees of freedom, and intercept $\beta_0$.

We obtain the following plots from the model, with the degrees of freedom of the smoothing spline $\nu$ equal to 9.

```
## This is mgcv 1.8-17. For overview type 'help("mgcv-package")'.
```

In order to address the hypotheses in question the derivative of the underlying smoothing trend must be computed.

### Summary of Results

After fitting the smoothing model, and computing the derivative and confidence interval of the underlying trend after removal of seasonal variation, we find that the carbon in the atmosphere is increasing. The increase may have slowed, however the confidence interval on the rate of increase indicates that this cannot be confirmed.

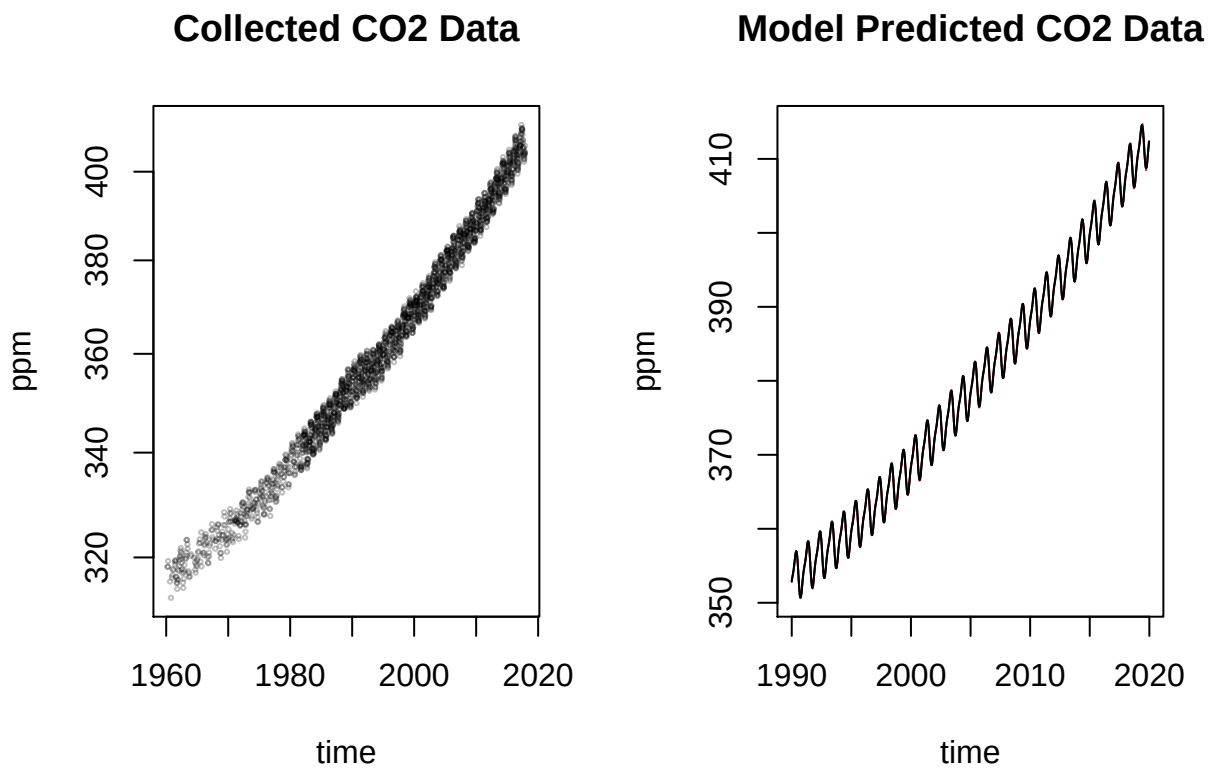


Figure 1: CO2 and Smoothing Model Trends. Prediction Displayed in Black With 95% Confidence Intervals in Red

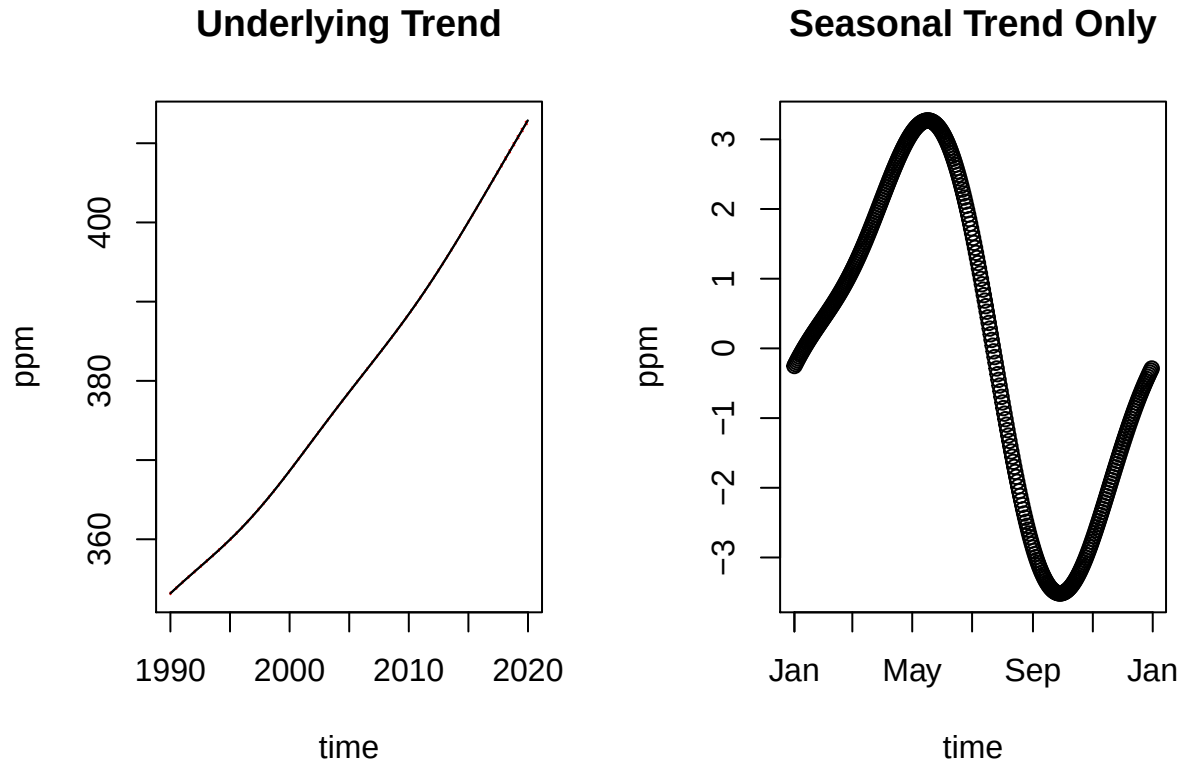


Figure 2: CO2 and Smoothing Model Trends. Prediction Displayed in Black With 95% Confidence Intervals in Red

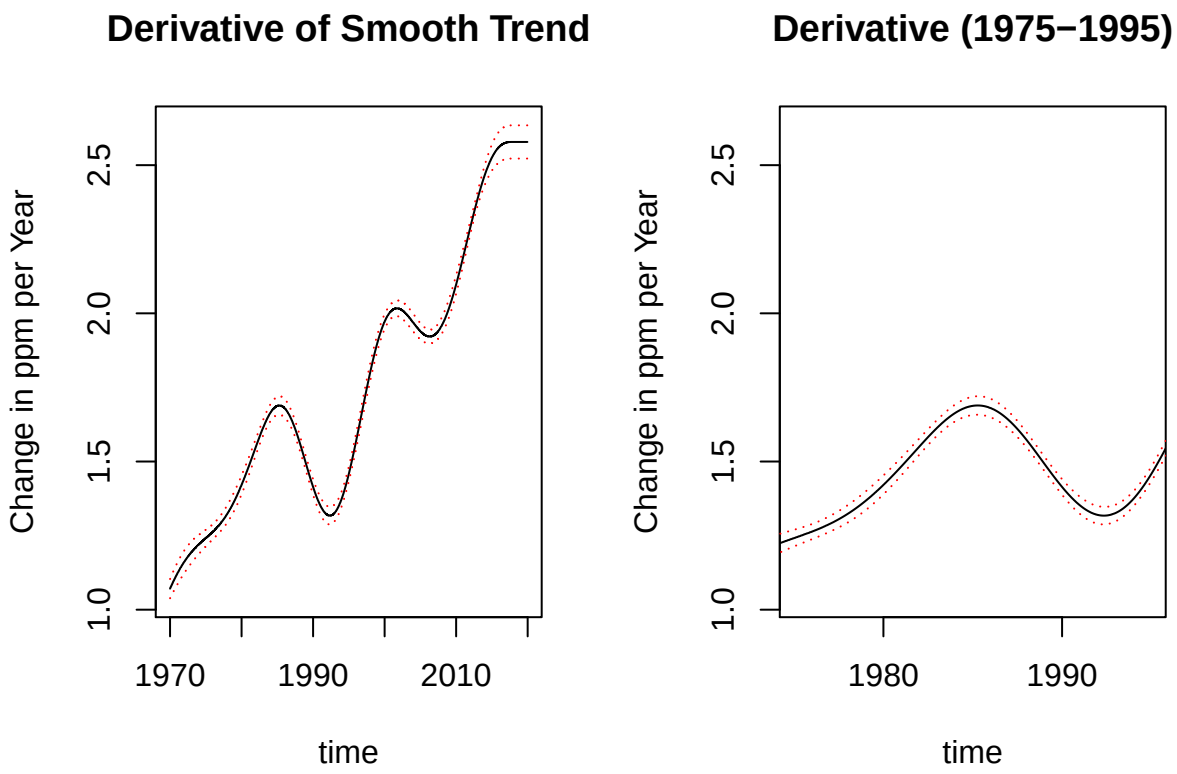


Figure 3: CO2 and Smoothing Model Derivative Trends. Prediction Displayed in Black With 95% Confidence Intervals in Red

From the underlying smoothed trend, we find that carbon emissions rate of increase is low from 1980-1982, however prior to this time, the rate of increase was also low. It can be seen that there is a decrease in the increase of the rate of emissions after 1990.

Thirdly, it is observed from modelling the data there is a trend in which carbon tends to be lower in October than March as seen in figure 1.

Lastly, we find that including the underlying trend surpasses 400 ppm by 2020. Additionally, even when we include regular seasonal variation, the carbon levels are still predicted to be above 400 ppm.

In future considerations, time series models may be an appropriate methodology for approaching these hypotheses.

## Math

First we remove negative scores.

We clearly have a well-defined left skew even after removal of the erroneous negative data. Observing a quick histogram can give us insight in to the structure of the data and an appropriate GLM to use.

The data suggests that this is a continuous distribution with positive support. This suggests a gamma generalized linear model may be appropriate for this data. The following model will be used.

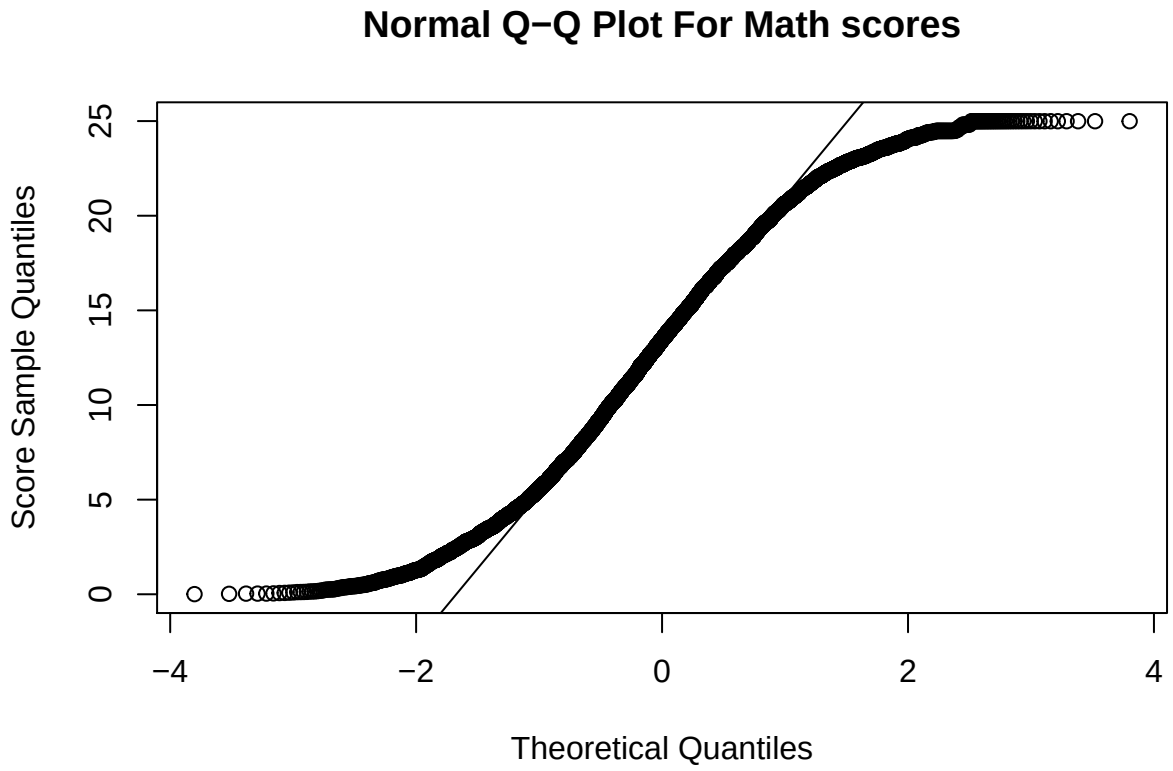


Figure 4: QQ Plot for Math Scores, Apparent Left Skew

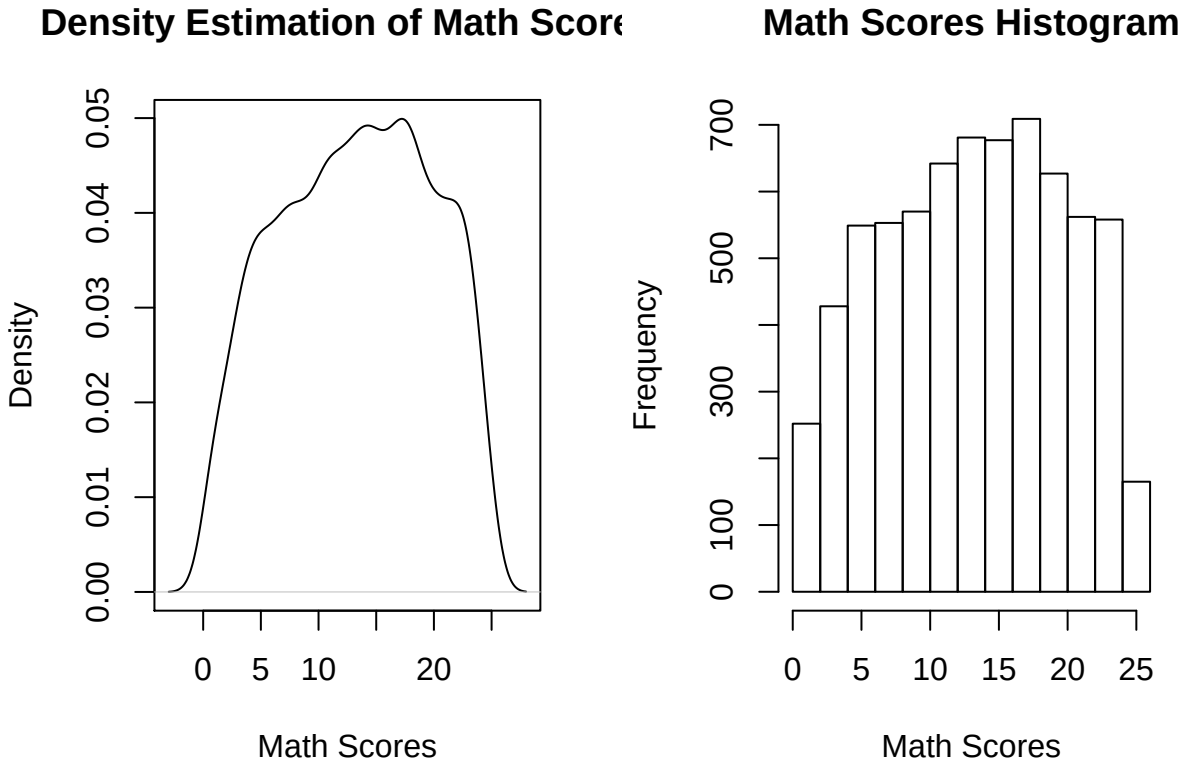


Figure 5: Density And Histogram of Math Scores

$$Y_i \sim Gamma(\frac{\lambda_i}{\nu}, \nu)$$
$$\lambda_{ij} = \beta_0 + \mathbf{X_{ij}}\beta + U_j$$
$$U_j \sim N(0, \sigma_U^2)$$

Here $Y_i$ is the math score, distributed with the gamma shape parameter $\nu$, $\mathbf{X_{ij}}$ is a vector of covariates including the minority status, socioeconomic status and their interaction effects, and the school random effect $U_j$.

Table 1: 95% intervals

| | 2.5% Quantile | 97.5% Quantile |
|---|---|---|
| School Average | 6.684 | 18.448 |
| Typical School | 8.293 | 17.314 |

Therefore since we find the 95% quantile interval of averages in schools, we find that there is a difference between schools that is greater than can be explained by within school variation.

```
## Loading required package: sp
```

```
## This is INLA_17.06.20 built 2017-06-20 03:44:36 UTC.
## See www.r-inla.org/contact-us for how to get help.
```

Lastly we find that the precision parameter for school had a 95% CI of [8.644,15.334]. This suggests that there is in fact a school effect beyond within school variation.

# Moss in Galicia Redux

1. A two dimensional spatial model is used according to the following. $Y_i$ is the measurement of lead in the soil at a particular location based on the population of the area, rainfall per year, and predominant soil type.

$$Y_i \sim Gamma[\frac{\lambda(s_i)}{\nu_\Gamma}, \nu_\Gamma]$$
$$\lambda(s_i) = \beta_0 + \beta_1 X_{log(\text{Pop})} + \beta_2 X_{\text{Rain}} + \beta_3 X_{\text{Soil}} + U(s_i)$$
$$\text{where } Cov(U(s_i), U(s_i + h)) = \sigma^2 \rho(h/\phi, \nu)$$

where $\rho$ is the Matérn family function with range parameter $\phi$ and shape parameter $\nu$.

$\nu_\Gamma$ is the shape parameter of the generalized linear model

The $\beta$ coefficients are associated with their respective linear parameters and the intercept of the model.

This is also a Bayesian model in which a PC (penalized complexity) prior was used for the $\sigma^2$ and $\nu_\Gamma$ parameters.

From figure 5 (in the assignment), we can estimate the 95% prior intervals to be the following:

Table 2: Table of Credibility Intervals of Priors

| Parameter | 2.5 CI | 97.5 CI |
|---|---|---|
| range | 5 | 50.0 |

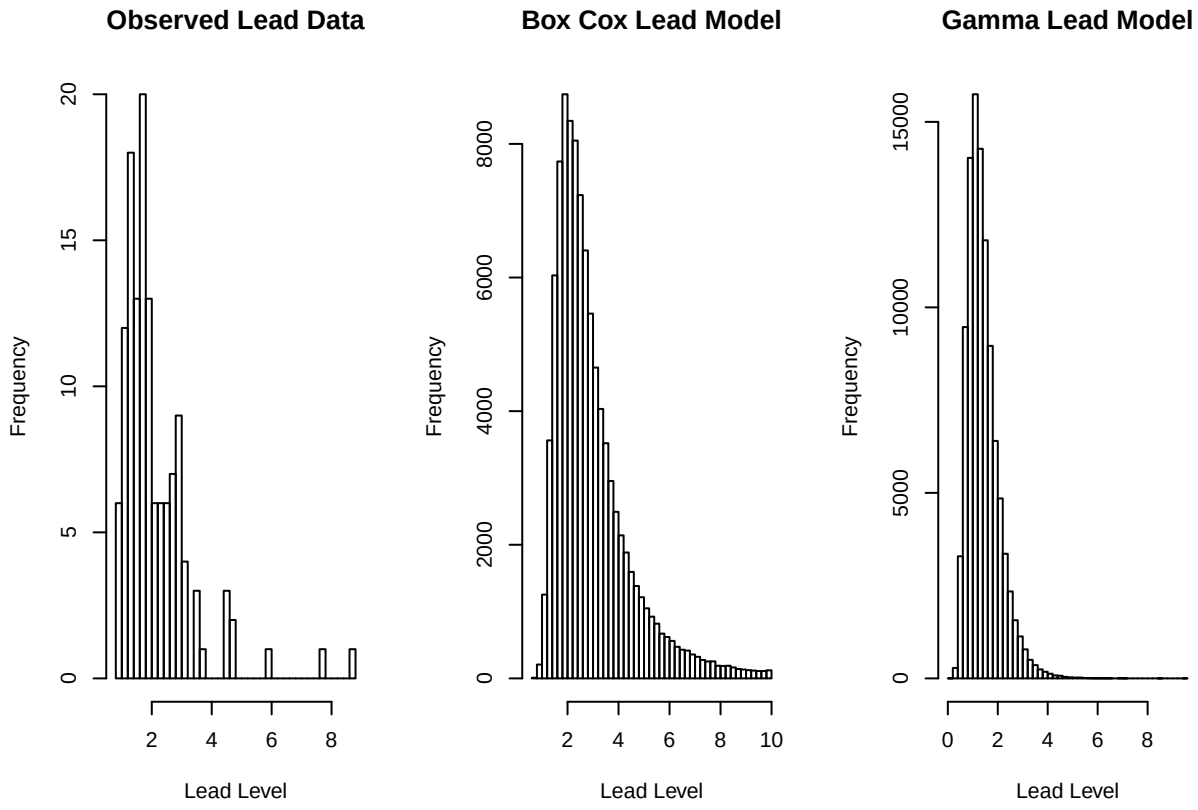


Figure 6: Comparison Of Lead Distribution and Models

| Parameter | 2.5 CI | 97.5 CI |
|---|---|---|
| sd | 0 | 0.7 |
| gamma shape | 3 | Inf |

This also includes the transformation of the gamma shape when displayed in the prior.

2. Based on the 95% credibility intervals, it appears that the population influences the lead levels, and though the credibility interval for the rain coefficient overlaps 0, it only narrowly does and likely has an effect.

3. We can display the Box-Cox transformed data, as well as the gamma transformed data from homework 2 and 3 respectively as seen in figure 6. The data seems to match the Box-Cox transformed distribution due to the heavier tails.

4. Both models attempt to explain the variation in the lead in the soil based on the population, soil type and rain of the area in Galicia using a spatially correlated model. The homework 2 model however firstly assumes a normal distribution (with a Box-Cox transformation of -0.52) and the homework 3 example uses a gamma distribution. The Homework 2 model additionally is a frequentist model, while the Homework 3 approaches it from a Bayesian perspective and uses a prior. This can be used to incorporate prior knowledge about a situation into a model. In order to test whether particular covariates are suitable predictors in a model, a likelihood ratio test was used in Homework 2, this found that the model which incorporated spatial variation and the population was the most appropriate, and

thus rain and soil type did not have an effect on the lead levels. In Homework 3's Bayesian model, population was a significant predictor, and though rain was found to be an insignificant predictor according to the 95% credibility region [-0.0001,0.0021] it was nearly significant, and may worth still being considered in the model. Additionally, the range parameter changed from 30.91 $km$ to 34.40 $km$, and the spacial variance from 0.25 to 0.4224. I prefer the second model in this case, because working with non-negative data (lead levels) it seems to be a more appropriate model. Additionally the simplicity of interpreting the significance of the model parameters is convenient in the Bayesian model.

# Application

## Introduction

Smoking and tobacco use is a major concern for the public and medical professionals. Health consequences such as increased rates of cancer are considered general knowledge about such products, however despite this knowledge, many individuals continue to use such products. Tobacco use habits most often begin prior to the legal age of purchase in many countries, (18- 21 depending on the state) and thus some health researchers and physicians consider tobacco use a pediatric disease. The 2014 American National Youth Tobacco Survey (NYTS2014) seeks to investigate the tobacco use habits of American middle and high school students. This report will seek to answer the following questions. Firstly, are the geographic variations (between states) in the mean age that children first try cigarettes greater than the variation between schools? Secondly, does cigarette smoking have a flat hazard function? In other words, is the risk of a child beginning smoking in the next month independent of age, given the known confounders of sex, rural/urban status, ethnicity, school and state are the same? Additionally, we wish to convey the difference between white urban males and white rural males in their smoking uptake habits.

## Methods

The survey data provides self-reported nationally representative data about middle and high school students' tobacco use habits. In particular, we are interested in the time that a student begins smoking. The time until initiating smoking is modeled as a survival analysis. The data is both left and right censored using the following model:

$$
\begin{aligned}
Z_{ijk}|Y_{ijk},\ A_{ijk} &= min(Y_{ijk}, A_{ijk}) \\
E_{ijk}|Y_{ijk},\ A_{ijk} &= I(Y_{ijk} < A_{ijk}) + 2I(Y_{ijk} \leq 8) \\
Y_{ijk} &\sim Weibull[\lambda_{ijk}, \alpha] \\
\lambda_{ijk} &= exp(-\eta_{ijk}) \\
\eta_{ijk} &= \mu + X_{ijk}\beta + U_{jk} + V_k \\
U_{jk} &\sim N(0, \sigma_U^2) \\
V_k &\sim N(0, \sigma_V^2)
\end{aligned}
$$

Where $A_{ijk}$ is the age of the respondant, $Y_{ijk}$ is the age at which a student began smoking, $E_{ijk}$ is an indicator for the censored data coded as 0 - They have never tried smoking (right censored data), 1- They have a recorded age of smoking after age 8, and 2- They have begun smoking prior to age 8 (Left censored Data). $U_{jk}$ is the school random effect and $V_k$ is the state level random effect. $X_{ijk}$ is a vector of covariates which include the sex of the students, the student's race and their interaction effects, as well as the rural or urban geographical status of the student. The hyperparameters include the rate parameter $\alpha$ as well as the school variance $\sigma_U^2$ and state variance $\sigma_V^2$.

The following information was given regarding the hyperparameters.

- The variability in the rate of smoking initiation between states is substantial, with some states having double or triple the rate of smoking update compared to other states for comparable individuals. It is not expected to see the 'worst' states having five or 10 times the rate of the 'healthiest' states.
- Within a given state, the 'worst' schools are expected to have at most 50% greater rate than the 'healthiest' schools, and differences of 10% to 20% in rates is more typical.
- Although a flat hazard function is expected, it is more likely that the hazard increases with age than decreases with age. The prior probability the hazard falls with age is less than 10%. It would not be unusual to see a quadratic or cubic increase in the hazard with age, but polynomial increases with age involving 5th or 6th powers is improbable.
- Here 'worst' or 'unlikely' refers to 10th percentile or 10% probability are of the right order of magnitude.

Applying these to the rate parameter, the statements regarding the variance of the random effects have been derived below:

$$\begin{aligned} P(-z_{0.05}\sigma < \eta < z_{0.05}\sigma) &= 0.9 \\ \text{Where } z_{0.05} &= 1.6448 \\ \Rightarrow P(e^{-z_{0.05}\sigma} < \lambda < e^{z_{0.05}\sigma}) &= 0.9 \end{aligned}$$

When considering the magnitude of the difference is at most $\omega = e^{2z_{0.05}\sigma}$. Therefore the following is required of the prior:

$$P(\sigma \in [0, \frac{log(\omega)}{2z_{0.05}}]) = 0.9$$

Therefore we can use a penalized complexity prior with parameters $(u, a)$ with which defines $P(\sigma > u) = a$. For the rate parameter $\alpha$ we wish to achieve $P(\alpha < 1) = 0.1$ as this corresponds to a decreasing hazard function and $P(\alpha > 6) \approx 0$. If it is not unusual to have a quadratic or cubic increase in hazard function we wish to have $P(1 \leq \alpha \leq 4) = 0.8$. A normal prior with $\mu = 2.5,\ \sigma = 1.170$ achieves this. We verify these assumptions and therefore the following priors were used.

Table 3: Model Covariates Expressed as Rates as well as random effects and alpha parameter

| Assumption | Verification | Value |
|---|---|---|
| Hazard function Decreases < 10% | Prob(alpha < 1) | 0.100 |
| Hazard function Unlikely To Be Greater Than Quintic Increase | Prob(alpha > 6) | 0.001 |

Table 4: Model Priors

| Parameter | Prior |
|---|---|
| alpha | N(2.5,sd = 1.170) |
| School SD | PC(u = 0.12326, a = 0.1) |
| State SD | PC(u = 0.33397, a = 0.1) |

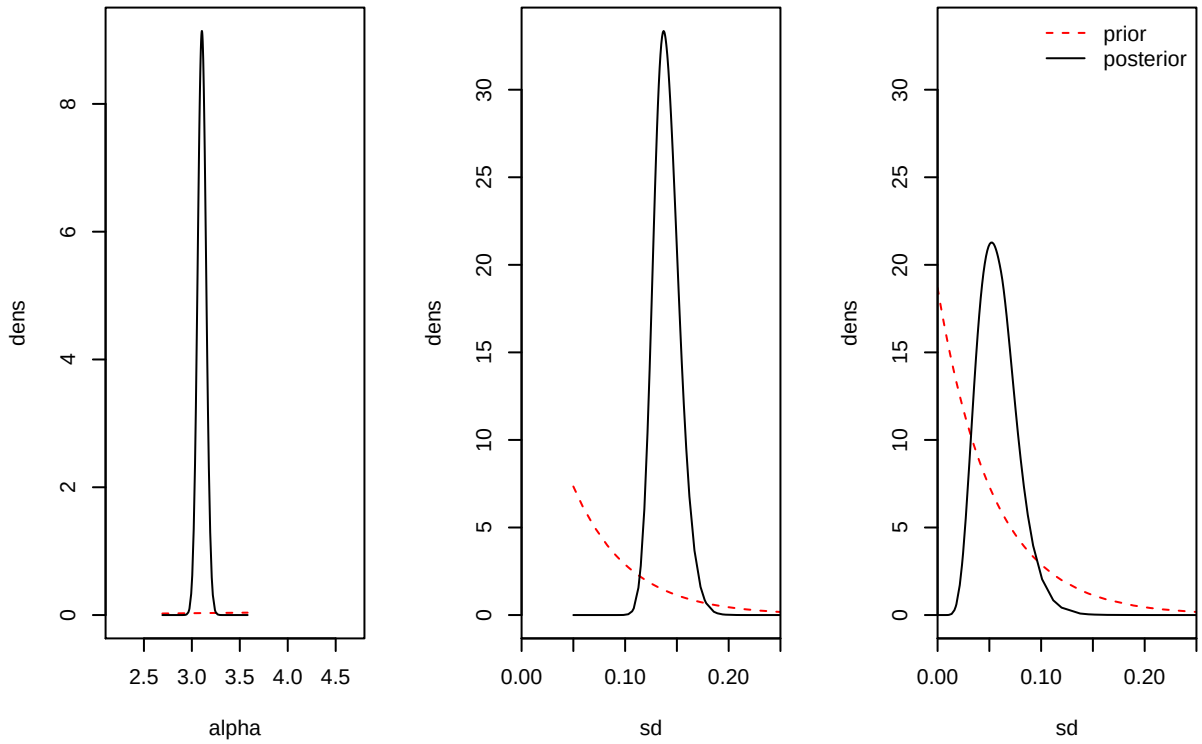


Figure 7: Prior And Posterior Distributions (from left to right), Alpha parameter, school level random effect Standard Deviation, State Level Random Effect Standard Deviation

## Results and Analysis

Along with these priors we have the following credibility intervals.

Thus when investigating the first hypothesis as seen in figure 7, we find that the geographic variation is in fact smaller than the variation within schools. Thus the tobacco control programs should not focus on states as a whole, but on the individual school where it plays a greater role.

Secondly, a flat hazard function corresponds to a value $\alpha = 1$. We have observed that the 95% credibility interval does not in fact overlap 1 ($3.105 \pm 0.086$). Therefore we find a closer to a quadratic hazard function.

The following hazard function was obtained.

$$h(x) = \alpha x^{\alpha-1} e^{-\eta_{intercept}}$$

Next to investigate the effects of the model parameters on the rate of smoking we obtain the following:

Table 5: Model Covariates Expressed as Rates as well as random effects and alpha parameter

| | mean | 0.025quant | 0.975quant |
|---|---|---|---|
| (Intercept) | 1.835 | 1.735 | 1.940 |
| RuralUrbanRural | 0.883 | 0.830 | 0.939 |
| SexF | 1.051 | 1.022 | 1.080 |
| Raceblack | 1.015 | 0.961 | 1.073 |
| Racehispanic | 0.958 | 0.918 | 1.000 |
| Raceasian | 1.217 | 1.109 | 1.342 |

| | mean | 0.025quant | 0.975quant |
|---|---|---|---|
| Racenative | 0.904 | 0.787 | 1.050 |
| Racepacific | 0.902 | 0.739 | 1.134 |
| SexF:Raceblack | 1.013 | 0.959 | 1.071 |
| SexF:Racehispanic | 0.981 | 0.938 | 1.025 |
| SexF:Raceasian | 1.003 | 0.888 | 1.134 |
| SexF:Racenative | 1.043 | 0.899 | 1.213 |
| SexF:Racepacific | 1.192 | 0.897 | 1.643 |
| RuralUrbanRural:Raceblack | 1.058 | 0.991 | 1.129 |
| RuralUrbanRural:Racehispanic | 1.029 | 0.976 | 1.085 |
| RuralUrbanRural:Raceasian | 0.937 | 0.818 | 1.079 |
| RuralUrbanRural:Racenative | 0.990 | 0.840 | 1.160 |
| RuralUrbanRural:Racepacific | 0.870 | 0.662 | 1.132 |
| SD for school | 0.140 | 0.119 | 0.167 |
| SD for state | 0.058 | 0.027 | 0.101 |
| alpha parameter for weibullsurv | 3.105 | 3.019 | 3.191 |

Thus from these covariate parameters we have found that white male students from rural areas tend to have a smoking rate of 0.883 of their urban counterparts with a 95% credibility region of (0.830,0.939). It was also found that females were found to have a higher smoking rate as well as Asians when compared to white males.

We were able to find a difference in the smoking rates of rural and urban white males, however, like all conclusions from Bayesian models, this does depend on initial assumptions. As moreinformation regarding the priors becomes available, this question may be worth revisiting with these considerations.

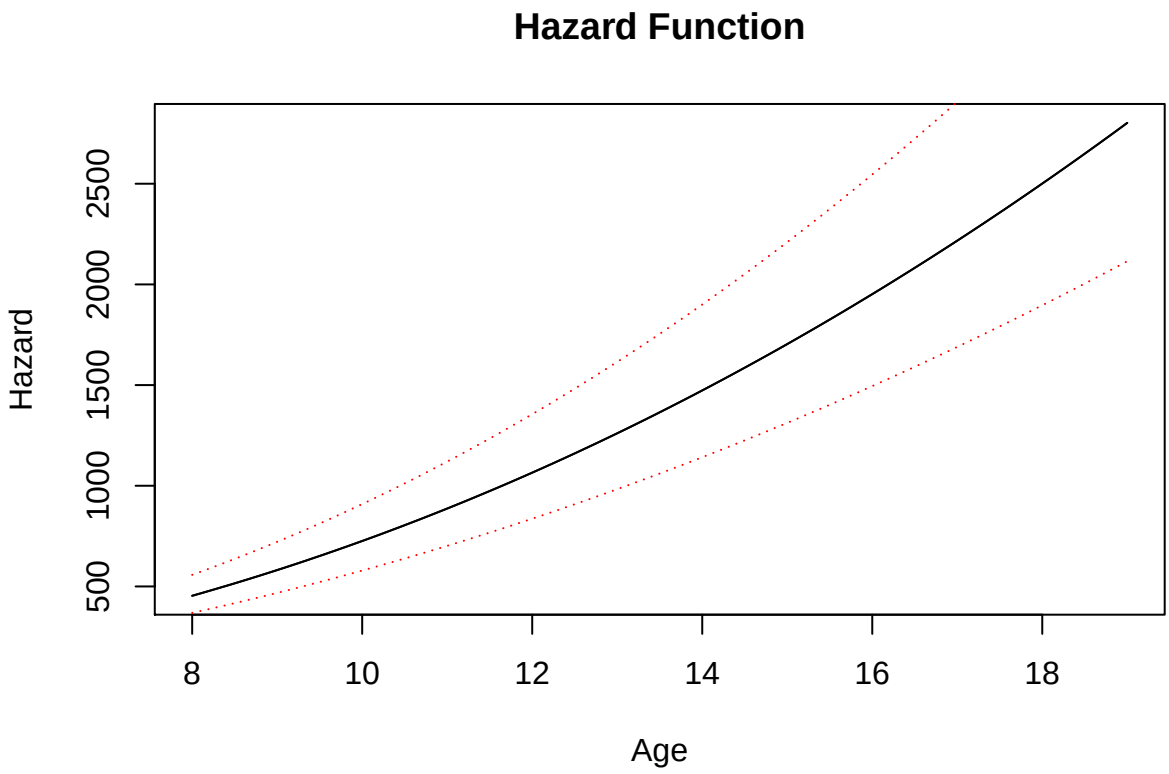


Figure 8: Hazard Function With 95% Baysian Credibility interval on alpha