Analysis of Smoking Uptake Age in the United States

Fiona McLean

Introduction

Cigarette use among adolescents is a major public heath concern. Tobacco product use is started and established primarily during adolescence. Nearly 9 out of 10 cigarette smokers first try cigarette smoking by age 18, and 98% first try smoking by age 26 (General 2014). Each day in the U.S. about 2,000 youth under 18 years of age smoke their first cigarette and more than 300 youth under 18 years of age become daily cigarette smokers (McCance-Katz 2017). In order to understand cigarette use among the youth and to better target prevention efforts, two main research questions were proposed and investigated. It was investigated if geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools. It was also investigated if two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages but provided the known confounders such as sex, rural/urban, ethnicity and school and state are identical. The 2014 American National Youth Tobacco Survey, a survey conducted by the CDC, which provides representative data regarding American youth's beliefs and behaviors towards tobacco, was used to investigate the above questions ("National Youth Tobacco Survey (Nyts)" 2019).

Methods

The data set used in the analysis is the 2014 American National Youth Tobacco Survey, a survey conducted by the CDC on young adults in the United States. The survey contains responses from 22,007 young adults on 162 questions regarding their use, perception of, and experience with tobacco.

As an exploratory tool, tables of mean age first tried cigarette and state were created, as well as tables comparing mean age first tried cigarette and school. Data for 9 year olds was removed due to suspicious results concerning the grade level of many 9 year olds respondents.

In order to determine if (1) if geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools and if (2) two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages but provided the known confounders (sex, rural/urban, ethnicity) and random effects (school and state) are identical, a Bayesian mixed censored survival model was constructed.

Bayesian models allow prior information to be incorporated into a model by specifying a prior and allow for complex models with many random variance factors to be fit, providing advantages over a frequentest model. The fixed effects in the model are the urban vs. rural location of the respondent and an interaction term between sex and race. We include the interaction between sex and race because we believe that men and women cigarette smoking behavior may be different between different races. The random effects are the school and state the respondent is from. School and state are random effects, as it is likely that the age an adolescent first tries cigarettes in different schools and states follow the same distribution. As the model is Bayesian, all factors have a prior distribution. We use a censored model in order to include information about respondents who have yet to try smoking, reducing bias in the model.

The model is:

$Z_{ijk}|(Y_{ijk}, A_{ijk}, U_j, V_k = min(Y_{ijk}, A_{ijk})$
$E_{ijk}|(Y_{ijk}, A_{ijk}, U_j, V_k = I(Y_{ijk} < A_{ijk})$
$Y_i \sim Weibull(\lambda_{ijk}, \alpha)$
$-log(\lambda_{ijk}) = \nu_{ijk} = X_{ijk}\beta + U_j + V_k$
$U_j \sim N(0, \sigma_U^2)$
$V_k \sim N(0, \sigma_V^2)$, where:

- Y_{ijk} is the age individual i in state j in school k smokes for the first time
- A_{ijk} is the age of individual i in state j in school k during data collection
- E_{ijk} is the indicator of if individual i in state j in school k has ever smoked
- λ_{ijk} is the scale parameter , α is the shape parameter
- $X_{ijk}\beta$ is the matrix of covariates including an intercept
- U_j is a random effect for state
- V_k is a random effect for school

With priors:
- $\theta_U \sim Exp(\frac{-log(.02)}{log(10)})$
- $\theta_V \sim Exp(\frac{-log(.05)}{4log(1.5)})$
- $\alpha \sim LogNormal(log(1), .7)$
- $\beta_0 \sim N(0, \infty)$
- $\beta_i \sim N(0, 1000)$

The priors for the fixed effects were chosen to follow a normal distribution with mean 0 and standard deviation 1000 and standard deviation ∞ for the intercept, since the prior is uninformative. Leading scientists in the industry determined that the variability of smoking initiation between states is substantial, with some states seeing double or triple the rate of smoking uptake compared to others, however seeing 10 times the smoking uptake is unlikely. Therefore, we put a prior an exponential prior on state standard deviation so that there is only a 2% chance of seeing a standard deviation between states being greater than 10. That is μ= log(10) and $\alpha = .02$. To determine the the λ parameter for the exponential distribution with these traits, the integral $\int_0^{log(10)} \lambda e^{-\lambda x} dx = 1 - .02$ was evaluated to yield $\frac{-log(.02)}{log(10)}$. Scientists also determined that the worst schools are expected to have at most 50% greater rate than the healthiest schools. Therefore, we put an exponential prior on the standard deviation between schools with μ= 4log(1.5) and $\alpha = .05$. That is, schools 4 standard deviations away from each other (or the best and worst schools) are expected to have more than a 50% difference in uptake rates 5% of the time. To determine the λ parameter for the exponential distribution with these traits, the integral $\int_0^{4log(1.5)} \lambda e^{-\lambda x} dx = 1 - .05$ was evaluated to yield $\frac{-log(.05)}{4log(1.5)}$. We also expect a flat hazard function, so the prior on the Weibull shape parameter should allow for 1, but is not believed to be 4 or 5. Therefore, we put a lognormal prior on the shape parameter with mean 1 and standard deviation .7. We do so since evaluation of the exponential of the normal distribution with mean log(1) and standard deviation .7 results in a 97.5 quantile value of 3.94. Therefore, it is very unlikely for us to see a shape parameter of 4.

Table 1: The table provides the distribution of the exponential of the normal with mean log(1) and standard deviation .7. We use this prior, since the mean for the shape parameter is 1, whcih we expect to see, and the 97.5 quantile is 3.94, making a shape parmater of 4 or 5 highly unlikely.

	2.5%	50%	97.5%
Shape Prior	0.25	1	3.94

Results

To gain some intuition as to how the uptake of smoking differs between schools and states, graphs were created to show the average age of smoking uptake in each school and state, as well as the range.

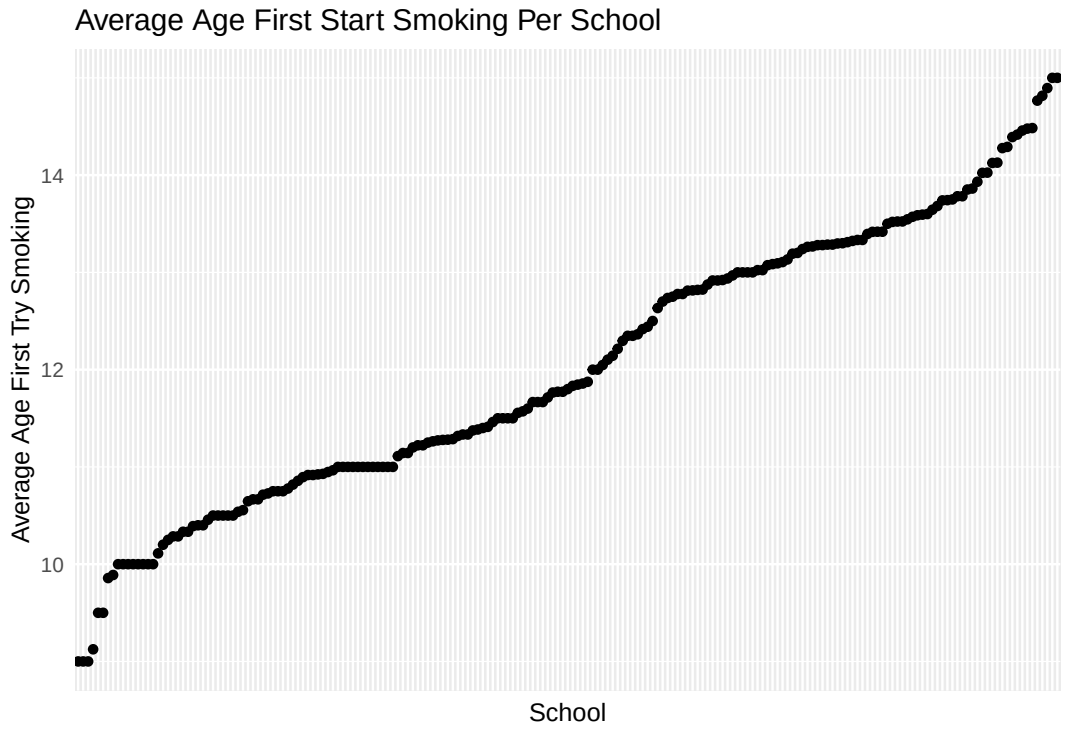

Figure 1: The graph above shows the range of mean ages that children start smoking in each school. The mean age ranges from approximately 9 to 15.

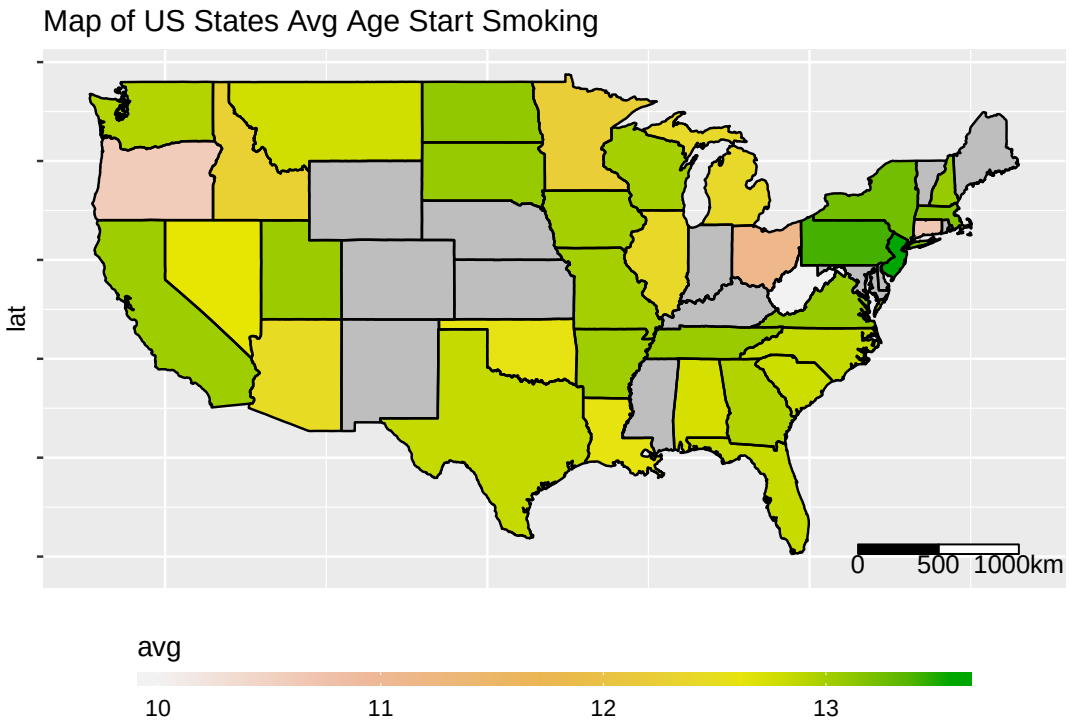

Figure 2: The map above shows the mean age of smoking uptake in each state. We can see that the mean age in certain states, such as Virginia is very low, nearly 10, while other states such as New Jersey have a mean uptake age of nearly 14. However, it appears the vast majority of states have mean uptake age from 12.5 to 13.5. States colored in dark grey do not have data.

It appears that the deviation in uptake age is smaller among states than among schools, although we will need statistical tests to confirm this prediction.

We now run the Bayesian mixed censored survival model described above to determine if geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools. We also use the model to determine if the hazard is flat, or if two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages.

Table 2: The table below provides the output for Bayesian mixed censored survival model. If the coefficient paramter is greater than 1, then the rate of uptake is slower for the group compared to the baseline of white males. For example, the rate of smoking uptake for rural residents is nearly 10.8% faster than the uptake rate for non rural children. For women, the rate of uptake of smoking is nearly 5.15% slower than for men.

	mean	0.025quant	0.975quant
(Intercept)	1.858	1.965	1.757
RuralUrbanRural	0.892	0.947	0.841
SexF	1.051	1.082	1.022
Raceblack	1.049	1.095	1.006
Racehispanic	0.975	1.009	0.942
Raceasian	1.215	1.333	1.114
Racenative	0.896	0.995	0.812
Racepacific	0.839	0.992	0.723
SexF:Raceblack	1.017	1.077	0.961
SexF:Racehispanic	0.984	1.030	0.940
SexF:Raceasian	0.994	1.130	0.876
SexF:Racenative	1.045	1.222	0.896
SexF:Racepacific	1.185	1.651	0.884
SD for school	0.151	0.127	0.178
SD for state	0.058	0.025	0.104

We can see that the mean of the posterior distribution for the standard deviation for the log relative rate of school is .151 and the mean for the posterior distribution of the standard deviation of for the log relative rate of state is .058. Notice the C.I do not overlap. Therefore we can conclude that there is higher deviation in the age of smoking uptake between schools than between states. Although we observe a wide CI interval for state, as the the standard deviation could be as low as .025 or as high as .104. However the standard deviation is lower than school, as the .025 quantile for school standard deviation is .127. Therefore, the rate at which children start smoking is between 2.53% and 10.36% faster in some states when compared to other states, while the rate of smoking uptake between schools is between 12.67% and 17.76% faster in some schools compared to others.

In order to determine if the hazard is flat, or if two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages, we analyze the alpha parameter.

Table 3: The CI for the alpha parameter for the weibull distribution is shown below.

	mean	sd	0.025quant	0.975quant
alpha parameter for weibullsurv	2.991	0.043	2.905	3.075

Since the mean of the posterior distribution of the alpha parameter is 3, and 1 is not in the credible interval, we do not have a flat hazard, the hazard is increasing. There is no evidence that as age increases, the propensity to smoke is constant.

The prior and posterior densities for the standard deviation of the log relative rate of school (left) and state

(right) is shown below.

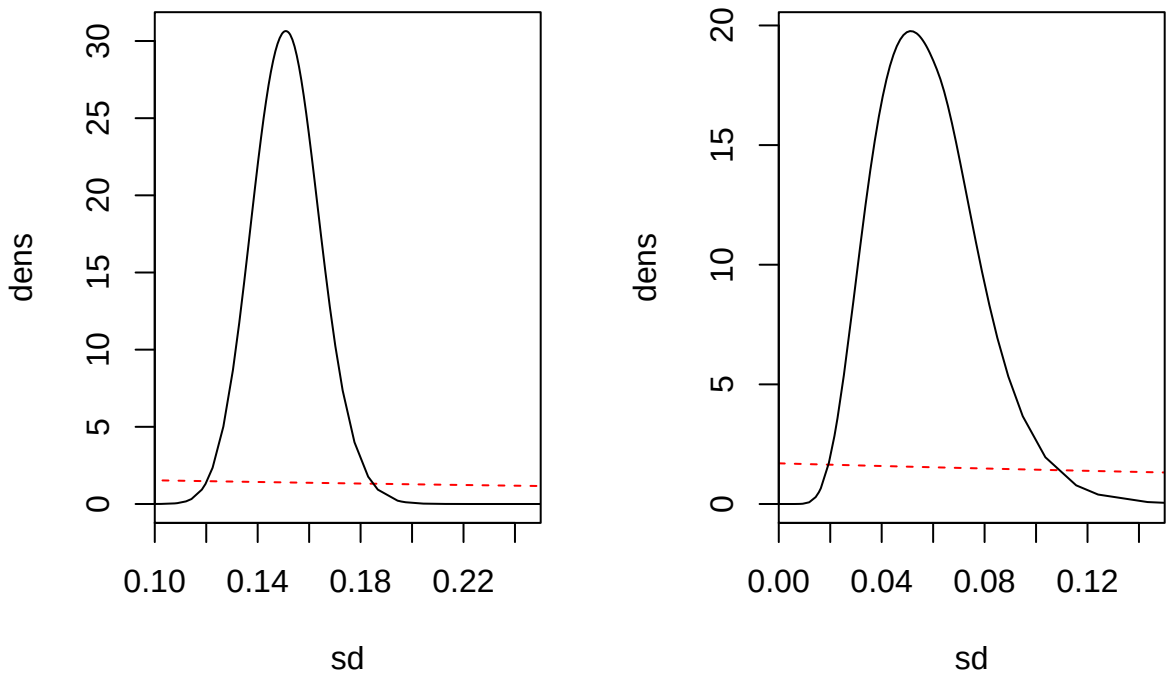

Figure 3: The prior and posterior density for school is shown on the left while the prior and posterior desnity for state is shown on the right. The prior density is the dashed red line while the posterior desnity is the black line.

Conclusion

By analyzing the posterior distribution of the standard deviation of the log relative rate in the uptake of smoking between schools and states, it was determined that school variation in the mean age children first try cigarettes is substantially greater than variation amongst states. As a result, tobacco control programs should target the schools with the earliest smoking ages and not concern themselves with finding particular states where smoking is a problem. Since the 95% CI of the posterior distribution of the alpha parameter does not include 1, we can conclude that there is no evidence that the uptake of smoking has a flat hazard function. As the hazard function is increasing, there is evidence to suggest that the older the child, the more likely the child is to try smoking within the next month.

References

General, Surgeon. 2014. "The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General." In *US Department of Health and Human Services*. Citeseer.

McCance-Katz, Elinore F. 2017. "The National Survey on Drug Use and Health: 2017." *Rockville, Maryland, USA: Substance Abuse and Mental Health Services Administration (SAMHSA)*.

"National Youth Tobacco Survey (Nyts)." 2019. https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm.