STATS 551 Report

Brody Erlandson, Andrew Heldrich, James Tang

April 2022

1 Introduction

This project is based on the Youth Risk Behavior Survey Surveillance System data set, collected by the CDC every two years. The data set includes responses to questions about risky behavior, ranging from violence and mental health. Our goal is to create a regression model to perform inference on car-related riskiness using Bayesian methodologies. Car crashes are a leading cause of death for young people in the U.S.[2], so being able to model the riskiness of a person's behavior can help inform policy or other education measures to reduce risky behavior. The regression model will use responses to other behavioural questions and demographic information as the covariates, while the response is a car-risk score constructed from four questions related to driving behavior. Many of the survey respondents do not answer every question, so the project will also compare different imputation methods including Bayesian model based imputation to estimate the joint probability of observing only a partial set of the input variables. For our final model, we use a generalized linear mixed effect model with a random intercept for the county. A gamma distribution is used for the likelihood model of the response to capture the heavily skewed data. After running a Monte Carlo Markov Chain on the data, the model resulted in several drug-use related questions strongly impacting the distribution. Additionally, there appeared to be distinct differences in predicted risk score across survey locations, with San Francisco and San Diego separating into a lower risk group.

2 Data Overview

2.1 Youth Risk Behavior Surveillance System (YRBSS)

The data is collected every two years, going back to 1991. Around 80 risk-related behavioral questions are asked to high-school students, ranging in topics from: sexual behaviors, alcohol and drug use, dietary behaviors, physical activity, and other injury causing behaviors. The full data set includes data from 4.9 million high school students in more than 2,100 separate surveys and many different counties. Our initial analysis will only use 2017, but we suggest ways to make a better model using multiple years in the extensions section. Below we can see a table summarizing some of the 2017 data. There are a total of 24 behavioral questions used in our analysis in addition to the demographic variables. Below we display a few, the full table is in the appendix. The list of all questions and response options are contained in the user guide[**yrbss**].

Table 1: Over view of a few variables used

Variable	Mean	Std	50%	IQR
Age	4.80	1.29	5.00	2.00
BMI	23.46	5.53	22.22	5.79
Q13	1.09	0.52	1.00	0.00
Q18	1.17	0.69	1.00	0.00
Q32	1.12	0.66	1.00	0.00
Q41	1.40	0.94	1.00	0.00

We can see from the table above that the questions are highly skewed. Figure 41 is also presented as an example of question responses. The distribution of the responses is typical of many questions, with a majority of the responses being the lowest risk category.

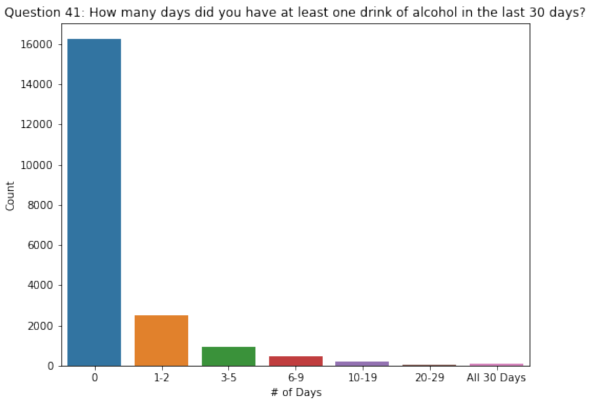


Figure 1: Responses to Question 41

In addition to the numerical variables, we did exploratory data analysis on the categorical variables such as sex, grade, and ethnicity. The result is presented at Table 2. The `grade` variable has a uniform distribution across all four grades from 9th to 12th. While `sex` is skewed towards males with 53% of our sample versus 47% that of females, `race` attribute also has a skewed distribution with white being the predominant race following by Black, Hispanic and Others.

Table 2: Decriptive statistics of categorical variabels, race, grade and sex

Race	Proportion
White	0.385
Black	0.230
Hispanic	0.200
Others	0.183

Grade	Proportion
9th grade	0.242
10th grade	0.260
11th grade	0.256
12th grade	0.240

Sex	Proportion
Male	0.538
Female	0.461

2.2 Car Risk Score Calculation

To create a car risk score, we use 4 questions that ask about car related activities. To make the this score evenly weighted for everyone, we only keep the panelists that answer all four questions. The car risk score is constructed using the following questions responses:

- How often do you wear a seat belt when riding in a car driven by someone else?

  a) Never 8

  b) Rarely 5

  c) Sometimes 3

  d) Most of the time 1

  e) Always 0

- During the past 30 days, how many times did you drive a car or other vehicle when you had been drinking alcohol?

  a) 0 times 0

  b) 1 time 3

  c) 2 or 3 times 5

  d) 4 or 5 times 7

  e) 6 or more times 10

- During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?

  a) I did not drive a car or other vehicle during the past 30 days 0

  b) 0 times 0

  c) 1 time 4

  d) 2 or 3 times ł6

  e) 4 or 5 times 8

  f) 6 or more times 10

- During the past 30 days, on how many days did you text or e-mail while driving a car or other vehicle?

  a) I did not drive a car or other vehicle during the past 30 days 0

  b) 0 days 0

  c) 1 or 2 days 2

  d) 3 to 5 days 3

  e) 6 to 9 days 5

  f) 10 to 19 days 7

  g) 20 to 29 days 8

  h) All 30 days 10

The answers are ranked from 0 to 10 (the score of the answer is highlighted in grey) according to the riskiness of the behavior. For example, driving a car when you had been drinking alcohol 6 or more times receives a score of 10. Each individual question score is then aggregated for the overall car risk score, then divided by the saturated score, where the saturated score is the maximum of each question. The score is between 0 and 1 with most scores close to 0, so we multiply by 1000 to make it an easier scale to see. Below we can see the formula for car risk score:

$$1000 * \frac{\sum_{i=1}^{4} A_{i\ score}}{\sum_{i=1}^{4} Q_{i\ max\ score}}$$

Where $A_{i\ score}$ is the panelist answer score for i^{th} question and $Q_{i\ max\ score}$ is the max score for question i. Figure 2. below shows the overall distribution for the car risk scores.

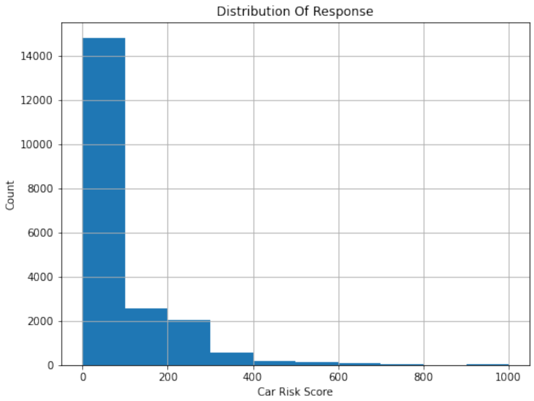


Figure 2: Distribution of car risk score

We can see from the graph that most people have a low score. It is important to note that since we go from 4 discrete answers to one continuous answer, we see a lot more variance than we would expect if the response was completely continuous. Moreover, this is survey data with high school students, and it is highly likely people answered more extreme or less extreme than reality. Given this high variance, we are particularly interested in a fairly biased model, where we can get some inference on what behaviors are cause a higher car risk. This, along with the fact that we have many different school districts, lead us to use a generalized linear mixed effect model.

3 Methods

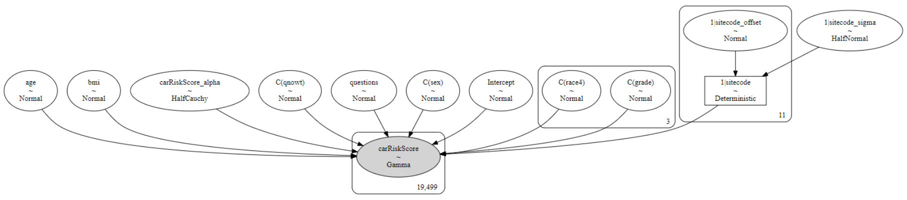
3.1 Motivation

The goal of our modeling is to capture what effects are associated with a higher car risk score. There are two main things we want to pull from our model, inference on which effects are statistically significant and creating a model that does not fit too much to the noise of the data. Our data is expected to be noisy, so using a regression model would be the best for predicting the outcome to capture general trends. Since the data is clearly skewed to the right, we decided to use gamma regression. Also, there are many different school districts across the U.S. used in the data set. It is highly likely that each district has some inner district correlations compared to the other counties, given students interact with other students in their own school district. Given this, we use a random effect of the district to counteract the inner district correlation.

3.2 Generalized Linear Mixed Effect Model (GLMM)

As mentioned above we are using a gamma mixed effect model with a random effect on district, which is a type of GLMM. We use the log link in our model. Our parameters are mostly other questions from the survey, which are asked in order of risk, so we treat them as ranks given the ordinal nature of the questions. There are some parameters which are not survey questions, that are continuous or nominal variables. For the nominal variables, we use dummy variables to have a parameter for each category. For the continuous and ordinal parameters, we treat them as continuous variables.

We do not expect the questions or panelist descriptive parameters to have a different effect on car risk across different counties; however, the car risk is likely higher or lower and have a different variance in across counties. For this reason, and for the computational benefit, we only use a random intercept. We can see a graph of the model below:



For each node in the graph, we see the parameter associated with it and the prior distribution used. For everything other than `carRiskScore_alpha`, `carRiskScore`, `1|sitecode_offset`, and `1|sitecode_sigma` the prior distribution is for the β coefficient of the model, not the prior distribution of the values of the parameter. The "questions" node represents all of the parameters for each question within the model. This was condensed into a single node for sake of the visualization.

Apart from graphically represent our model, we can explain the model in mathematical languages with the equation:

$$Y_{i,site} = \beta_0 + \beta_1 Sex_{i,site} + \beta_2 Age_{i,site} + \beta_3 Grade_{i,site} + \beta_4 Race_{i,site} + \beta_q Questions_{i,site} + b_{0,site} + \epsilon_{i,site}$$

Because the mixed effect model is composed of two parts, fixed effects and random effects. The fixed effect have a global mean and variance for the coefficients. Whereas, the random effects has its group specific site mean and variance apart from the global mean. We see this represented with $b_{0,site}$. Lastly, to save space we represent all questions like $\beta_q Questions_{i,site}$; however, there are many question that each have their own coefficient.

3.3 No Imputing Model

The first model we wanted to start with was a model with no missing values. Since most people are missing at least one value, if we dropped people with missing values we would end up with little data. Thus, we only took questions that had at least 90 percent responding panelists. This means we have fewer covariates than the other models. This will give us a baseline for the other models, and see how much better the later models.

3.4 Median Impute Model

The first imputation method we tried was a median impute. We use this instead of the mean because a majority of questions are skewed. As most people choose one of the least risky answers, but only a few pick the more extreme. The mean would capture these extremes too much.

With imputing, we can accept more questions than we could in the no imputing model. This adds a lot of questions, giving us a better ability to capture the types of behavior that is associated with risky driving.

Whenever you impute with one value, it reduces the variance of the data. This, especially when you are trying to estimate distributions, can lead to problems of being too biased. Also, given we are interested in the higher car risk, which is not the average person, assuming everyone has a particular mean can cause problems. With the next model, we can see if the median impute model may be too biased.

3.5 Model Based Impute Model

We also tested a Bayesian imputation model for missing data. Since this takes into account other parameters when assigning the missing values, it could better capture the more extreme responses as we expect the panelists who are extreme in similar questions to be extreme in the missing question. This will, hopefully, keep the variance at its "true" level.

For the model based imputation, we will assume that data is missing at random, a reasonable assumption given that we are dealing with survey data. A more complex analysis would not hold this assumption and we would need to model the relationship between the parameters and the missingness of the data. Additionally, for ease of computation and model setup, we will assume a multivariate normal for the variables which results in the need to approximate the posterior distribution of unobserved values:

$$p(\theta, \Sigma, Y_{missing} | Y_{observed})$$

To accomplish this we construct a Gibbs sampler that will sample data for the missing values conditioned on the observed values. The imputed samples are drawn from the following distribution[1]:

$$y_{[b]} | y_{[a]}, \theta, \Sigma \sim MVN(\theta_{b|a}, \Sigma_{b|a})$$

where b refers to the indices of missing data and a is the indices of observed data. Note that this could result in values over and below the score bound. We will then reassign values that happen to be beyond the bounds to the appropriate bound.

4 Results

Four separate markov chains are run to draw 10,000 samples from the posterior distributions of the model parameters using 19,499 data points. Table 3. below summarizes the posterior mean for parameters with a high density interval that did not include zero. The high density interval is the range from the 2.5th quantile to the 97.5th quantile. The full posterior distribution summary is listed in the appendix.

Table 3: Posterior means and quantiles for selected variables with a high density interval that does not include zero.

	Posterior Mean and 95% High Density Interval		
	Median Impute	Model Impute	No Impute
Term			
Intercept	3.530 [3.184, 3.879]	3.752 [3.408, 4.100]	3.568 [3.206, 3.931]
Q47	0.060 [0.035, 0.085]	0.051 [0.026, 0.076]	0.205 [0.172, 0.238]
Q57	-0.109 [-0.160, -0.058]	-0.116 [-0.168, -0.065]	-0.196 [-0.265, -0.127]
Q70	-0.027 [-0.041, -0.013]	-0.028 [-0.041, -0.014]	Not Included
Q15	0.071 [0.028, 0.114]	0.063 [0.020, 0.106]	Not Included
Q16	0.088 [0.045, 0.132]	0.069 [0.026, 0.112]	Not Included
Q17	0.057 [0.030, 0.084]	0.056 [0.028, 0.084]	Not Included
Q67	-0.098 [-0.125, -0.070]	-0.097 [-0.125, -0.068]	Not Included
Q68	-0.045 [-0.066, -0.025]	-0.040 [-0.061, -0.020]	Not Included
Q46	0.046 [0.033, 0.059]	0.039 [0.027, 0.052]	Not Included
Q19	-0.128 [-0.210, -0.046]	-0.118 [-0.200, -0.036]	Not Included
Q12	0.060 [0.030, 0.091]	0.063 [0.031, 0.094]	Not Included
Q41	0.154 [0.121, 0.187]	0.157 [0.123, 0.190]	Not Included
Q40	0.049 [0.036, 0.061]	0.047 [0.035, 0.058]	Not Included

The parameter summaries are generated for each imputation method. The model that excludes questions with no responses results in a much smaller list of estimated parameters. In general, there is high similarity in the parameters between using median and bayesian model imputation. Of note are question coefficients with a larger magnitude:

- Question 19: Have you ever been physically forced to have sexual intercourse when you did not want to?
- Question 41: During the past 30 days, on how many days did you have at least one drink of alcohol?
- Question 57: During the past 12 months, has anyone offered, sold, or given you an illegal drug on school property?

Moreover, when looking into the topic of each of the questions we see the following:

Table 4: Topic of questions that do not include zero in the 95% interval

Question	Topic
Q47	Drug Use
Q57	Drug Use
Q70	Diet
Q15	Violence
Q16	Violence
Q17	Violence
Q67	Health
Q68	Health
Q46	Drug Use
Q19	Sexual Violence
Q12	Violence
Q41	Drug Use
Q40	Drug Use

We can see from the table above that the questions that seem to be associated with the car risk score mostly have to do with drug use and violence. There are some health and sexual violence related questions associated also. This seems reasonable to be associated with the response. The specific questions are listed in the YRBS documentation[3].

Taking a closer look at question 41 in particular, Figure 4. shows the distribution of the coefficient and the trace of the markov chain. The posterior distribution shows the coefficient is clearly non-zero and a stationary trace of the sampling demonstrates convergence.

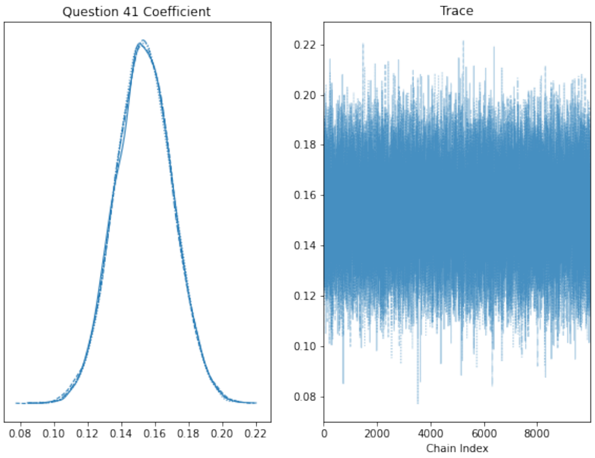


Figure 3: Posterior distribution and monte carlo trace for question 41 coefficient in the median imputation model.

The hierarchical model allows us to analyze possible differences across sites or regions in the data. Table 5. and Figure 4. below summarize the posterior distributions for the site intercept offsets. There appears to be significant overlap across most sites, but three groups somewhat emerge: a lower risk group of San Diego and San Francisco, a higher risk group of Cleveland, and the rest of the sites in a neutral/average group.

Table 5: Posterior Means and Quantiles for Site Intercept Offsets

	Posterior Mean and 95% High Density Interval		
	Median Impute	Model Impute	No Impute
Site			
Cleveland, OH	0.273 [0.137, 0.415]	0.277 [0.120, 0.437]	Not Included
Chicago, IL	0.025 [-0.111, 0.162]	0.043 [-0.115, 0.199]	0.038 [-0.127, 0.207]
Duval County, FL	-0.016 [-0.147, 0.112]	0.003 [-0.152, 0.151]	0.015 [-0.139, 0.170]
Broward County, FL	-0.056 [-0.205, 0.092]	-0.049 [-0.220, 0.119]	0.003 [-0.176, 0.186]
Fort Worth, TX	-0.013 [-0.142, 0.117]	0.012 [-0.139, 0.160]	0.081 [-0.073, 0.236]
Miami-Dade County, FL	0.021 [-0.111, 0.152]	0.044 [-0.109, 0.194]	0.101 [-0.054, 0.260]
Orange County, FL	-0.020 [-0.163, 0.122]	-0.006 [-0.170, 0.153]	0.038 [-0.128, 0.208]
Philadelphia, PA	0.193 [0.056, 0.336]	0.209 [0.051, 0.371]	0.214 [0.051, 0.388]
San Diego, CA	-0.276 [-0.414, -0.147]	-0.317 [-0.475, -0.168]	-0.222 [-0.382, -0.067]
San Francisco, CA	-0.222 [-0.361, -0.090]	-0.340 [-0.500, -0.189]	-0.364 [-0.533, -0.207]
Shelby County, TN	0.091 [-0.046, 0.230]	0.103 [-0.056, 0.260]	0.103 [-0.061, 0.274]

With the figure below we can see the above distributions plotted. On the left-hand side we see that most counties are right around zero, with a few going above. There are two that are distinctly below, as previously mentioned. We can also see the trace suggests the site intercepts have converged.

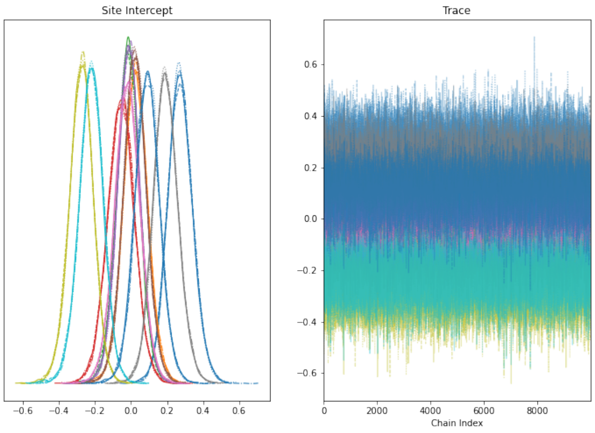


Figure 4: Posterior distributions of site intercept offsets in the median imputation model

Using the above information, we can calculate the full posterior predictive distribution of a new person from each site. This will enable us to determine the likelihood of a random person having a higher risk score than people from other parts of the country. We will use every 10th sample from the posterior predictive distribution to reduce auto-correlation across samples. Table 6. below displays the probability that a random person samples from a site has a risk score greater than the maximum risk score from a random sample across all other sites: $Pr(Y_j^* > max\{Y_1^*, ..., Y_{-j}^*\}|Y)$. As intimated in the site intercept figure above, the sites separate into three groups: low, medium, and high. SD and SF have the lowest chances of having the maximum risk score, while Cleveland (CE) has the highest.

Table 6: Posterior predictive probability that a random person from a site will have a higher risk score than a random person from all other sites.

	Median Impute	Model Impute	No Impute
Site			
Cleveland, OH (CE)	0.140	0.137	Not Included
Shelby County, TN (ST)	0.117	0.116	0.129
Philadelphia, PA (PH)	0.113	0.118	0.125
Chicago, IL (CH)	0.101	0.101	0.11
Miami-Dade County, FL (MM)	0.100	0.102	0.116
Duval County, FL (DU)	0.089	0.091	0.1
Fort Worth, TX (FW)	0.088	0.088	0.111
Broward County, FL (FT)	0.083	0.082	0.1
Orange County, FL (OL)	0.080	0.084	0.098
San Diego, CA (SA)	0.049	0.046	0.065
San Francisco, CA (SF)	0.042	0.034	0.046

For easier visualization, we created kernel density estimates of the full posterior predictive distribution of the highest and lowest probability sites from above. The distribution for the CE site has a much heavier tail indicating a higher chance of a random person having a high car risk score.

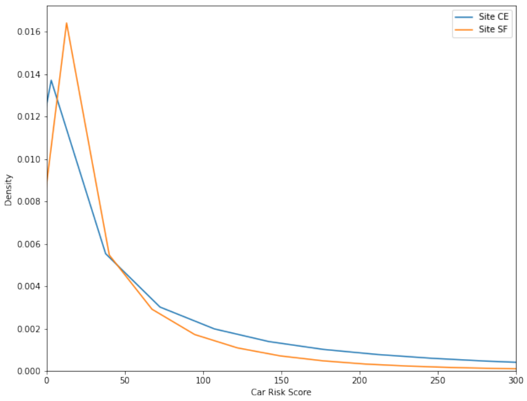


Figure 5: Kernel density estimates of the posterior predictive distributions for two selected sites in the median imputation model.

To asses the quality of our models and to provide an additional comparison of imputation methods we calculate a marker that estimates the probability that a sample will have a car risk score greater than or equal to 200: $t(y) = P(Y^* \geq 200|Y)$. The response is heavily skewed and decays quickly, so we want our model to estimate this behavior well. Table 7. summarises these quantities below compared to the observed proportion. The values are generally close to the empirical value, a good sign that the model is at least not completely mis-specified. Moreover, due to the high noise expected of the data, the models seem to capture the trend without capturing the noise. Though we should look into why this may be happening more closely, it does not indicate huge concern of out models at their current state.

Table 7: Posterior predictive probability of having risk score greater than or equal to 200 compared to empirical proportion.

Observed Data	10.9%
Median Impute	8.28%
Model Impute	7.59%
No Impute	6.94%

From the table above, we can see there does not appear to be any clear advantage to using a bayesian imputation method. The computational costs do not provide major benefits and using a simpler imputation strategy still provides similar results. The asymptotic complexity of using a gibbs sampler to impute missing data results results in $O(ns)$, with s being the number of monte carlo samples. Additionally, the assumption that the data is missing at random has a strong chance of being violated. Due to the sensitive nature of the survey, it is possible certain questions are not answered due to how the previous questions are answered. We also treated the responses as continuous and placed a normal distribution on their priors. An ordered probit regression model that utilizes a latent variable is likely a more appropriate model. In this type of model, the latent variable serves as the regression response and the observed values are a function of this latent variable.

5 Extensions

5.1 The Idea

Given the data is temporal, it would be nice to use the data over the years. One way to do this is to add a random effect for year; however, often questions are added over the years, so this causes problems with having one big model. In addition, the entire data set is quite large and could cause computational or storage issue. Particularly, there is enough data in one year to capture what is happening in the year, and we get new data every two years. Given this, we can have a model that "forgets" about past data.

"Forgetting" about past data allows us to not have a random effect on year, thus allowing the model to be a bit more simple. It can also be more efficient when updating the model, because we have less data to look at. The idea is to use the last year's model as a starting point, and allow the model to change with the new years data.

5.2 What we attempted

Given this kind of model is very particular to our data, we had to create it ourselves from scratch. So, as a disclaimer, we had some issues with our model converging and getting it to run. Though we get results, they are not as reliable as the above results that use a package made by someone else. For this reason, we will only give a model explanation, and provide the code on the project github. We will not use these results to interpret. We do suggest this as a model that could be of better use than the ones above, if we were looking for "real world" usage.

When creating the model, we set it up very similar as the models above. On a singular year, we have a gamma GLM with a random intercept on county. We initialize the first years model with a frequentest gamma regression coefficients and estimated standard deviation. Then we run a Metropolis-Hastings MCMC on the model; this is in contrast with the above models, as they were run with the NUTS algorithm. After the MCMC is done running, we save the results and use them to initialize the next years fitting. We do this for 2013, 2015, and 2017, to test out the model. As mentioned before we run into convergence issues, but have a working model.

5.3 What could be done further

The above was our first attempt on doing a model that can update over the years, but there are more sophisticated ways to do this we did not attempt. We wanted to mention our ideas of how these could be used to better this model here. The first one is to use sequential MC (SMC). Since we have temporal data, the model fits pretty naturally into this SMC. Using this mode of computation could make for quicker convergence, though it does take more effort to set up. As we would need to come up with a way to deal with the extra questions added in later years.

Another thing we could be used is weighting. Since the model forgets about the data used in the previous year, and only used it as a starting value. The model will weight the current years data pretty strongly, meaning the model is just fitting to the current years data. Though this could be good in cases where the model changes a lot year to year, it can be inefficient and noisy if it doesn't. Its likely that year to year, the model doesn't change dramatically (though it might many years later, probably not for adjacent years), so having a model that takes into account how much data was used so far may make for better convergence.

6 Conclusion

We present ideas for further analysis, where using data across the years would be the main idea. Then updating as new data comes in, and we supply a starter code for this on github. However, even though there is a lot that could be done further, we do find reasonable results with the models we use. Particularly, we find that the models seem to capture the distribution fairly well, as the extreme values seem to be captured fairly. Even though it was not super close, we get a decent result given the noisiness of the data. Furthermore, we found that some of the counties are distinctly different from others, with San Francisco and Sand Diego having a lower car risk score on average than the others, and Cleveland having a higher car risk score on average than the others. Lastly, we find that the questions that have to do with drug use, violence, health, and sexual violence tend to be significant. This indicates that there is some kind of association between car riskiness and these types of behaviors. With this information from our model we could use the YRBSS to try to bring awareness to what is associated with car riskiness.

7 Contributions

Andrew: Did part of the data cleaning. Implement gibbs sampler for imputing missing data. Analyze posterior distributions and generate parameter summaries for report. Write introduction, results for report, and did the bibliography.
Brody: Did part of the data cleaning. Created car risk score, and code that goes with it. Did EDA on questions and response. Made scripts to run the model that works across the year, and the single year models. Wrote part of the data overview. Wrote most of the methods. Wrote all of the extensions and conclusions.
James: Finish EDA and apply a more complex random effects baysian modeling with diagnostic plots . Draft the slides. Write the mathematical representation, explanations of the model, and categorical data table.

8 Appendix

All code for this project is available on the following github repository: `https://github.com/brodyee/STATS551_Project`

Variable	Mean	std	50%	IQR
age	4.80	1.29	5.00	2.00
bmi	23.46	5.53	22.22	5.79
grade	2.44	1.11	2.00	2.00
q13	1.09	0.52	1.00	0.00
q18	1.17	0.69	1.00	0.00
q25	1.68	0.47	2.00	1.00
q26	1.84	0.37	2.00	0.00
q47	1.45	1.12	1.00	0.00
q57	1.73	0.44	2.00	1.00
qnowt	1.83	0.37	2.00	0.00
race4	2.64	0.93	3.00	1.00
sex	1.48	0.50	1.00	1.00
q70	2.94	1.67	2.00	2.00
q50	1.12	0.60	1.00	0.00
q15	1.18	0.65	1.00	0.00
q16	1.18	0.83	1.00	0.00
q17	1.54	1.26	1.00	0.00
q67	3.10	0.93	3.00	1.00
q68	2.03	1.13	2.00	2.00
q23	1.85	0.36	2.00	0.00
q51	1.16	0.66	1.00	0.00
q53	1.09	0.54	1.00	0.00
q32	1.12	0.66	1.00	0.00
q69	2.50	1.65	2.00	2.00
q46	2.37	2.04	1.00	4.00
q19	1.91	0.29	2.00	0.00
q12	1.27	0.88	1.00	0.00
q41	1.40	0.94	1.00	0.00
q40	2.88	2.12	2.00	4.00
q27	1.86	0.35	2.00	0.00

	Posterior Mean and 95% High Density Interval		
	Median Impute	Model Impute	No Impute
Term			
Intercept	3.530 [3.184, 3.879]	3.752 [3.408, 4.100]	3.568 [3.206, 3.931]
Age	0.021 [-0.010, 0.053]	0.021 [-0.013, 0.054]	0.068 [0.019, 0.117]
C(Sex)	0.019 [-0.026, 0.065]	0.021 [-0.025, 0.067]	-0.011 [-0.069, 0.048]
Bmi	0.012 [0.007, 0.017]	0.012 [0.007, 0.017]	0.007 [0.001, 0.013]
C(Qnowt)	-0.008 [-0.071, 0.054]	-0.018 [-0.078, 0.042]	0.026 [-0.054, 0.106]
Q13	0.006 [-0.054, 0.069]	-0.027 [-0.090, 0.038]	0.052 [-0.025, 0.135]
Q18	0.031 [-0.022, 0.084]	0.026 [-0.029, 0.083]	0.169 [0.095, 0.246]
Q25	-0.008 [-0.061, 0.045]	-0.015 [-0.068, 0.038]	-0.056 [-0.126, 0.015]
Q26	-0.058 [-0.131, 0.015]	-0.031 [-0.109, 0.046]	-0.148 [-0.241, -0.056]
Q47	0.060 [0.035, 0.085]	0.051 [0.026, 0.076]	0.205 [0.172, 0.238]
Q57	-0.109 [-0.160, -0.058]	-0.116 [-0.168, -0.065]	-0.196 [-0.265, -0.127]
Q70	-0.027 [-0.041, -0.013]	-0.028 [-0.041, -0.014]	Not Included
Q50	0.028 [-0.033, 0.092]	-0.016 [-0.077, 0.047]	Not Included
Q15	0.071 [0.028, 0.114]	0.063 [0.020, 0.106]	Not Included
Q16	0.088 [0.045, 0.132]	0.069 [0.026, 0.112]	Not Included
Q17	0.057 [0.030, 0.084]	0.056 [0.028, 0.084]	Not Included
Q67	-0.098 [-0.125, -0.070]	-0.097 [-0.125, -0.068]	Not Included
Q68	-0.045 [-0.066, -0.025]	-0.040 [-0.061, -0.020]	Not Included
Q23	-0.046 [-0.113, 0.018]	-0.046 [-0.112, 0.020]	Not Included
Q51	0.063 [0.015, 0.114]	0.045 [-0.002, 0.094]	Not Included
Q53	0.057 [-0.019, 0.137]	-0.003 [-0.075, 0.073]	Not Included
Q32	0.032 [-0.016, 0.083]	0.015 [-0.032, 0.065]	Not Included
Q69	0.015 [0.000, 0.030]	0.014 [-0.001, 0.028]	Not Included
Q46	0.046 [0.033, 0.059]	0.039 [0.027, 0.052]	Not Included
Q19	-0.128 [-0.210, -0.046]	-0.118 [-0.200, -0.036]	Not Included
Q12	0.060 [0.030, 0.091]	0.063 [0.031, 0.094]	Not Included
Q41	0.154 [0.121, 0.187]	0.157 [0.123, 0.190]	Not Included
Q40	0.049 [0.036, 0.061]	0.047 [0.035, 0.058]	Not Included
Q27	0.000 [-0.082, 0.081]	-0.024 [-0.107, 0.057]	Not Included
1—Sitecode_Sigma	0.187 [0.114, 0.321]	0.222 [0.135, 0.381]	0.206 [0.120, 0.365]
Carriskscore_Alpha	0.435 [0.428, 0.442]	0.435 [0.428, 0.442]	0.410 [0.401, 0.418]

References

[1] Peter Hoff. "A First Course in Bayesian Statistical Methods". In: (), p. 118.

[2] "Leading Causes of Death". In: (). URL: https://www.cdc.gov/injury/wisqars/animated-leading-causes.html.

[3] "YRBS National, State, and District Combined Datasets User's Guide". In: (). URL: https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2019/2019_YRBS_SADC_Documentation.pdf.