1 Introduction and Data Description

Central nervous system stimulants, such as amphetamines, have seen widespread use over the past century, for legal and illegal purposes. A substantial portion of amphetamine transactions are carried out in the "black market", hence it is important to study their spatial and temporal trends, both for public health and law enforcement purposes. In this case study, I investigated factors that affect the price (per mg) of the drug amphetamines based on a dataset taken from the website StreetRx.com, which gathers user-submitted information on street prices of diverted prescription or illicit drugs.

The dataset for amphetamines consists of 76065 observations with 13 variables. I removed entries with NA values (58 rows), and split the date variable into year and month. The variables api_temp, form_temp and country has only one level for amphetamines and hence are discarded. The variable city has about 8000 levels, which is not relevant/interesting for the purposes of this study without further geographic information linking them to states and regions. Hence I also removed the city column. For certain levels within the variables source, state and year, the number of entries are too small (<10) to be studied in a satisfactory manner, hence I also dropped those levels. Since the distribution of ppm is extremely positively skewed, I removed outliers that are three standard deviations away from the mean after normalization (this corresponds to a 99.7 percent interval) and performed a log-transformation on the value of ppm. The cleaned dataset consists of 75207 observations with 11 columns: both the original and log-transformed ppm(Price per mg), state, country, USA_region, source, mgstr, bulk_purchase, Primary_Reason, month and year.

2 Exploratory Data Analysis

I explored the dataset by first looking at how various grouping structures might show different height patterns. Figure 1 shows the box plots of price per mg of amphetamines vs. region (a), state (b), bulk_purchased or not (c), dosage strength (d), source (e), primary reason (f), year (g) and month (h). I observed that geographic factors like region and state show different mean prices and variances that are heterogeneous. In particular, the variability with Alaska, American Samoa and Vermont have relatively higher and larger ranges than other states, which might potentially be related to their lower population density and smaller sample sizes. I also saw a small difference in mean prices and variability for bulk purchases vs non-bulk purchases. I observed that dosage strength, while a numerical covariate, takes only a fixed number of values, so for exploration purposes I visualize it with the side-by-side box plot as shown, which seem to demonstrate a slightly negative correlation (with a lot more dispersion at higher dosages). I also observed higher variability for primary reasons of purchase 4 and 12, which could be related to their smaller number of entries. Last but not least, an inspection of the year by year box plot indicated that in the earlier years (before 2014), which again have much smaller samples, there are substantially more variability than the latter years. I did not observe any substantial patterns in the mean or variance across months.

3 Modeling

I noted that the response of interest, price per milligram (ppm), is strictly non-negative and also has drastically different values, ranging from 10^{-1} to 60. I therefore adopted a log-transformation of the ppm response, which improves the normality of the response variable and therefore should theoretically allow for better fit of the Bayesian hierarchical normal. I chose to implement such a Bayesian model because I would like to have a coherent way to account for uncertainty. The only continuous covariate in the data set is mgstr, which I modelled as a fixed effect. There are multiple groupings that I took into consideration: states, sources, reason of purchase, month and year. There is also a binary bulk purchase indicator in the dataset, which I incorporated in the model using an additional intercept and an additional variance effect.
I had the following model:

$$\begin{align}
\log(\text{ppm}_{ijkmpq}) &\sim N(\mu + \beta_1 \text{mgstr}_i + \text{state}_j + \text{source}_k + \text{reason}_m + \text{month}_p + \text{year}_q \tag{1}\\
&\quad + \beta_2 I_{\text{i bulk purchased}}, \sigma^2 + \gamma^2 I_{\text{i bulk purchased}}) \tag{2}\\
\text{state}|\Omega_{state} &\sim MVN(\mathbf{0}, \Omega_{state}) \tag{3}\\
\text{source}|\Omega_{source} &\sim MVN(\mathbf{0}, \Omega_{source}) \tag{4}\\
\text{reason}|\Omega_{reason} &\sim MVN(\mathbf{0}, \Omega_{reason}) \tag{5}\\
\text{month}|\Omega_{month} &\sim MVN(\mathbf{0}, \Omega_{month}) \tag{6}\\
\text{year}|\Omega_{year} &\sim MVN(\mathbf{0}, \Omega_{year}) \tag{7}\\
\Omega_{state} &\sim \text{InvWishart}(\omega_{0,state}, \Omega_{0,state}) \tag{8}\\
\Omega_{source} &\sim \text{InvWishart}(\omega_{0,source}, \Omega_{0,source}) \tag{9}\\
\Omega_{reason} &\sim \text{InvWishart}(\omega_{0,reason}, \Omega_{0,reason}) \tag{10}\\
\Omega_{month} &\sim \text{InvWishart}(\omega_{0,month}, \Omega_{0,month}) \tag{11}\\
\Omega_{year} &\sim \text{InvWishart}(\omega_{0,year}, \Omega_{0,year}) \tag{12}\\
1/\sigma^2 &\sim \text{gamma}(\omega_{0,overall}/2, \omega_{0,overall}\sigma_0^2/2) \tag{13}\\
1/\gamma^2 &\sim \text{gamma}(\omega_{0,bulk}/2, \omega_{0,bulk}\gamma_0^2/2) \tag{14}\\
\beta_1 &\sim N(0, 1/\tau_\beta^2) \tag{15}\\
\beta_2 &\sim N(0, 1/\tau_\beta^2) \tag{16}\\
\mu &\sim N(\mu_0, \delta_0^2) \tag{17}
\end{align}$$

After the data cleaning procedure, I still have more than 75000 data points in the dataset, with many subgroups having hundreds/thousands of data points. I therefore included almost all of the available grouping information as crossed random effects. Due to the large number of available data, the results of the sampling should in theory be similar to that of a frequentist mixed effects model.

For the response, I fitted the log-transformed ppm according to a normal model. Each of the random effects are assumed to follow a multivariate normal distribution that is zero mean-ed with a covariance matrix that is in turn assumed to follow the conjugate Inverse Wishart priors with an empirical scale matrix (this approach is justified by Raff's work, among others) and a minimally informative choice of the degree of freedom parameter (equal to the dimension of the scale matrix). This setup allows us to account for heterogeneous variances across subgroups while enforcing our prior assumption that the subgroups are independent of each other. For the fixed effect coefficients, I weakly informative prior where the precision is set to be low (0.0001). For the overall variance, which is a scalar, I adopt the typical inverse gamma assumption, again with weakly informative hyper parameter choices of 0.001.
I run the above model using Gibbs Sampling, implemented through JAGS in R. Results are shown below.

4 Model Results and Interpretation

Assessment of autocorrelation and effective sample size show no evidence of non-convergence. I shall proceed with further analysis with caution assuming that the results have converged.

Our results are shown in Figure 2 and Figure 3a,b. I remark that overall, the results of the bayesian analysis closely resembles that of the empirical data analysis. I observe that states such as American Samoa, Vermont and "Unknown" (user did not enter state), which have high variance in the EDA, also exhibited high variance in the results. Similarly, I observed that in the earlier years, both the EDA and the model results showed higher variances when compared to the latter years, which are much more uniform. For month subgroups, both EDA and model results show a constant pattern. For the reason of purchase subgroup, I observe that reason 12, which has a very small sample size, has high variances in both the EDA and the model. The above random effect analysis suggest that the EDA and the modeling shared similar patterns.

As for the fixed effects, I observe that the values, after exponentiation, closely matches that of the empirical mean. For example, the posterior mean of μ is around 0.5, and the empirical mean of ppm is around 0.47.

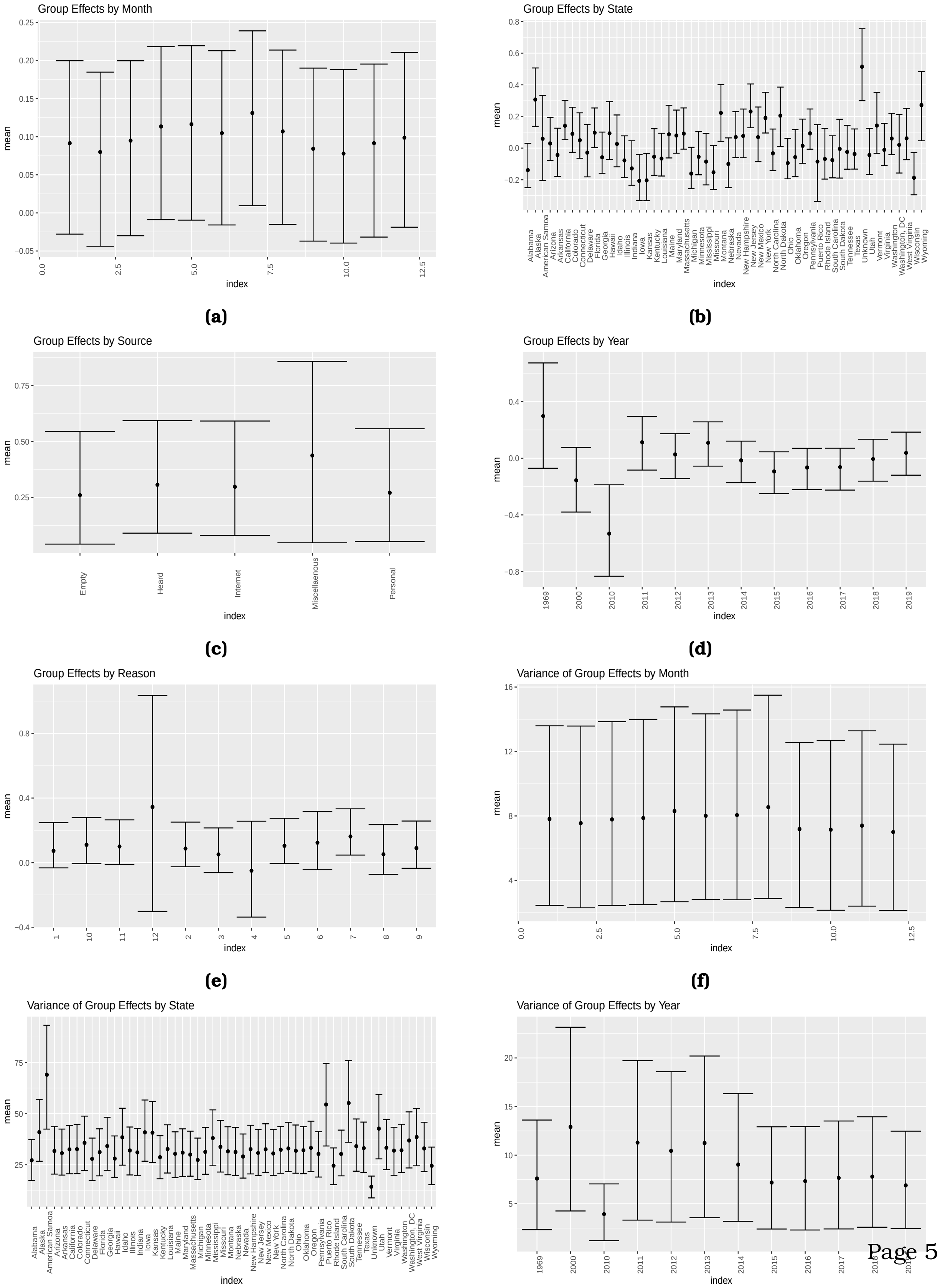
Overall, I can conclude that geographical effects such as states contribute substantially to the variation of price of amphetamines, while months and years (as long as it is after 2011) do not appear to show as much variation.

5 Model Validation

I conduct a posterior predictive check of the model (Figure 3c,d), and the resulting comparison of the histograms (3 simulated datasets) and the empirical CDF's indicate that our model closely capture the mean of the data. However, I see that our model did not account for all the variance observed in the data and was unable to capture the shape of the data. A closer analysis reveals that while the ppm/logppm values displayed a roughly symmetric bell shape curve under a coarse histogram analysis, when I plot them under finer scales (Figure 3e,f), the density of the ppm/logppm no longer resembles a bell curve. To the contrary, they exhibited multiple sharp peaks. This multimodality, combined with the highly dispersed nature of the data likely contributed to the perceived discrepancies between the actual and the simulated data. Since I am utilizing a hierarchical normal model, these multimodality and large dispersion are more difficult to capture. In conclusion, our model fitted the means very well but due to the limitation of the distributional assumptions as well as the multimodality of the data, I should treat any interpretation of the variances of the model with a grain of salt.

6 Conclusions and Future Directions

I examine the variability of price per mg of amphetamines across different states, sources, reasons of purchase. The results from our Bayesian hierarchical normal model matches that of the exploratory data analysis closely. I are able to draw the following general conclusions: 1) of all the random effects groups considered, geographical location at the state level showed substantial contribution to the variation of price, where temporal factors (month, year) and personal factors (reason of purchase, source) showed less substantial variation (except for several unique exceptions). 2) posterior predictive model checking suggests that our model fitted the means well, but was unable to fully capture the variation present in the data and the full distributional shape of the response variable. This is likely due to the multimodality as well as the skewed and highly dispersed nature of the dataset, which a simple hierarchical normal model was not able to fully capture. I have tried various transformations of the data such as the log, but for a fully satisfactory treatment, one would probably have to consider nonparametric models or multiple mixtures of normals.



(a)

(b)

(c)

(d)

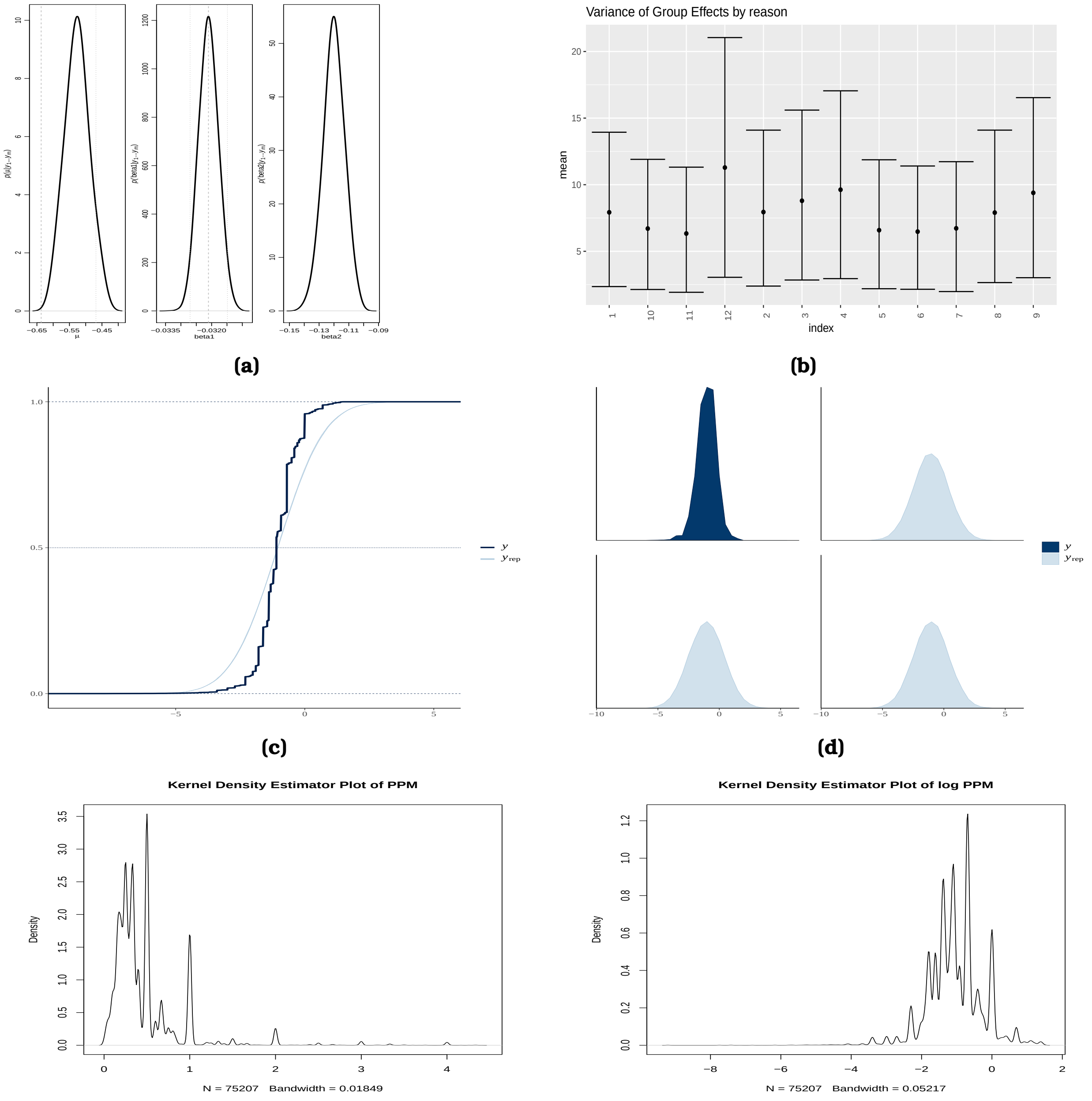
(e)

(f)

(g)

(h)

Figure 1: Model Results Plots



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 2: Model Results and Diagnosis Plots