Data 102 Final Project Report

Marcel Angelo Daniel Gunadi
marceladg@berkeley.edu

Charlie Kortleven
charliekortleven@berkeley.edu

Parth Patel
parth773@berkeley.edu

Nicholas Toar Turangan
nikot@berkeley.edu

Group 31

December 2025

1 Data Overview

For this project, we used the Chronic Disease and Air Quality dataset from the CDC, which was graciously provided by the course staff. This data represents a census and, to our knowledge, does not systematically exclude groups of pepole. Each row of our data represents a topic (e.g., asthma, tobacco use), year, the state where the data were collected, and reports results for measures of health. For cleaning and filtering, we first assigned each State to a region based on the US Census Bureau's nine divisions. We did not include any U.S. territories such as Guam or Puerto Rico in our data. Thus our resulting data represents all 50 US states, the District of Columbia. From this we derived a United States average. Afterward, we filtered by topics, specifically to include: crude smoking and asthma prevalence. During our EDA, we found that Florida and New Jersey were missing data for 2021 and 2019, respectively. Because our research focuses on each state individually, we elected to ignore those missing values. We did not use any supplemental data beyond our CDC dataset.

2 Research Questions

The two research questions we will be answering are

1. **Which states show a statistically significant linear trend in asthma prevalence over time (from 2011 - 2021)?**
   We use multiple hypothesis testing to answer this question. This research question has the potential to influence each state's asthma policy, provide insight into ways to reduce it, and identify other factors that may be at play. For this question, we will

be testing 51 null hypotheses, one for each US state and the District of Columbia. $H_0(i) : B_1(i) = 0$ For state i Asthma Prevalence $= \beta_0 + \beta_1(Year) + \epsilon$.
We assess the significance of trends using two-sided p-values for the year coefficient from the regression outputs, testing the null hypothesis that no trend exists (at a p-value threshold of 0.05) across the 50 states and DC. We want to test for association at the state level because, by disaggregating the data, we can partition associations that may not appear when aggregated at the national level. Furthermore, to generate meaningful results that can inform policy, we need to understand where asthma prevalence is worsening or improving over time. One limitation we have is that the range might be too small and can lessen our statistical power.

2. **How do state-level asthma prevalences in adults differ after partial pooling, and how is it associated with adult smoking prevalence?** For this research question, we will be using Bayesian hierarchical Modeling with partial pooling. The real-world decisions we can make from it are that we can make targeted health campaigns in states with high prevalence and smoking rates, which can potentially lead to lower asthma rates. This is a good fit for this question, as we can measure how smoking prevalence influences asthma prevalence. One limitation of this method is that it relies heavily on prior distributions. If chosen distributions are incorrect, we cannot reliably interpret our results.

3 Prior Work

Trends in adult current asthma prevalence and risk factors in the US by state: 2000–2009. Zhang et al. examine asthma trends in the US with a particular interest in the various factors that may influence these trends, such as sociodemographic characteristics, smoking status, and weight status. They use a logistic regression models to determine if there is a significant linear change in current asthma prevalence over time for each state. They found that between 2000 and 2009, there was a significant increase in state-specific asthma prevalence, with obesity being a significant determinant. However, this paper only examines asthma prevalances between 2000-2009. Given our access to a dataset that extends beyond this timeframe, we want to know whether these increases will continue through the next decade and whether smoking status impacts asthma prevalence. This is directly relavent to our first reasearch question as we try to use a regression model to see if there are any linear trends in asthma prevalnce for 2010-2021.

Percentage of People with Asthma who Smoke
This CDC report, based on 2010 BRFSS data, details the percentage of individuals with asthma who smoke, as well as the percentage of individuals without asthma who smoke. This report suggests that the percentage of people with asthma who smoke varies significantly from state to state, and gives a US total average as well. This report directly motivates our second research question by documenting state-level variation in smoking prevalence among adults with asthma. One key difference, however, is that this study is only descriptive, whereas our research uses Bayesian hierarchical modeling to examine associations between

the two.

4 EDA

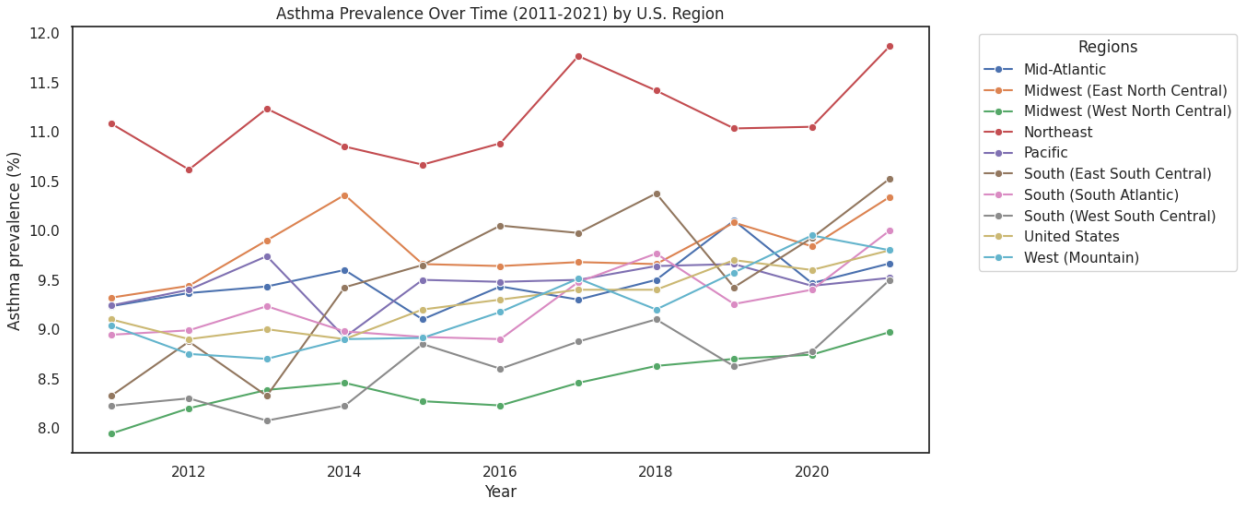


Figure 1: Asthma Prevalence Over Years by U.S. Region.

In this figure, we plotted asthma prevalence over time for 10 years (2011-2021) in the U.S. Prevalence rates are aggregated by region, and each region is defined by the United States Census Bureau. We additionally plotted the national average ("United States").

We observe that across all regions in the U.S., asthma prevalence is increasing. The Northeast region, in particular, has the highest prevalence, while the Mountain and the West South Central regions have the lowest. The remaining regions show greater variability in their asthma prevalence over time.

The visualization is relevant to our research question, which asks whether and which regions and states show a linear trend in their association between asthma and smoking prevalence. Results from this EDA suggest that on the regional scale, there are some linear trends in asthma. This motivates and justifies our later research, which will identify which states in these regions also exhibit linear associations between asthma and smoking over time.

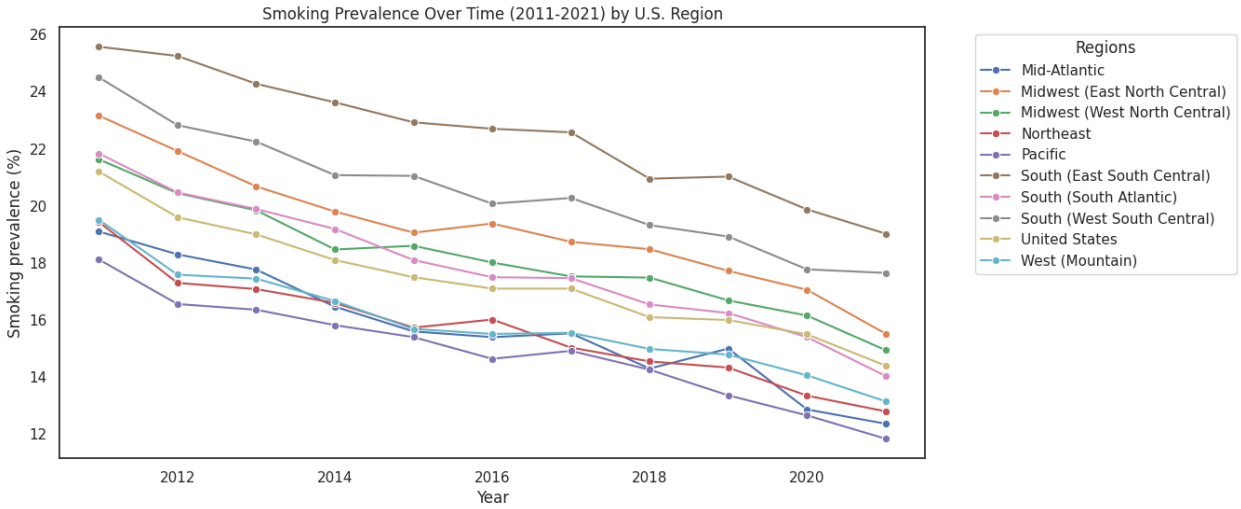


Figure 2: Smoking Prevalence Over Years by U.S. Region.

In this figure we visualize the trends of smoking prevalence from 2011-2021 aggregated by region and the United States national average. We can see that all regions have a decreasing linear trend. However there is variability in this decrease for instance some regions such as the Mid Atlantic and the Northeast show trends that increase and decrease for smoking prevalence. It is unclear which states within these regions might be responsible for this phenomenon. This is why our research will show which states have linear relationships. Looking at figure one we can see that even though smoking prevalence is decreasing asthma prevalence keeps increasing. However if you take a look at the Mid Atlantic in 2019 the upward trend in asthma prevalence corresponds with a spike in the smoking prevalence graph suggesting that there is some correlation between the two variables.

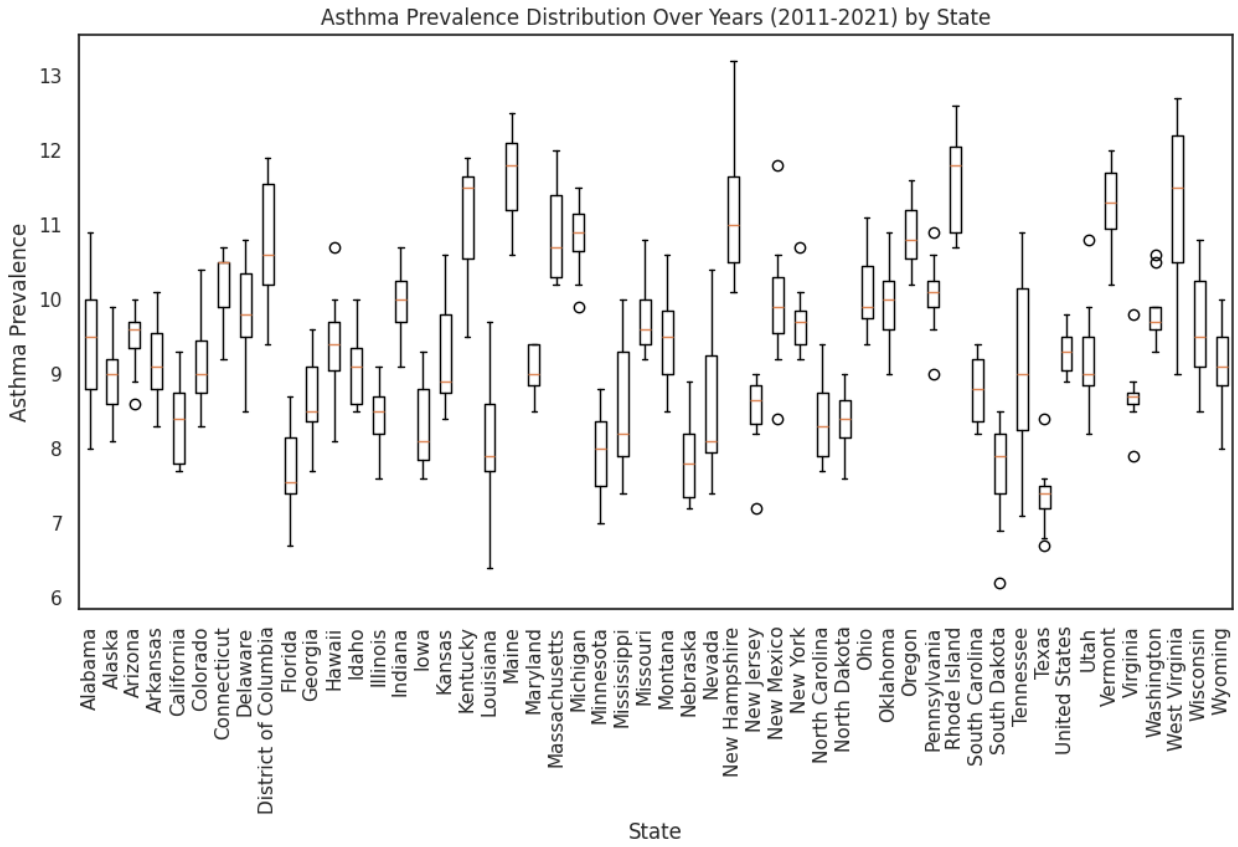


Figure 3: Asthma Prevalence Distribution Over Years by State

In this visualization, we plot asthma prevalence by state from 2011-2021. We observe substantial variation in asthma prevalence across states, and that some states have much higher asthma prevalence (e.g., District of Columbia, New Hampshire) while other states have much lower prevalence (e.g., Florida, Louisiana, and Texas). We also observe a loose association between state population size and prevalence, where smaller populations tend to show more extreme (either higher or lower) prevalence. We hypothesize that these estimates are more extreme because there are fewer people/data points, and thus estimates are more easily skewed. Although the prevalence accounts for population size (and is mathematically defined as: #positive cases/#people), estimates in very small populations are more likely to appear noisy and thus less reliable.

These observations from our EDA are directly related to our research question, which asks whether state-level asthma prevalences in adults differ after partial pooling. This visualization motivates the need for partial pooling (regional-level), because small-population states may have unreliable prevalence estimates. As a result, we will specifically investigate whether partial pooling shrinks extreme estimates in low-population states and can provide a means of generating more reliable estimates.

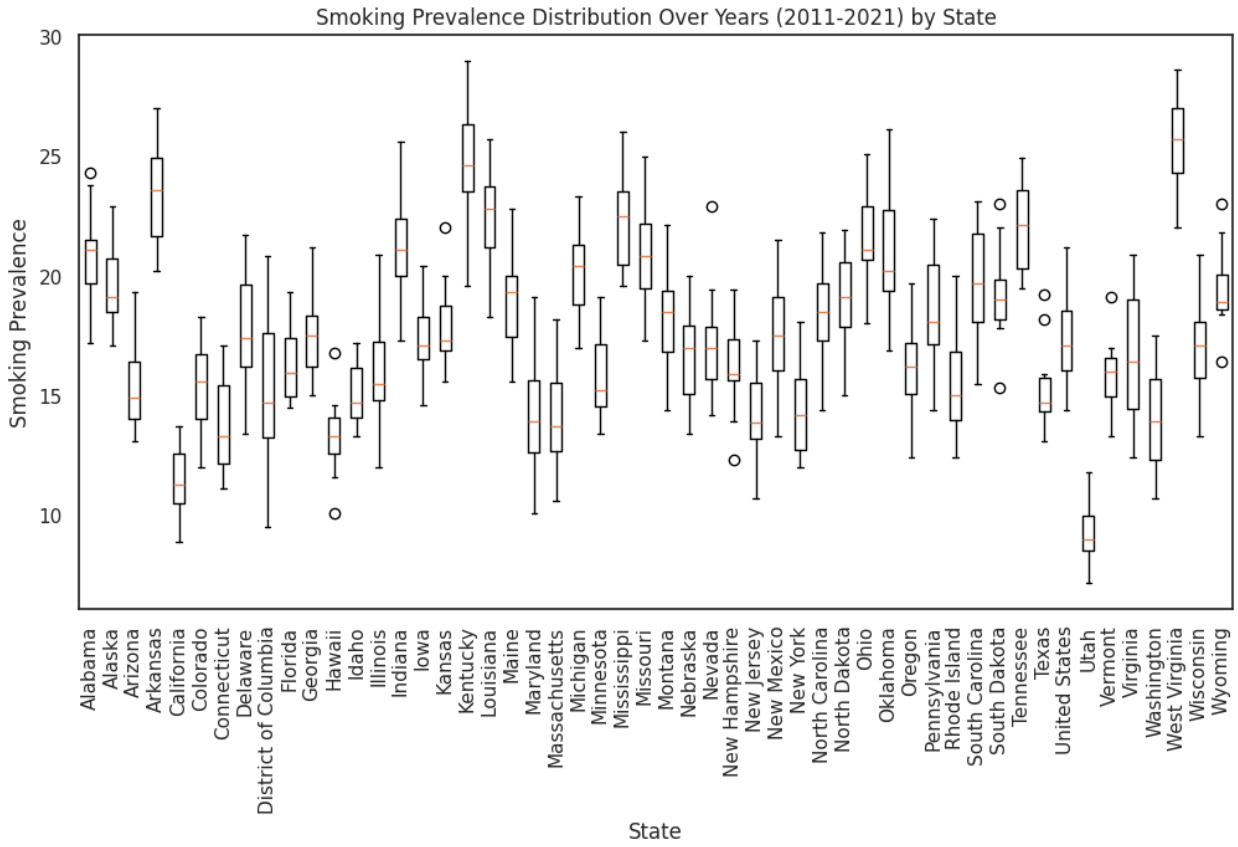


Figure 4: Smoking Prevalence Over Years by State

In this visualization, we plot smoking prevalence by state from 2011-2021. We observe similar variability in state-level prevalence estimates as observed in asthma prevalences (Figure 3-4). When comparing the two figures jointly, there is an approximate positive association in states between smoking and asthma. However, we are not claiming that this association is significant, solely that we observe a general visual trend. Furthermore, smoking prevalence is generally higher (10-30%) in all states compared to asthma prevalence (6-13%) (Figure 3-4). Notably, Maine has the highest asthma prevalence out of all states (Figure 3), but has moderate smoking levels (Figure 4), suggesting that there may be additional factors (e.g, societal norms, environmental factors, lifestyle choices) that influence more "extreme" prevalence estimates.

These observations directly relate to the second part of our research question, which asks what is the association between adult smoking prevalence and asthma. After applying partial pooling, we will re-examine this relationship and determine which regions reveal the strongest positive association between smoking and asthma prevalence. We are interested in identifying regions where smoking is closely linked to asthma, such that our research can help target regions where policy (e.g., public health intervention) could be most beneficial.

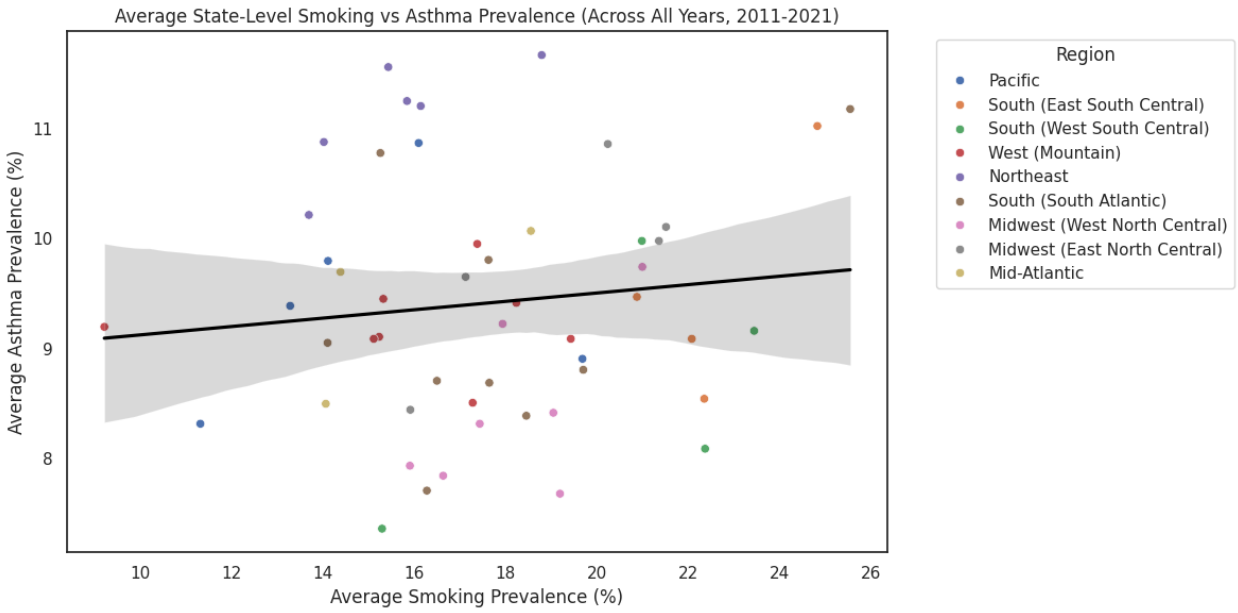


Figure 5: Average State-Level Smoking vs Asthma Prevalence (Across all Years)

In this figure, we observe a slight positive correlation between average state-level smoking prevalence and average state-level asthma prevalence, suggesting a potential link worth further exploration. However, this relationship is weak and shows substantial scatter across the states. The fitted regression line is sloped upwards, indicating that states with higher smoking rates tend to have slightly higher asthma prevalence. This figure supports our research questions by showing that smoking may be associated with asthma prevalence. However, the association is weak, so that's why in our research we are using partial pooling to get rid of some of the state level noise.

5 Research Question 1

Research Question Which states show a statistically significant linear trend in asthma prevalence over time (from 2011 - 2021)?

5.1 Overview

The aim of this study is to test whether and which states show a statistically significant linear trend in asthma prevalence over time (from 2011 - 2021). Our null hypothesis is that there is no significant linear association ($\beta_1 = 0$) over time in asthma prevalence in any given state (i) (including DoC). Our alternative hypothesis is that there is a linear association ($\beta_1 \neq 0$) over time in asthma prevalence in any given state (i) (including DoC). Where β_1 is our year coefficient

We will be using the β_1 slope coefficient as our test statistic. We want to test for association at the state level because, by disaggregating the data, we can partition associations that may not appear when aggregated at the national level. Furthermore, to generate meaningful results that can inform policy, we need to understand where asthma prevalence is worsening or improving over time.

5.2 Methods

5.2.1 Hypothses

We will be testing all 50 states and the District of Columbia (DoC) in our hypothesis testing. One of our alternative hypotheses is for the state of California: H_Ai=California, there is a significant linear association ($\beta_1 \neq 0$) between time in years and asthma prevalence in the state of California.

During our exploratory data analysis, we found that many states had subtle increases in asthma prevalence over time. Because we have noticed that public health studies typically don't use very high correlation values, we decided on 0.5. These values (per state) were estimated from the slope of our simple linear regression model. We are using β_1 as our value because we want to know whether there is significance in either direction, rather than just a linear trend, and our coefficient tells us that.

5.2.2 Power Calculation

To calculate the power of our tests, we used a Monte Carlo simulation as described in Estimating Power with Monte Carlo Methods. We estimate distribution parameters from the data and generate a set of observations for each sample. Then we calculate the p-value of our test and compare it to our alpha. If we reject the null, we add one to our success and repeat until we reach the end of our trials. Then our power is the proportion of samples that rejected the null hypothesis, divided by the number of trials. We use Monte Carlo simulation to calculate the power of our tests because we have a small sample size of only 11 yearly observations, which can make other analytical methods less reliable. Plus, this way it captures individual trends for each state. Another advantage of using Monte Carlo simulation is that it mitigates errors when our data are autocorrelated. We ended up

calculating the power for each state. For our purposes, we took 1000 random samples, used the regression coefficients, and sampled the error from a normal distribution with mean 0 and standard error equal to the standard error of each state. Then we fit another regression model on the simulated y and calculated the p-value from that model, and then compared it to the significance level of .05. If it was less than .05, we added one to the rejection counter and finally computed power by dividing the number of rejections by the number of iterations(1000).

To test our hypothesis, we fit a simple ordinary least squares linear regression model to each state, with asthma prevalence as the dependent variable and year as the independent variable.

$H_0(i) : B_1(i) \neq 0$ For each state i Asthma Prevalence $= \beta_0 + \beta_1 * (Year) + \varepsilon$

We selected the linear regression model because it is the gold standard of quantifying linear trends in data. We reject the null if our p-value for each test i is < 0.05.

5.2.3 Error Correction Methods

We employed three different methods to control error rates:

(A) Bonferroni Correction. Family-wise error rate (FWER) is being controlled. FWER = alpha/m, where alpha is the false positive rate (0.05), and m is the number of hypotheses being tested (51).

(B) Benjamini-Hochberg, for when hypotheses are independent. The false discovery rate (FDR) is being controlled here and was set to 0.05.

(C) Benjamini–Yekutieli: for when hypotheses are dependent. The false discovery rate (FDR) is controlled here and set to 0.05.

5.3 Assumptions

Linear regression makes six assumptions about the data:

1. The relationship between the independent and dependent variables is linear.
2. The variance of residuals remains constant across all levels of the independent variables, i.e., Homoscedasticity of Residuals.
3. The residuals follow a normal distribution when multiple predictors are involved.
4. Residuals must not correlate with each other across observations, i.e., independence of our errors.
5. Lack of Multicollinearity
6. Absence of Endogeneity

To confirm that our data is linear, we examined residual plots for all states and found no patterns that would indicate linearity. We have a couple below.

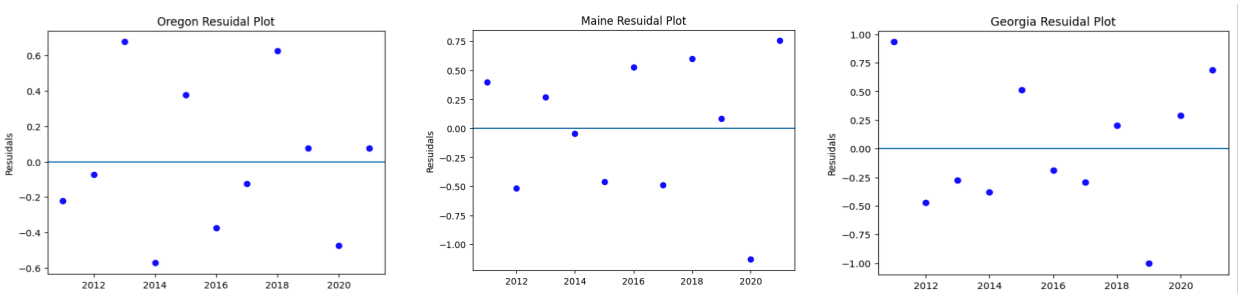


Figure 6: Resudial Plots

To confirm that homoscedasticity was not violated, we used the Breusch–Pagan test on the residuals, which tests whether the variance of errors from a model depends on our Year variable. If a state score is above .05 on the Breusch–Pagan test, the data are not homoscedastic. We found that all states follow homoscedasticity.

The Normal Distribution and Multicollinearity are satisfied by only having one variable to predict from. We used the Durbin–Watson test to confirm the independence of our errors since we are working with a time series. The absence of Endogeneity is implicit, since year is an exogenous variable.

For the correction algorithms, we used only the Benjamini–Hochberg procedure, which assumes that all tests are independent, as we will discuss later.

Our Monte Carlo simulation makes the same assumptions as our linear regression, as discussed above. Also, it assumes that the residuals follow the same distribution as our original data.

5.4 Results

Before correcting for FDR or FWER, we found that 18 states reject the null hypothesis: Alabama, Colorado, Connecticut, Idaho, Kansas, Louisiana, Michigan, Minnesota, Mississippi, Nebraska, Nevada, Oklahoma, Pennsylvania, South Carolina, Tennessee, Utah, and West Virginia.

Alabama, California, Hawaii, Illinois, Indiana, Kansas, Maine, Nevada, New Jersey, North Dakota, Ohio, Oregon, South Carolina, and Utah all failed the Durbin-Watson Test, which is around 27% of our data, indicating autocorrelation and residual correlations across observations. However, we proceed with our test regardless, even though it violates an assumption of linear regression, since it's only a small portion of the dataset, all violations are at most 1 point above or below the threshold, and we are performing error Correction on the entire test, which should account for any inflated false positives.

Here are all the results from our tests.

	State	Beta0	Beta1	Std_error	Pvalue	sigma	R^2	N_years	Pass_Thres	trend_der	Power	Bon_.05	BH_.05	BY_.05
0	Alabama	-322.250909	0.164545	0.069946	0.043131	0.695956	0.380765	11	True	Inc	0.621	False	False	False
1	Alaska	-115.716364	0.061818	0.048500	0.234373	0.482569	0.152910	11	False	Inc	0.241	False	False	False
2	Arizona	-104.174545	0.056364	0.036207	0.153963	0.360252	0.212141	11	False	Inc	0.316	False	False	False
3	Arkansas	-49.483636	0.029091	0.054978	0.609512	0.547025	0.030171	11	False	Inc	0.084	False	False	False
4	California	-44.830909	0.026364	0.054122	0.637822	0.538508	0.025687	11	False	Inc	0.070	False	False	False
5	Colorado	-320.781818	0.163636	0.028844	0.000304	0.286991	0.781476	11	True	Inc	1.000	True	True	True
6	Connecticut	-178.552727	0.093636	0.034111	0.022659	0.339398	0.455713	11	True	Inc	0.744	False	False	False
7	Delaware	-85.492727	0.047273	0.072358	0.529885	0.719949	0.045278	11	False	Inc	0.082	False	False	False
8	District of Columbia	-88.185455	0.049091	0.083915	0.572914	0.834941	0.036633	11	False	Inc	0.077	False	False	False
9	Florida	111.538788	-0.051515	0.065944	0.457178	0.564708	0.070877	10	False	Dec	0.110	False	False	False
10	Georgia	-0.472727	0.004545	0.057795	0.939033	0.575049	0.000687	11	False	Inc	0.053	False	False	False
11	Hawaii	134.016364	-0.061818	0.065046	0.366746	0.647204	0.091203	11	False	Dec	0.148	False	False	False
12	Idaho	-192.509091	0.100000	0.039811	0.033210	0.396118	0.412125	11	True	Inc	0.640	False	False	False
13	Illinois	-61.198182	0.034545	0.040151	0.411911	0.399500	0.076000	11	False	Inc	0.128	False	False	False
14	Indiana	-34.003636	0.021818	0.043392	0.627164	0.431741	0.027324	11	False	Inc	0.086	False	False	False
15	Iowa	-158.460000	0.082727	0.053594	0.157080	0.533249	0.209328	11	False	Inc	0.320	False	False	False
16	Kansas	-360.983636	0.183636	0.031978	0.000279	0.318177	0.785599	11	True	Inc	1.000	True	True	True
17	Kentucky	-100.769091	0.055455	0.087522	0.542103	0.870836	0.042701	11	False	Inc	0.088	False	False	False
18	Louisiana	-393.276364	0.199091	0.059705	0.008735	0.594054	0.552673	11	True	Inc	0.887	False	False	False
19	Maine	-17.650909	0.014545	0.058935	0.810595	0.586391	0.006723	11	False	Inc	0.057	False	False	False
20	Maryland	-84.414545	0.046364	0.031954	0.180733	0.317934	0.189577	11	False	Inc	0.273	False	False	False
21	Massachusetts	31.041818	-0.010000	0.064553	0.880309	0.642290	0.002659	11	False	Dec	0.049	False	False	False
22	Michigan	-172.409091	0.090909	0.040156	0.049857	0.399545	0.362845	11	True	Inc	0.578	False	False	False
23	Minnesota	-221.154545	0.113636	0.043403	0.027895	0.431856	0.432346	11	True	Inc	0.705	False	False	False
24	Mississippi	-447.803636	0.226364	0.055330	0.002712	0.550529	0.650315	11	True	Inc	0.976	False	True	False
25	Missouri	125.207273	-0.057273	0.047210	0.255947	0.469729	0.140545	11	False	Dec	0.204	False	False	False
26	Montana	-221.505455	0.114545	0.052542	0.057169	0.522790	0.345581	11	False	Inc	0.526	False	False	False
27	Nebraska	-204.750909	0.105455	0.040252	0.027817	0.400500	0.432669	11	True	Inc	0.712	False	False	False
28	Nevada	-374.530909	0.190000	0.070482	0.024566	0.701291	0.446727	11	True	Inc	0.726	False	False	False
29	New Hampshire	-289.358182	0.149091	0.079798	0.094545	0.793977	0.279468	11	False	Inc	0.439	False	False	False
30	New Jersey	16.554745	-0.003996	0.056743	0.945586	0.535247	0.000620	10	False	Dec	0.045	False	False	False
31	New Mexico	-78.016364	0.043636	0.087026	0.628113	0.865899	0.027176	11	False	Inc	0.072	False	False	False
32	New York	42.689091	-0.016364	0.043000	0.712366	0.427849	0.015836	11	False	Dec	0.064	False	False	False
33	North Carolina	-61.252727	0.034545	0.056656	0.557113	0.563721	0.039670	11	False	Inc	0.083	False	False	False
34	North Dakota	-55.727273	0.031818	0.040763	0.455096	0.405586	0.063406	11	False	Inc	0.109	False	False	False
35	Ohio	-24.712727	0.017273	0.052187	0.748238	0.519256	0.012025	11	False	Inc	0.066	False	False	False
36	Oklahoma	-242.934545	0.125455	0.039506	0.011264	0.393076	0.528413	11	True	Inc	0.859	False	False	False
37	Oregon	-89.927273	0.050000	0.042245	0.266904	0.420335	0.134684	11	False	Inc	0.193	False	False	False
38	Pennsylvania	-193.360000	0.100909	0.037359	0.024351	0.371716	0.447711	11	True	Inc	0.743	False	False	False
39	Rhode Island	-177.207273	0.093636	0.060811	0.157996	0.605062	0.208510	11	False	Inc	0.300	False	False	False
40	South Carolina	-236.776364	0.121818	0.025361	0.000969	0.252335	0.719391	11	True	Inc	0.996	True	True	False
41	South Dakota	-192.085455	0.099091	0.061761	0.143083	0.614514	0.222407	11	False	Inc	0.321	False	False	False
42	Tennessee	-610.370909	0.307273	0.082531	0.004748	0.821174	0.606326	11	True	Inc	0.947	False	True	False
43	Texas	-137.421818	0.071818	0.038377	0.094093	0.381850	0.280117	11	False	Inc	0.433	False	False	False
44	Utah	-282.203636	0.144545	0.052046	0.021497	0.517854	0.461501	11	True	Inc	0.739	False	False	False
45	Vermont	-102.374545	0.056364	0.049826	0.287209	0.495764	0.124482	11	False	Inc	0.179	False	False	False
46	Virginia	-104.920000	0.056364	0.040447	0.196919	0.402447	0.177470	11	False	Inc	0.258	False	False	False
47	Washington	-89.167273	0.049091	0.038071	0.229390	0.378802	0.155936	11	False	Inc	0.238	False	False	False
48	West Virginia	-635.770909	0.320909	0.070125	0.001335	0.697730	0.699422	11	True	Inc	0.994	False	True	False
49	Wisconsin	-195.610909	0.101818	0.070115	0.180409	0.697632	0.189831	11	False	Inc	0.270	False	False	False
50	Wyoming	-176.014545	0.091818	0.051067	0.105724	0.508107	0.264275	11	False	Inc	0.407	False	False	False

Figure 7: Testing Results

5.5 Discussion

After Bonneferri correction, only Colorado, Kansas, and South Carolina remained significant. After Benjamini-Hochberg correction, Colorado, Kansas, Mississippi, South Carolina, Tennessee, and West Virginia were significant. Finally, after Benjamini–Yekutieli correction, we only found Colorado and Kansas to have significant associations between time (2011-2021) and asthma prevalence. In both states, asthma prevalence increased over time. All other states were statistically insignificant, but only the following showed some negative correlation Florida, Hawaii, Massachusetts, New Jersey, Missouri, and New York. However, none of these fail to reject the null hypothesis, indicating no linear trend.

There is some uncertainty in our findings due to our small sample size and the violation of the independence-of-errors assumption in linear regression.

Colorado and Kansas should start studying the factors and populations with higher

asthma rates and explore ways to mitigate them. While we think the trends may be related to factors such as increased wildfire frequency in Colorado between 2011 and 2021, likely driven by higher temperatures and drought caused by climate change, and possible increases in asthma risk due to wildfire smoke containing pollutants such as PM2.5 and other irritants, our analysis does not establish causality. Similarly, although Kansas ranks very high among U.S. states for obesity, with obesity prevalence rising from 30% in 2014 to 35% by 2020 , and global research suggests obesity contributes to asthma, it is only possible that increasing obesity in Kansas is one source driving the higher prevalence of asthma. Our dataset and analysis cannot confirm these causal relationships. Overall, we found that asthma prevalence is increasing in about 88% of the United States, and we need to continue funding research into why this is happening and how to mitigate it. As a starting point, we should investigate which policies and environmental factors are driving the decrease in asthma prevalence in the states mentioned above and consider applying them to other states.

5.5.1 Limitations

Our analysis has several limitations; we have a small sample size of only around 11 observations per state, which substantially limits our power to see more moderate trends in the data. Secondly, the violation of our data's independence complicates how we interpret it. When errors are autocorrelated, our standard errors can be underestimated, increasing the risk of false positives. Third, since we are only focused on time prevalence, we cannot see the factors that go into changes in asthma prevalence. Such as environmental, policy, population, and other changes that may explain why asthma prevalence is increasing or decreasing. Fourth, our simple linear regression suggests that the trends are strictly linear and can't capture more complex patterns, such as exponential curves or drops that then rise. We avoided p-hacking by using the same sample size for each test (except for FL and NJ), applied many error-correction algorithms, and went with the one that best suited the data, not the results, and acknowledged all our assumption violations.

Having more granular, extensive data would definitely help our study. If we had county-level data, we could conduct geospatial analyses to pinpoint areas or groups that show a trend and identify potential cofounders. Also, having more data on potential confounders such as air quality, obesity, and policy would allow us to better analyze the data and apply more detailed solutions. Even having more time data would be beneficial, so we can see if there was a trend that could help shape policies.

5.5.2 Prior Work Comparison

Our results differ from prior work trends in adult current asthma prevalence and risk factors in the US by state 2000–2009. They discovered that there was a significant increasing trend in state-specific asthma prevalence. Our overall results indicate an increasing trend in most states; however, many are not linear, and some even show decreasing trends. We believe this is because many states are trying to adopt policies to reduce asthma prevalence and promote better health nationwide. However, we can still see asthma prevalence increasing, so there must be other reasons for that. Overall, we do not see a significant increase in asthma prevalence.

6 Research Question 2

Research Question How do state-level asthma prevalences in adults differ after partial pooling, and how is adult asthma prevalence associated with adult smoking prevalence?

6.1 Methods

6.1.1 Model Overview and Graphical Model

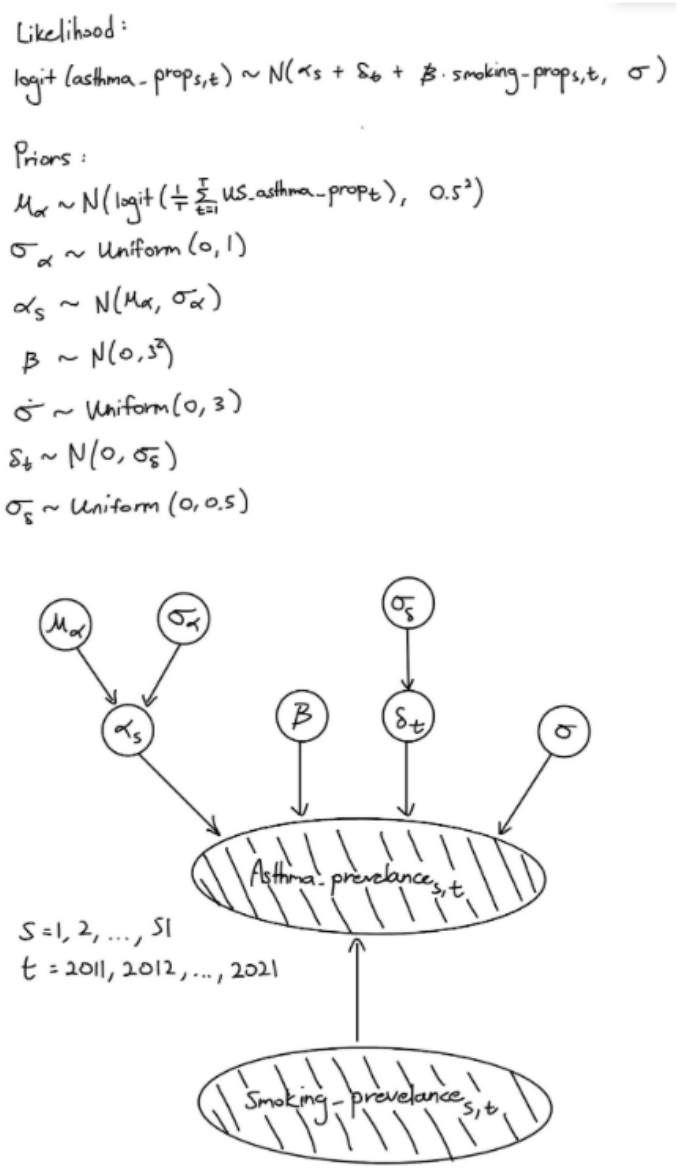


Figure 8: Graphical representation of the Bayesian hierarchical model.

6.1.2 Data Structure and Hierarchical Groups

The groups in the hierarchical model are the 50 U.S. states plus the District of Columbia, indexed by

$$s = 1, 2, \ldots, 51.$$

Within each state, we observe the prevalence of asthma and smoking for the years

$$t = 2011, 2012, \ldots, 2021.$$

In the model, these two observed variables are denoted

$$\text{Asthma\_prevalence}_{s,t} \quad \text{and} \quad \text{Smoking\_prevalence}_{s,t}.$$

Note we only draw one node for each for space purposes

Two state–year observations are missing in the dataset: Florida in 2021 and New Jersey in 2019.

6.1.3 Model Parameters and Hierarchical Structure

Running the model in PyMC produces posterior samples for each unobserved random variable. These include the global parameters

$$\mu_\alpha,\ \sigma_\alpha,\ \sigma_\delta,\ \beta,\ \sigma,$$

as well as the state-level vector

$$\alpha_s \quad \text{(one baseline parameter for each state)}$$

and the year-level vector

$$\delta_t \quad \text{(one effect for each year)}.$$

Below is a definition for each term:

- α_s: State s baseline asthma level on the log-odds (logit) scale.
- μ_α: The average baseline asthma level across states on the logit scale.
- σ_α: The between-state standard deviation in baseline asthma levels on the logit scale.
- δ_t: The effect of year t, relative to the average year in the dataset, on the log-odds of asthma prevalence after accounting for both state baseline asthma prevalence and smoking prevalence.
- σ_δ: The variation in year-specific effects.
- β: The global slope measuring how the log-odds of asthma prevalence change as smoking prevalence increases.
- σ: Residual noise capturing variation unexplained by the model.

The parameters α_s, μ_α, δ_t, and β are defined on the logit scale, while σ_α, σ_δ, and σ represent standard deviations measured in logit units. These are all the hidden variables we are trying to estimate.

6.1.4 Assumptions and Justification for Priors

We make the simplifying assumption that the conditional distribution of logit asthma_prevalence given smoking prevalence is normally distributed. We apply a logit transformation to asthma_prevalence so that we can interpret coefficients as log odds and map back to prevalence via the sigmoid function.

We set priors and hyperpriors to define our model. Note that the average U.S. asthma prevalence in our dataset is $\bar{p}_{\text{US}} = 9.3\%$. We set

$$\mu_\alpha \sim \mathcal{N}\big(\text{logit}(\bar{p}_{\text{US}}) = -2.277542957,\ 0.5^2\big)$$

because it centers the average baseline asthma prevalence at the U.S. average across the datapoints we have access to. The standard deviation of 0.5 coupled with the choice of a normal distribution, allows for a wide, weakly informative range that contains the 2011–2014 U.S. average of 8.8% among adults (CDC).

We set

$$\sigma_\alpha \sim \text{Uniform}(0, 1)$$

to express our initial uncertainty about the variance in the average baseline asthma prevalence. We select an upper bound of 1 because it captures enough variance to support the studies we looked through while maintaining our initial uncertainty. Consequently, since

$$\alpha_s \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha),$$

α_s can be drawn from a high- or low-variance distribution with equal probability. Since all state asthma prevalence baselines α_s come from the same shared distribution, after μ_α and σ_α are updated by the data, partial pooling is induced.

We model the year-specific effects as

$$\delta_t \sim \mathcal{N}(0, \sigma_\delta) \quad \text{and} \quad \sigma_\delta \sim \text{Uniform}(0, 0.5).$$

We center δ_t at 0 because we initially believe that no specific year t contributes a higher or lower asthma prevalence than the typical year in our dataset. We set σ_δ to be $\text{Uniform}(0, 0.5)$ to capture our initial uncertainty about year-to-year effect variation. With equal probability, δ_t could have very little variability or reasonably large variability.

The upper bound of 0.5 was chosen to allow potentially large variance in year-to-year effects while still encoding uncertainty in the variance. For instance, if a state's asthma prevalence was deterministically equal to 9% (-2.31 on the logit scale), a value of $\sigma_\delta = 0.5$ allows 68% of the year effect to change asthma prevalence to 5.7%–14.1% (corresponding to -2.81 to -1.81 on the logit scale), assuming a normal distribution for δ_t.

We place a

$$\beta \sim \mathcal{N}(0, 3^2)$$

prior centered at 0 because before observing any data, we do not believe smoking is associated with either increases or decreases in asthma prevalence. The standard deviation of 3 was chosen so that the data has a strong influence on the posterior.

For illustration, if asthma prevalence was deterministically equal to 9% (-2.31 on the logit scale) before accounting for smoking, then with 68% probability, a 10% increase in

smoking prevalence would change asthma prevalence to anywhere from 6.8% to 11.8% (-2.61 to -2.01 on the logit scale). With 95% probability, a 10% increase in smoking prevalence would change asthma prevalence to anywhere from 5.2% to 15.3% (-2.91 to -1.71 on the logit scale). Thus, this prior is very weakly informative.

We place a

$$\sigma \sim \text{Uniform}(0, 3)$$

prior on σ to encode high uncertainty about all other factors that could potentially affect asthma prevalence. The upper bound of 3 allows for large variation in observed asthma prevalence about the prediction.

At the extreme case where $\sigma = 3$, if asthma prevalence were predicted to be 9% (-2.31 on the logit scale) after adjusting for year and smoking effects, then with 68% probability the observed prevalence could range from 0.5% to 66.9% (-5.31 to 0.69 on the logit scale), providing significant flexibility in the residuals. With equal probability, σ takes on values between 0 and 3, expressing uncertainty in the variability of observed data.

Observing the posterior characteristics of α_s for each state directly answers the first part of the research question: *How do state-level asthma prevalences in adults differ after partial pooling?*

Observing the posterior characteristics of β directly answers the second part of the research question: *How is adult asthma prevalence associated with adult smoking prevalence?*

6.2 Results

6.2.1 Choice of Hierarchical Modeling Approach

We chose to use a full hierarchical model, not empirical Bayes. This is because we wanted uncertainty in our hyperparameters and global parameters. After reading the literature, it was difficult to pinpoint fixed numbers for μ_α, σ_α, σ_δ, β, and σ. Thus, we choose a weakly informative range for their priors to reflect our initial uncertainty. Using empirical Bayes may have been a good idea, but it would have produced point estimates and failed to capture our initial uncertainty. For example, it was difficult to find an exact value for σ_α which captures the strength of pooling toward the global mean μ_α. While empirical Bayes would have given us an estimate for σ_α from the data, we erred on the side of caution and assigned a very weakly informative prior for it, so that the data heavily determines the posterior. We also intentionally center β at zero to reflect our initial belief of no relationship between smoking and asthma.

6.2.2 State-Level Baseline Asthma Prevalence After Partial Pooling

After partial pooling, the states with both the lowest posterior mean and median baseline asthma prevalence were (in order from lowest to highest) Texas, South Dakota, Florida, Nebraska, and Minnesota. They had posterior median asthma prevalences of 7.5%–8.1%. The states with the highest posterior mean and median baseline asthma prevalence were (in order from lowest to highest) New Hampshire, West Virginia, Vermont, Rhode Island, and Maine. They had posterior median asthma prevalences of 11.3%–11.8%.

To assess the total spread in asthma prevalence after partial pooling, we take the difference between the state with the lowest posterior median prevalence (Texas) and the highest posterior median prevalence (Maine), which gives us

$$\text{Med}_{\text{Maine}} - \text{Med}_{\text{Texas}} = 11.8 - 7.5 = 4.3\%$$

prevalence, which is quite large.

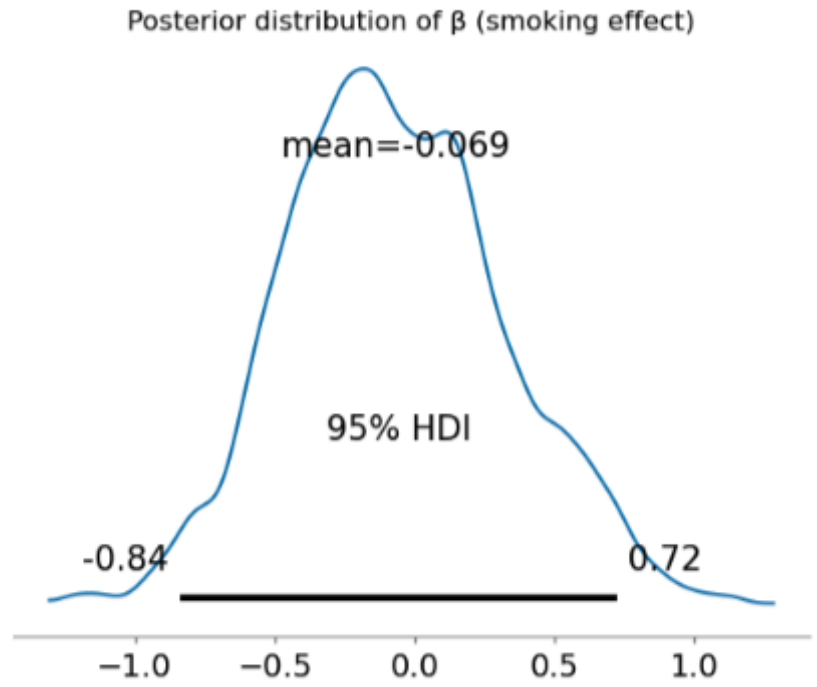
The 95% credible interval for Texas ranges from 6.6% to 8.4%, while the 95% credible interval for Maine ranges from 10.1% to 13.3%. Surprisingly, since they do not overlap, our model suggests a significant difference in baseline asthma prevalence across at least the lowest and highest state. The 95% credible interval for Texas tells us that, given the data and prior, there is a 95% posterior probability that the true baseline asthma prevalence for Texas is between 6.6% and 8.4%.

Region	average_median_prev
Midwest (West North Central)	8.600000
South (West South Central)	8.775000
West (Mountain)	9.325000
South (South Atlantic)	9.366667
Mid-Atlantic	9.533333
Pacific	9.580000
South (East South Central)	9.675000
Midwest (East North Central)	9.960000
Northeast	11.200000

Figure 9: Average posterior median asthma prevalence by U.S. region.

Further, we group states by region and take the average of the posterior median asthma prevalence in each region. The region with the lowest average median prevalence was the Midwest at 8.6%, and the region with the highest average median prevalence was the Northeast with 11.2%, giving a spread of

$$11.2 - 8.6 = 2.6\%.$$



Beta summary:

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta	-0.069	0.402	-0.841	0.723	0.04	0.016	110.0	443.0	1.04

```
Posterior mean for beta: -0.06918533441480403
Posterior median for beta: -0.09357786379687794
(2.5%, 97.5%) quantiles: -0.8182920684332458 0.7501427807609623
```

Figure 10: Posterior distribution and summary for β.

6.2.3 Credible Interval for the Smoking Effect

The posterior mean of β was -0.069 and the posterior median was -0.0936. Interpreting the median, this tells us that a 10% increase in a state-year's smoking prevalence is associated with a decrease of

$$0.0936 \times 0.1 = 0.00936$$

in the log odds of asthma prevalence.

Surprisingly, the 95% highest density interval (HDI) includes 0, which tells us that the direction of the change in asthma prevalence associated with increased smoking prevalence is uncertain. Interpreting the 95% HDI for β, we know that, given the data and prior, there is a 95% posterior probability that the true effect of smoking prevalence on the log odds of asthma prevalence lies between -0.84 and 0.72.

6.2.4 Sensitivity Analysis

To see whether our results depended heavily on the choice of prior, we experimented with different reasonable priors for the smoking effect parameter, β. In our main model, we used a weakly informative prior of

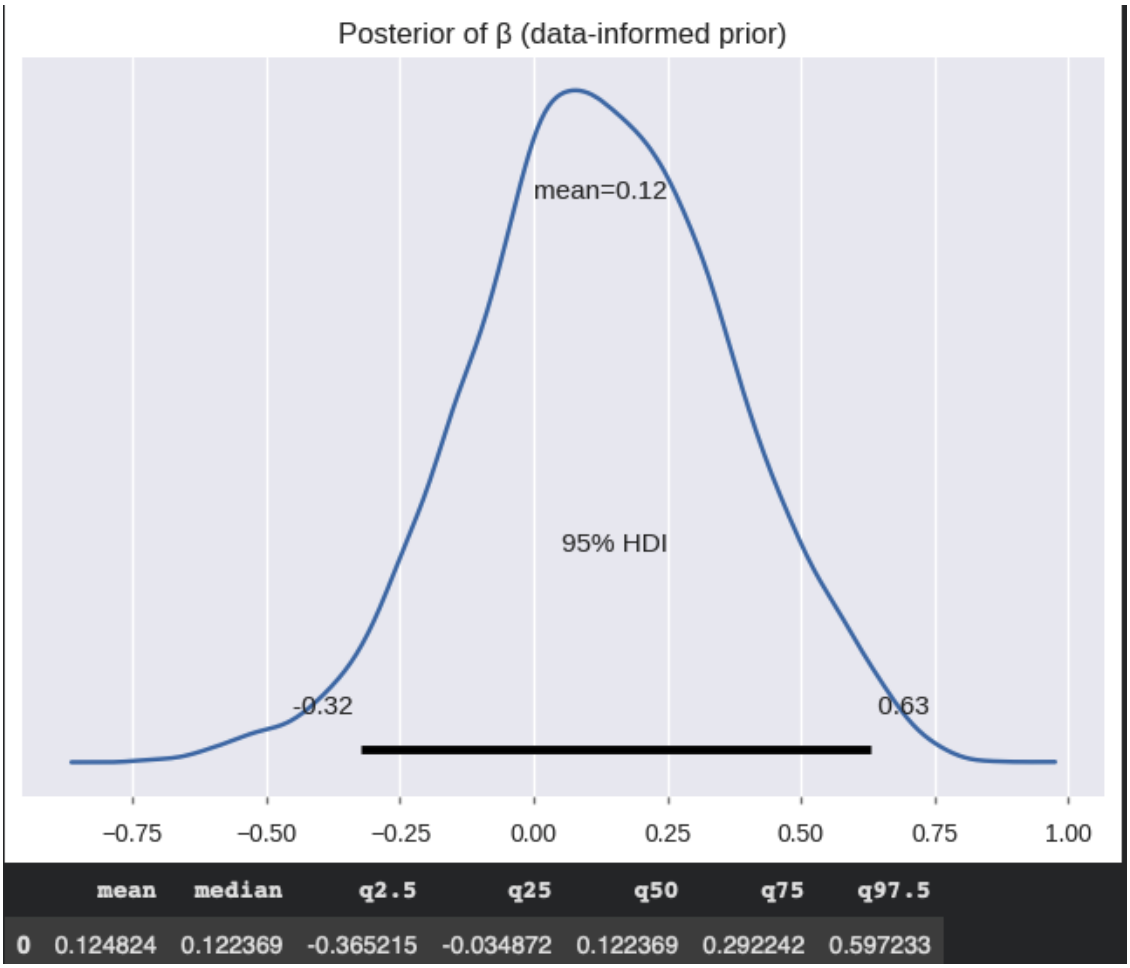
$$\beta \sim \mathcal{N}(0, 3^2),$$

which allows β to take on a wide range of values.

We then reran the model using a more data-informed prior,

$$\beta \sim \mathcal{N}(0.2, 0.3^2),$$

based on what we observed in our exploratory data analysis and findings reported in public health studies, which suggest that any association between smoking prevalence and asthma prevalence is small.



	mean	median	q2.5	q25	q50	q75	q97.5
0	0.124824	0.122369	-0.365215	-0.034872	0.122369	0.292242	0.597233

Figure 11: Posterior distribution of the smoking effect parameter β under the data-informed prior $\mathcal{N}(0.2, 0.3^2)$.

When using the data-informed prior, the posterior distribution of beta shifted upward. The posterior mean and median was approximately 0.12. Since the 95% credible interval

of -0.37 to 0.60 still included 0, the data still do not support a clear positive or negative effect of smoking prevalence on asthma prevalence. Hence, our overall conclusion remained the same. This suggests that our results are not sensitive to reasonable changes in the prior specification for β.

6.2.5 Final Result

Taking the results for β and each α_s together, we conclude that there is a meaningful difference in state baseline asthma prevalence, while the association between smoking prevalence and asthma prevalence is small and has an uncertain direction.

6.3 Discussion

6.3.1 Interpretation of Partial Pooling and Shrinkage

Since the likelihood of the logit asthma prevalence is modeled as conditionally normally distributed and α_s is normally distributed, we have approximate normal–normal conjugacy for each α_s. As a result, each baseline asthma prevalence α_s is an approximate convex combination of the state's own mean asthma prevalence and the shared prior mean μ_α. Consequently, there is shrinkage, where the no-pooling estimate (which uses only a state's own data) is pulled toward the global mean μ_α.

6.3.2 Limitations

Since we did not have access to exact prior information, we choose all of our priors to be broad and weakly informative. This would allow the posterior to be more heavily influenced by the data. Since the data analysis is performed at the state-year level, we cannot infer individual-specific effects of smoking on asthma. Perhaps the largest limitation is the high likelihood of omitted variable bias. Since many more variables affect asthma prevalence besides smoking prevalence, we cannot make causal claims. We model the conditional likelihood of the logit asthma prevalence for state years as normally distributed, which may not be true in practice. Our model assumes a global smoking prevalence slope for all state-years, which also may not be true in practice.

6.3.3 Convergence Diagnostics

Using PyMC, we drew 16,000 samples across 4 chains (4000 samples per chain). We intially used the parameter `target_accept = 0.95`, which resulted in convergence issues. Using this, approximately 4,000 samples were divergent, indicating that the sampler took steps that were too large to explore the posterior distribution adequately. To fix this, we increased the target acceptance rate to `target_accept = 0.99`, which caused the sampler to take smaller steps and resulted in zero divergent samples.

6.3.4 Alternative Model Formulations

Initially, we tried a different model that did not incorporate a temporal trend. Everything in the model stayed the same except for δ_t and σ_δ, which were removed. This did not

work because it did not make use of time at all. Omitting this would force the effects of nationwide changes throughout time in asthma prevalence to be absorbed into smoking prevalence, state-specific baseline asthma, and noise, which would have biased our estimates.

6.3.5 Additional Data and Graphical Model Extension

Data on confounding variables that vary across both state and time, including air quality, economic condition of the state, and healthcare, can help refine our estimate of the association between smoking prevalence and asthma prevalence as well as the logit baseline asthma prevalence for each state α_s. We could add these variables by including each confounder additively with its own regression coefficient. Below, each $Z_{k,s,t}$ is a confounding variable that affects both asthma prevalence and smoking prevalence. When we also condition on them, the estimate of the association becomes more refined as β and α_s absorb fewer omitted variables.

We can represent this extended model as

$$\text{logit}(\text{asthma\_prevalence}_{s,t}) \sim \mathcal{N}\left(\alpha_s + \delta_t + \beta \cdot \text{smoking\_prevalence}_{s,t} + \sum_k \gamma_k Z_{k,s,t},\ \sigma\right).$$

6.3.6 Comparison with Prior Work

Although we could not find any similar previous work that applies the same Bayesian hierarchical model as our work, a considerable amount of prior work exists on studying the association between asthma and smoking in the public health domain by employing other models. A considerable amount of prior work suggests that a positive association exists between the outcome of asthma and smoking, especially when assessed individually.

On the other hand, our study examines population-level variations based on state-year data. In this case, there is no significant relationship between smoking rates and asthma rates based on data because the 95% credible interval for the smoking rate variable encompasses zero. This discrepancy could arise because our study uses ecological data. A relationship that is significant at the individual level might not necessarily be important at a population level.

Therefore, our findings are not necessarily inconsistent with previous research but rather illustrative of the fact that the relationship between smoking and asthma may vary based on the unit of observation. Although previous research has shown that smoking is a key individual risk factor for the onset of asthma, our research illustrates that state-level variation in adult rates of asthma likely differ based on other factors beyond rates of smoking.

7 Conclusion

7.1 Outcomes Summary

In this project, we learned about the prevalence of adult asthma in the USA through two different methods.

For the first research question, we examined if there was a linear increase or decrease in the prevalence of asthma for each state from 2011 to 2021. After employing the Bonferroni correction for multiple hypothesis testing, we observed a statistically significant linear trend only for Colorado and Kansas. The observation indicates that although a few states have a slightly increasing prevalence of asthma or there could be other states with a slightly decreasing rate, there is no statistically significant linear change for most states.

Moving on to our second research question, we applied the Bayesian hierarchical model with the feature of partial pooling for the study of variations in the baseline prevalence of adult asthma by states and how it is influenced by the prevalence of adult smoking. When performing the partial pooling on our model, we identified that there is a considerable variation in the baseline prevalence of asthma by states. States with a lower baseline prevalence of asthma include Texas and South Dakota, and states with a relatively higher baseline prevalence of asthma include Maine and Rhode Island. There is a difference of 4.3 percentage points between the lowest and highest baseline asthma prevalence at the state level after the partial pooling is performed on the model.

Furthermore, we examined the association between adult prevalence of smoking and adult prevalence of asthma. The 95% credible interval for the effect parameter for smoking included zero in the posterior distribution. This implies that, while there could exist a possible association between the two variables, prevalence of asthma and prevalence of smoking are not known from a population perspective. Sensitivity analysis, performed assuming both weak and informed priors, was seen to arrive at the same conclusion and hence was deemed to be robust for a reasonable choice of priors.

Overall, from our findings, there is variation in the prevalence of asthma in adults within the United States, there is no significant relationship between smoking prevalence and asthma prevalence, and few states have a significant increase in asthma prevalence.

7.2 Critical Evaluation

7.2.1 Data Limitations

However, it should be noted that the results for our study can be limited by the fact that the data are not disaggregated. In other words, because the data are not disaggregated, it is not possible for us to conclude that there is a correlation between the data on people who smoke and the data on people with asthma. Moreover, there are a number of other factors that can influence the prevalence of people with asthma that are not present in the data. This can cause the results to suffer from the problem of omitted variable variables.

7.2.2 Missing Domain Knowledge

Although we looked into the public health literature for advice on how to make our model choices, we did not have deep knowledge within the area of asthma epidemiology. A thing that we might ask a domain expert for advice on is whether there is a possible population-level impact for smoking prevalence compared to the individual-level impact. A possible explanation for us finding no discernible link at a population level could very well be if the impact of smoking on asthma outcomes lies at the individual-level.

7.2.3 Robustness to Modeling Choices

Our findings are based on various modeling decisions such as the application of a linear trend model for Research Question 1 and the global smoking effect for Research Question 2. For the first research question, we assumed that the trend of change in the prevalence of asthma is linear with time. However, if the actual trends are nonlinear, then our model might not precisely represent the trends. For the second research question, we assumed that there is one overall smoking effect for all states and for all years. Even though this is simpler to interpret for our research questions, there might be state and time effects that are masked by our model. However, from our sensitivity analysis study, the principal finding on the smoking effect was the same for different possible prior distributions, and hence our findings are relatively robust on this point.

7.2.4 Generalizability of Findings

However, it must also be clarified that our results specifically hold for adult cases of asthma prevalence in the US from 2011 to 2021, and hence may not hold true for any other date range, geographic location, or age demographic. Finally, we point out that our results are based on population trends, and hence should not, and cannot, be applied for determining health outcomes for an individual person or patient.

7.3 Recommendations

7.3.1 Future Studies

We concluded that there is no association between smoking and asthma prevalence at the state-level (RQ2). We propose a follow-up study where data are collected from individuals who have asthma, and both those who do and do not have asthma. This data would ideally record hypothesized covariates such as demographic and socioeconomic information. In this way, we would be able to directly compare those who smoke with those who do not, and in turn, this follow-up study could help us confirm whether the lack of relationship is a result of state-level aggregation. For example, if we reject the null hypothesis at the individual-level, it may suggest that some other variables (e.g., environmental pollution, access to healthcare) are confounding our ability to detect the association.

7.3.2 Call to Action

Given our findings, we argue that public health strategies to target the reduction of asthma prevalence among adults should not exclusively consider policies related to smoking. Although smoking status can be considered one determinant of poor health outcomes, our results suggest that differences in asthma prevalence across states likely arise from factors not represented in our dataset. Instead, emphasis should be placed on region-specific asthma interventions that are informed by region-specific differences in the prevalence of asthma. In this way, funding should be allocated to those states that have a high baseline prevalence of asthma in the regions of the Northeast, even in regions with partial pooling of asthma.

7.3.3 Potential Impacts

The above recommendation is possible since state and national health organizations have the capacity to launch targeted strategies for the prevention and control of asthma. The state health organizations are the best to develop strategies according to the context. However, there may also be valid defense on how our attention is diverted from the current focus on the policy related to smoking, especially in the context that the negative effects of smoking are already well-established. The response to the defense is that our argument does not undermine the significance of the strategy to reduce the problem of smoking but only underscores the significance that strategies for controlling and preventing asthma cannot benefit from the strategy of reduced smoking only. The effects of the recommendation are most probably going to differ according to the groups. Those are the states where the prevalence of asthma is high will benefit greatly from the intervention, and those are the states where the prevalence is low will benefit least from the intervention. From the moral perspective, our recommendation is framed by the principles of allocation according to what is observed in the health outcomes. Regions in diverse places may probably have the diverse levels of the outcomes, and this is the strategy to improve the allocation of resources in the decisions on issues of importance in the improved health.

Bibliography

"API Reference — Statsmodels." n.d. Www.statsmodels.org. https://www.statsmodels.org/dev/api.html.

Brockwell, Peter J, and Richard A Davis. 2010. *Introduction to Time Series and Forecasting*. New York: Springer.

CDC. 2019. "Asthma Stats - Percentage of People with Asthma Who Smoke | CDC." Www.cdc.gov. January 2, 2019. https://www.cdc.gov/asthma/asthma_stats/people_who_smoke.htm.

Frost, Jim. 2023. "Monte Carlo Simulation: Make Better Decisions." Statistics by Jim. February 2023. https://statisticsbyjim.com/probability/monte-carlo-simulation/.

GeeksforGeeks. 2022. "Assumptions of Linear Regression." GeeksforGeeks. December 27, 2022. https://www.geeksforgeeks.org/machine-learning/assumptions-of-linear-regression/.

Graebner, Robert. n.d. "Study Design with SAS ® : Estimating Power with Monte Carlo Methods." https://support.sas.com/resources/papers/proceedings/proceedings/sugi24/Posters/p236-24.pdf.

"Introductory Overview of PyMC — PyMC 5.13.1 Documentation." n.d. Www.pymc.io. https://www.pymc.io/projects/docs/en/stable/learn/core_notebooks/pymc_overview.html.

Kane, Natalie. 2022. "Revealing the Racial and Spatial Disparity in Pediatric Asthma: A Kansas City Case Study." *Social Science & Medicine* 292 (January): 114543. https://doi.org/10.1016/j.socscimed.2021.114543.

Lee, Sarah. 2025a. "A Comprehensive Guide to Durbin-Watson Test in Regression Models." Numberanalytics.com. 2025. https://www.numberanalytics.com/blog/comprehensive-guide-durbin-watson-regression-models.

———. 2025b. "The Ultimate Guide to Autocorrelation." Numberanalytics.com. 2025. https://www.numberanalytics.com/blog/ultimate-guide-autocorrelation-regression#correction-techniques.

Noah, Terry L, Cameron P Worden, Meghan E Rebuli, and Ilona Jaspers. 2023. “The Effects of Wildfire Smoke on Asthma and Allergy,” May. https://doi.org/10.1007/s11882-023-01090-1.

“RMA Intelligence.” 2024. Nifc.gov. 2024. https://gacc.nifc.gov/rmcc/intelligence.php.

“SciPy API — SciPy V1.9.3 Manual.” n.d. Docs.scipy.org. https://docs.scipy.org/doc/scipy/reference/index.html#scipy-api.

Zhang, Xingyou, Teresa Morrison-Carpenter, James B Holt, and David B Callahan. 2013. “Trends in Adult Current Asthma Prevalence and Contributing Risk Factors in the United States by State: 2000–2009.” *BMC Public Health* 13 (1). https://doi.org/10.1186/1471-2458-13-1156.