CEJS Activity

Armed with all of that background knowledge on the CEJS dataset and on statistical inference, we can proceed with a detailed analysis of the data. In particular, we are interested in assessing the following hypothesis:

> *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

In this part of the session you will interpret results from a statistical model fitted to datasets from different U.S. States.

> Instructor Note: Introduction
>
> These are instructor notes; they will be removed from the student-facing assignment file. This is the **Bayesian** form of the activity.

Overview

Throughout this activity, you will be studying a statistical model fitted to data from the CEJS dataset. As a reminder, we are interested in the `Energy Burden Percentile` (higher values correspond to a higher burden) and the `Percent Black Census Tract` (which measures the number of people in a region who are Black).

From this scatterplot, we can see that the energy burden seems to increase as the percent Black increases. However, we can make this rough observation more formal by using a statistical model.

Analyze a Bayesian Model

To analyze the dataset, we will use the following statistical model,

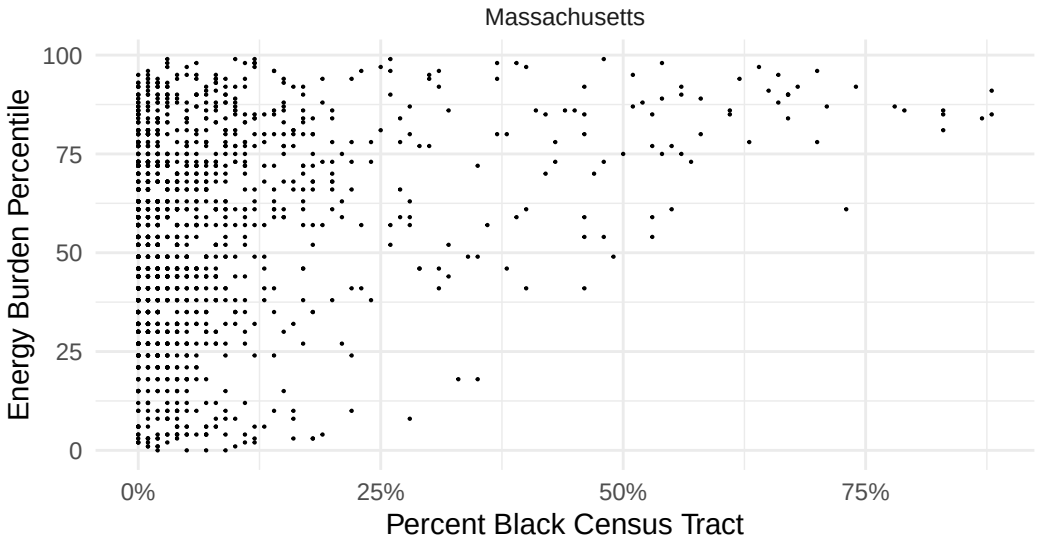


Figure 1: CEJS data for Massachusetts.

$$B = mP + b + \epsilon$$

where B is the energy burden percentile, P is the percent Black, m is the slope parameter, b is the intercept parameter, and ϵ is a *residual* term that represents factors not accounted in the model. The residual term is assumed to be normally distributed $\epsilon \sim N(0, \sigma^2)$ with an unknown parameter σ^2. All three parameters have a prior distribution, defined via

$$\begin{aligned} m &\sim N(\mu_m, \sigma_m^2), \\ b &\sim N(\mu_b, \sigma_b^2), \\ \sigma^2 &\sim \text{Exponential}(1/s_y), \end{aligned}$$

where m, b, σ^2 are independent.[1] We will discuss how to set the prior through its parameter values $\mu_m, \mu_b, \sigma_m^2, \sigma_b^2$ later in this activity.

Instructor Note: Model Assumptions

Note that this model makes a number of important assumptions, which students may identify and question. We recommend validating student input, but try to maintain a focus on the assumptions that are aligned with the lesson's learning objectives. We enu-

[1] Note that s_y is determined based on the standard deviation in the observed data. This is a way of *autoscaling* the prior.

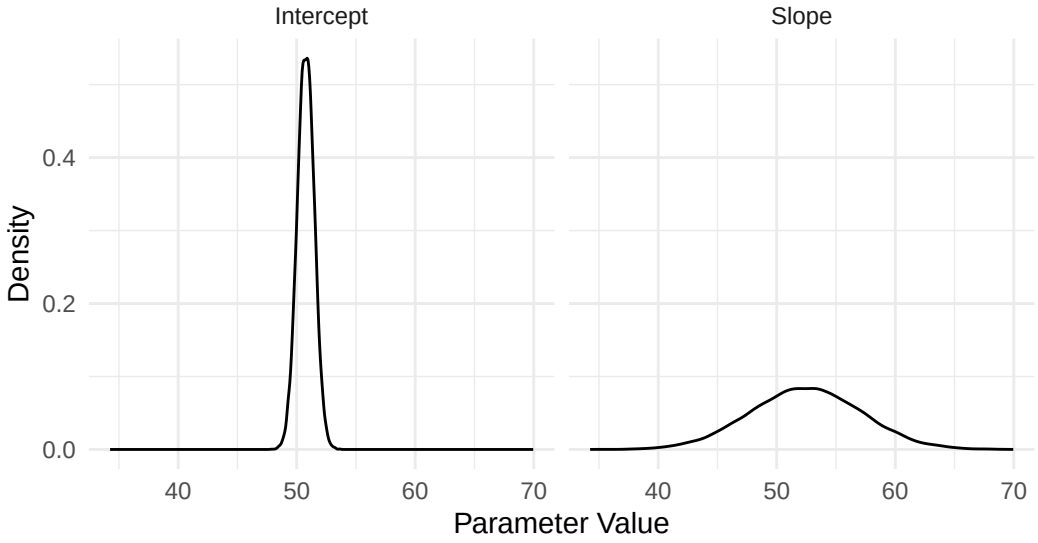
merate important model assumptions, ramifications, and relevance to learning objectives here:

- The responses B_i are independent when conditioned on the percent Black P.
  - This is almost surely not true as there are a variety of other factors that affect one's energy burden, such as State-level economic policies. These other factors are not entirely captured in our lone predictor (P), which may manifest as association between the observed responses (B_i). This will lead to an *omitted variable bias* in our estimates.
  - While omitted variable bias is an important consideration, it is outside the learning objectives in this lesson since this assumption is shared between the Frequentist and Bayesian approaches.
- The structure of the response is linear; that is $B = mP + b + \epsilon$.
  - This discounts the possibility of nonlinearity; for instance, there could be little change in the mean energy burden at small percent Black, but much larger change at higher values.
  - While the structure of the response is an important consideration, it is outside the learning objectives in this lesson since this assumption is shared between the Frequentist and Bayesian approaches.
- Residuals are normally distributed $\epsilon \sim N(0, \sigma^2)$ with constant σ^2.
  - This will never be exactly true, which we can check by inspecting the residuals. This assumptions has implications for our predictive uncertainty; for instance, assuming a constant σ^2 discounts the possibility of heteroskedasticity.
  - While the residual distribution is an important consideration, it is outside the learning objectives in this lesson since this assumption is shared between the Frequentist and Bayesian approaches.
- The intercept b and slope m parameters are treated as random variables with a distribution that represents our state of knowledge.
  - This is a fundamental component of the Bayesian approach, and hence is directly related to the lesson's learning objectives.

Study the posterior distribution

Results from a Bayesian model take the form of a *posterior distribution* for the model parameters. The model has two parameters that are closely related to our hypothesis: The slope m and intercept b of the fitted line. "Fitting" the Bayesian model to the Massachusetts dataset will result in a posterior distribution for the parameters, with an example posterior given

below:



The posterior distribution helps us determine **how confident** we should be in conclusions drawn from the model. The next exercise will help you assess confidence in results based on the fitted model.

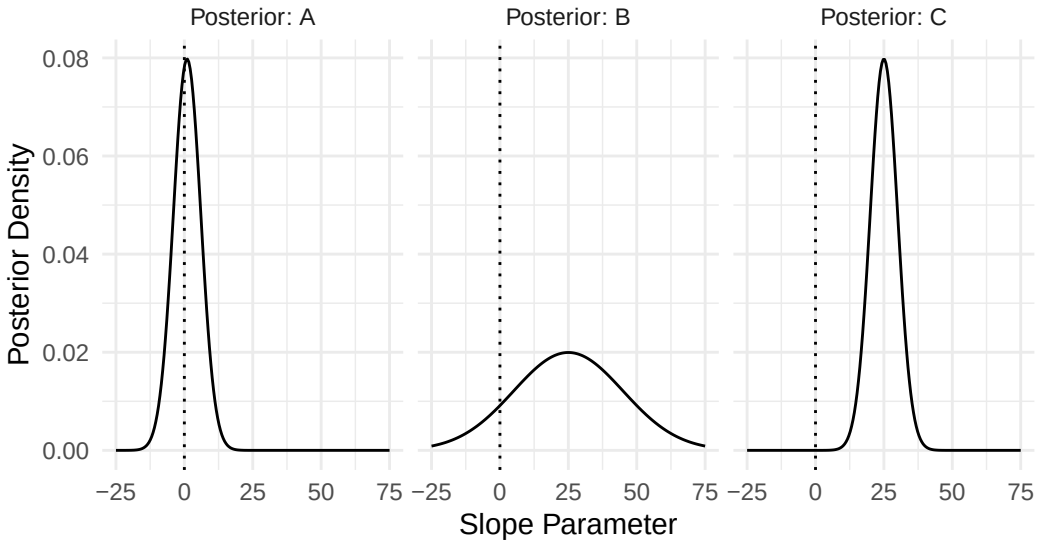
Assessing confidence

Let's imagine three different posterior distributions for the posterior (marginal) distribution for the slope parameter m.

Model summaries

Questions for the Class

- Roughly, what fraction of *Posterior Distribution A* is greater than zero?
  - (Write your response here):
- Roughly, what fraction of *Posterior Distribution B* is greater than zero?
  - (Write your response here):
- Roughly, what fraction of *Posterior Distribution C* is greater than zero?
  - (Write your response here):
- Which posterior gives the *highest confidence* (*highest probability*) that the slope parameter is positive? How can you tell?



– (Write your response here):

Questions for the Class

- Roughly, what fraction of *Posterior Distribution A* is greater than zero?
  - Precisely 57.93%, so roughly 60%
- Roughly, what fraction of *Posterior Distribution B* is greater than zero?
  - Precisely 89.44%, so roughly 90%
- Roughly, what fraction of *Posterior Distribution C* is greater than zero?
  - Essentially 100%
- Which posterior gives the *highest confidence* (*highest probability*) that the slope parameter is positive? How can you tell?
  - Posterior C, as the probability is essentially 100% (as we identified above).

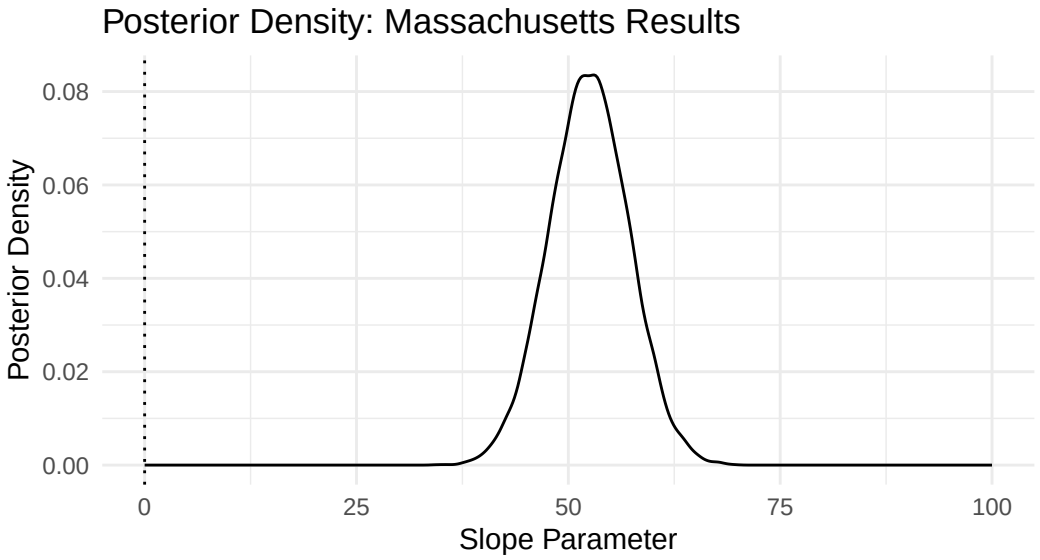
General inference

- How does the slope parameter relate to our hypothesis? As a reminder, our hypothesis is:

  As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).

- (Write your response here):
- The slope parameter is directly related to our hypothesis. A positive slope is in agreement with our hypothesis.

Let's return to the posterior from our model for the Massachusetts data and use the same reasoning as above to make sense of the results.



Questions for the Class

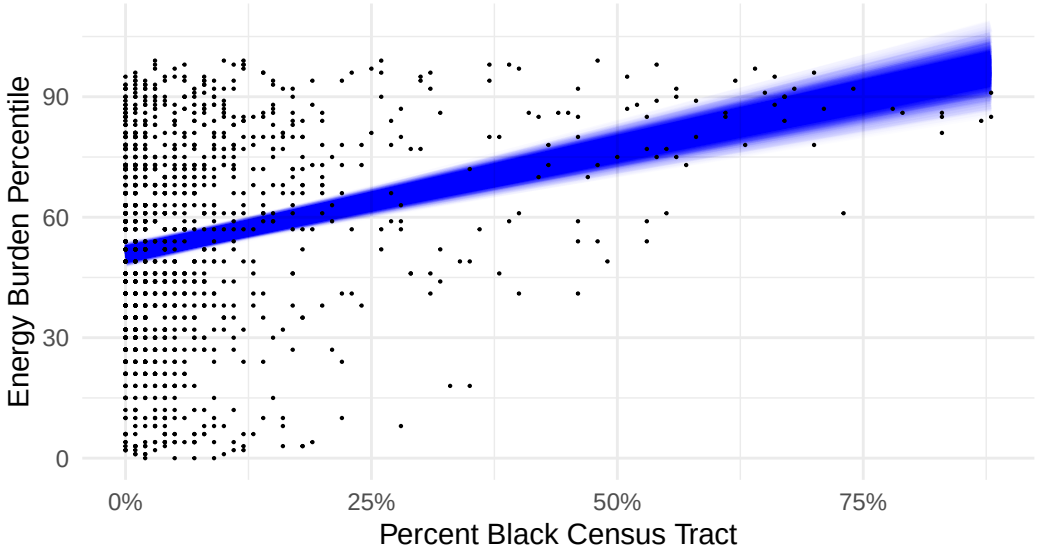
- For Massachusetts, does the fitted model support or contradict our hypothesis? As a reminder, our hypothesis is:

  > *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

- (Write your response here):
- How confident are you in the model results?
  - (Write your response here):
  - The fitted model supports our hypothesis.
- How confident are you in the model results?
  - The model assigns a nearly 100% probability to a positive slope, which suggests a high confidence in a positive slope.

Study the posterior predictions

Bayesian analysis does not produce a "best" line; rather, the posterior distribution implies a *family* of lines (each with a different chance). We call this the *posterior predictive distribution*. For instance, the following visualizes the posterior predictive distribution of lines against the Massachusetts data. This object appears as a "cone" with darker regions corresponding to lines with higher probability (density).



We can use this kind of plot (predictions against observed data) as a way to sanity check the model. You'll do this in the following questions.

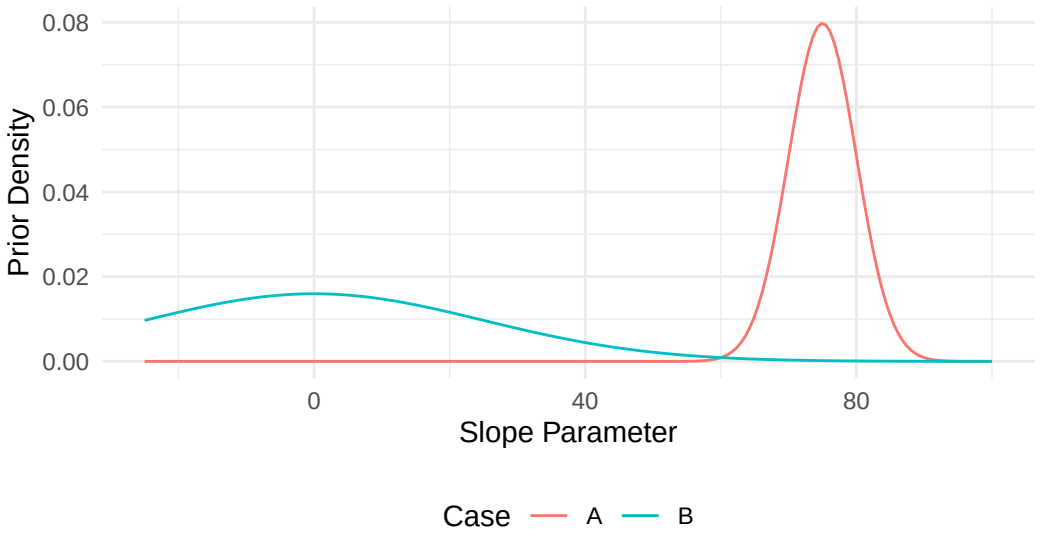
Questions for the Class

- How well do the posterior predictions (transparent blue lines) agree with the data (black dots)? Do they follow the same general trend as the data?
  - (Write your response here):
  - The posterior predictions tend to agree with the data.

The Prior Distribution

Above, we glossed over how we arrive at a posterior distribution. In addition to the equation for the line we must also provide a *prior distribution* for the model's parameters. In a Bayesian analysis the prior represents all of our *prior knowledge* about the problem.

For instance, if we were confident that the slope of the Energy Burden vs Percent Black line is positive, we could represent that with a prior distribution that was tightly concentrated at a positive value (Case A below). If we were highly uncertain about the slope, we might represent this case with a prior distribution centered at zero with a very large standard deviation (Case B below).



Fitting a model: Data + Prior = Posterior

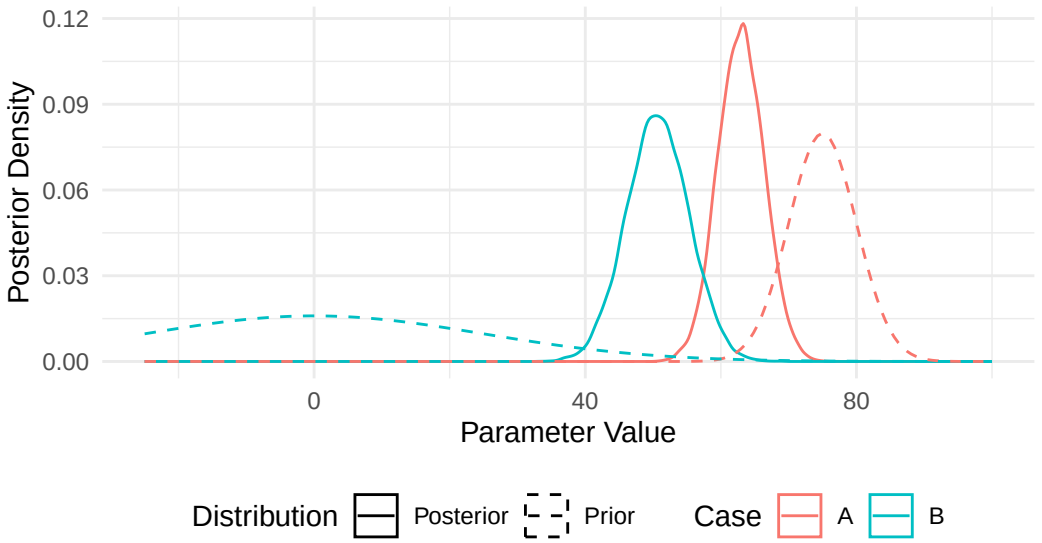
The mathematical details of fitting a Bayesian model are outside the scope of this activity. However, the basic "formula" is:

[Data + Model = Posterior]

Note that the Model contains both the formula for the model $B = mP + b + \epsilon$ and the prior distribution for the parameters m, b, σ^2. With a small dataset the posterior distribution will largely depend on the prior. With a larger dataset, the posterior will depend more on the data.

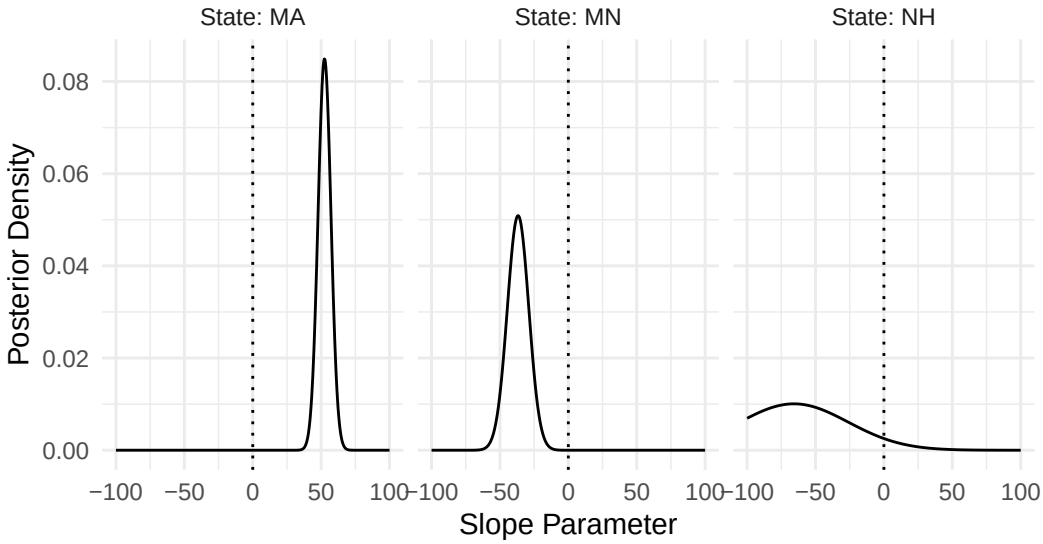
For instance, the following figure shows what happens when we use two different prior distributions for the slope when fitting the Massachusetts data:

Note that the highly uncertain Case B is shifted quite a bit more (from prior to posterior) than the Case A results. If we were to obtain a larger and larger dataset, using either prior (Case A or B) would converge to the same posterior distribution.



Three options

Bayesian approaches to statistics are particularly useful when we have limited data, as they allow us to incorporate prior knowledge. For the rest of the activity we'll consider a scenario where our access to the CEJS data is limited: Suppose we are conducting our analysis while the data are actively being gathered. In this case, we may have access to the data for some states before others. In this context, we can conduct a *sequential* Bayesian analysis by using the posterior from one analysis as the prior for a new analysis.



Pick a State

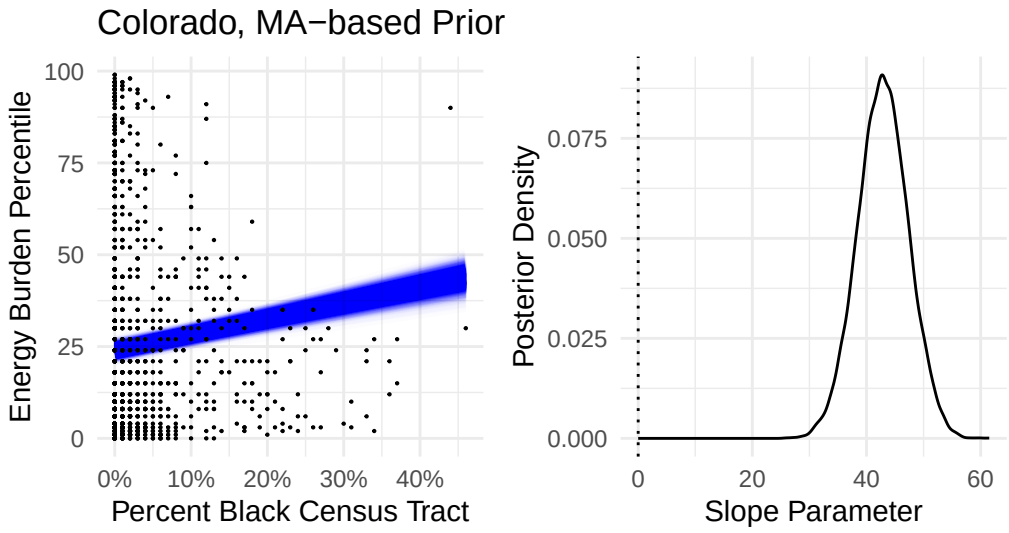
Study the posteriors above carefully; you will use this as a *prior distribution* for the slope for the rest of the activity. This means you will combine *new data* with a *prior distribution* to form a new *posterior distribution* for the model parameters. The *prior distribution* should reflect your beliefs about what you think the slope parameter should be.
Pick *one state for your group*, then come ask the instructor for your chosen state's packet.

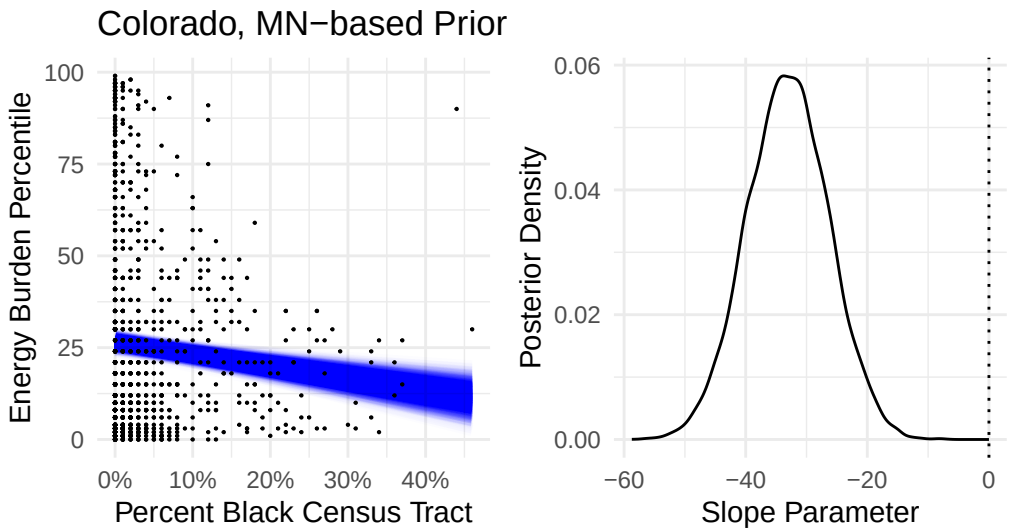
Colorado

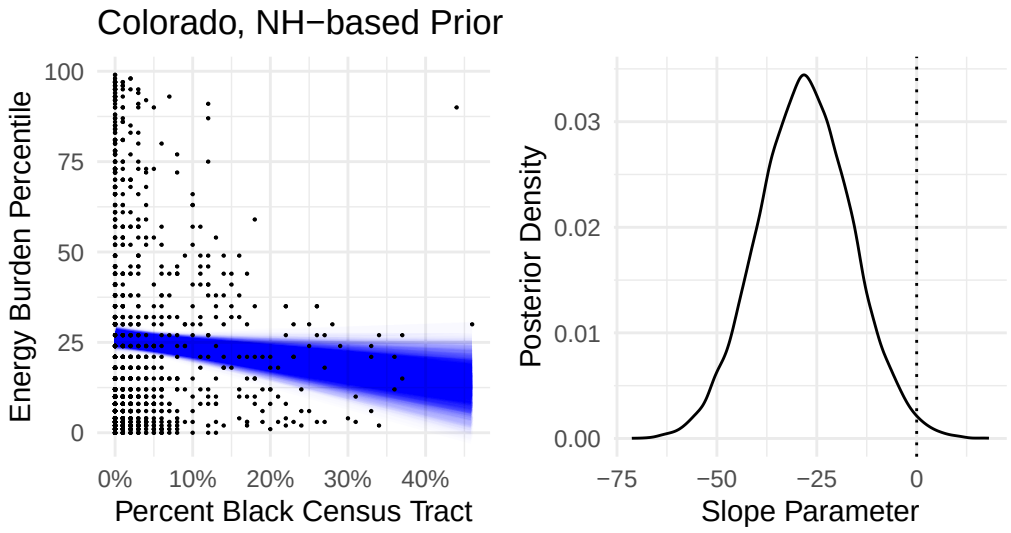
(PREP) Print CO results

Print the following three graphs and place them in envelopes labelled for the State-based prior.

MA-based Prior




Colorado, MN−based Prior
Energy Burden Percentile
100
75
50
25
0
0%
10%
20%
30%
40%
Percent Black Census Tract
Posterior Density
0.06
0.04
0.02
0.00
−60
−40
−20
0
Slope Parameter


Colorado, NH−based Prior
Energy Burden Percentile
100
75
50
25
0
0%
10%
20%
30%
40%
Percent Black Census Tract
Posterior Density
0.03
0.02
0.01
0.00
−75
−50
−25
0
Slope Parameter

MN-based Prior

NH-based Prior

Study the results

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black in Colorado?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

- For the MA-based prior, the posterior predictions do not match the data well. The posterior is extremely confident in a positive slope, which agrees with our hypothesis.
- For the MN-based prior, the posterior predictions do match the data well. The posterior is extremely confident in a negative slope, which contradicts our hypothesis.
- For the NH-based prior, the posterior predictions do match the data well. The posterior is quite confident in a negative slope, which contradicts our hypothesis.

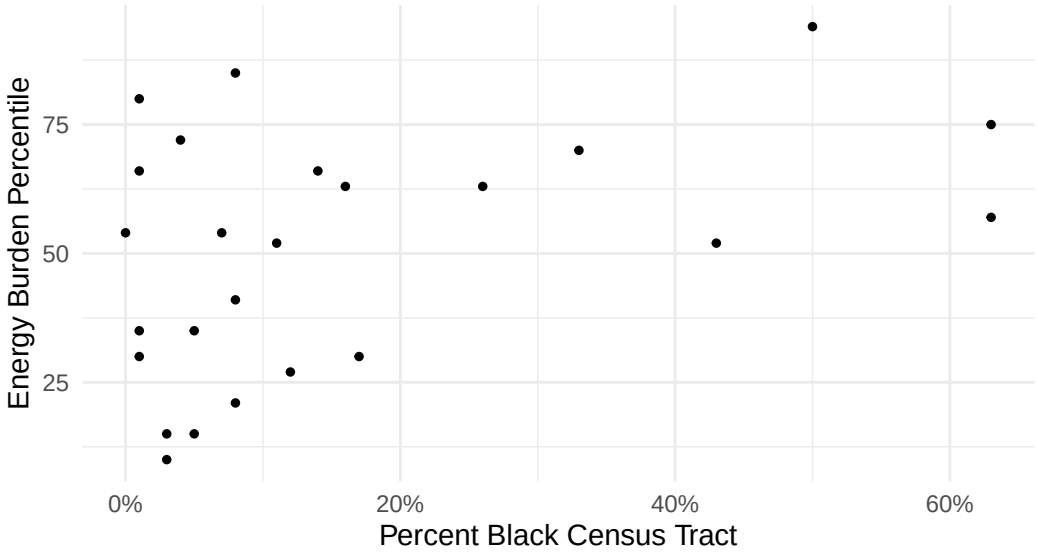
Florida

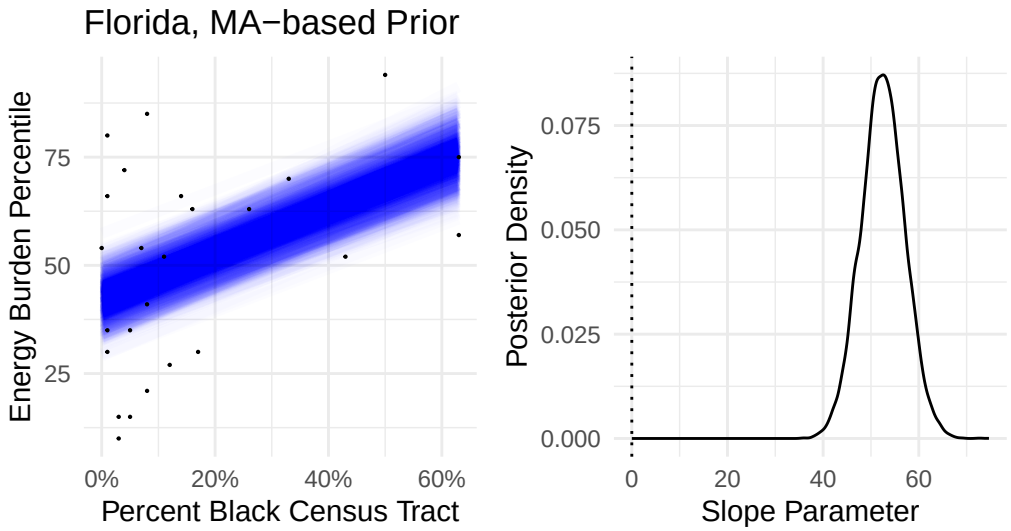
In some cases, we may find that gathering more data is simply not possible. Let's suppose that, for some reason, Florida is unwilling to provide all of their energy burden data. Therefore, we must figure out what to do with only $n = 25$ observations:

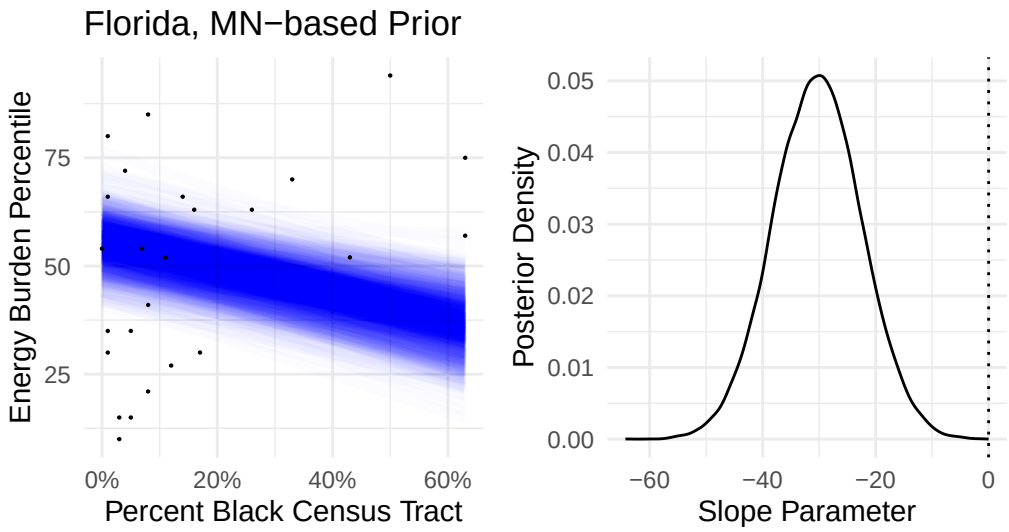
Given the limited data, our results will depend much more strongly on our prior distribution.

(PREP) Print FL results

Print the following three graphs and place them in envelopes labelled for the State-based prior.


Energy Burden Percentile
75
50
25
0%
20%
40%
60%
Percent Black Census Tract

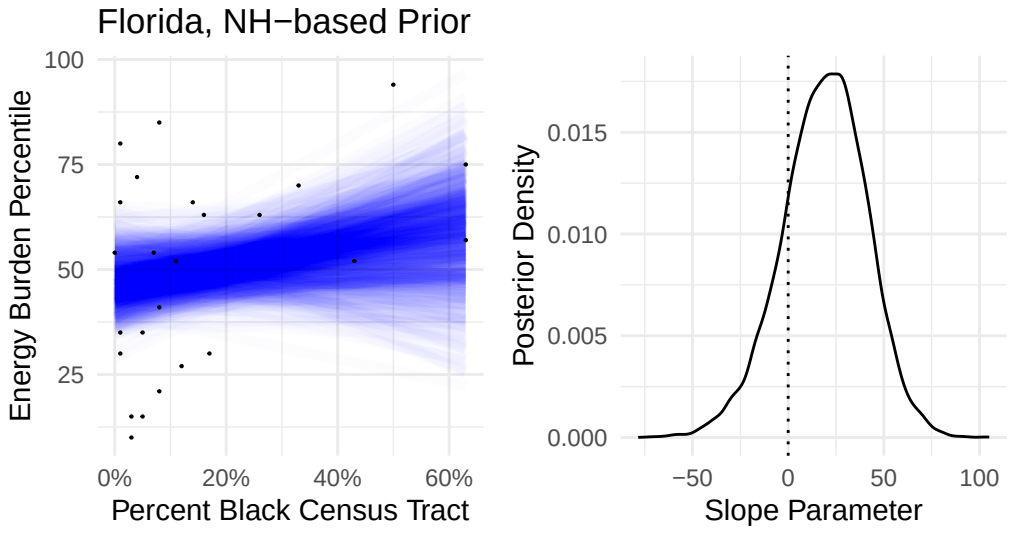

Florida, MA−based Prior
Energy Burden Percentile
75
50
25
0%
20%
40%
60%
Percent Black Census Tract
Posterior Density
0.075
0.050
0.025
0.000
0
20
40
60
Slope Parameter



MA-based Prior

MN-based Prior

NH-based Prior



Study the posterior predictions

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black in Florida?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

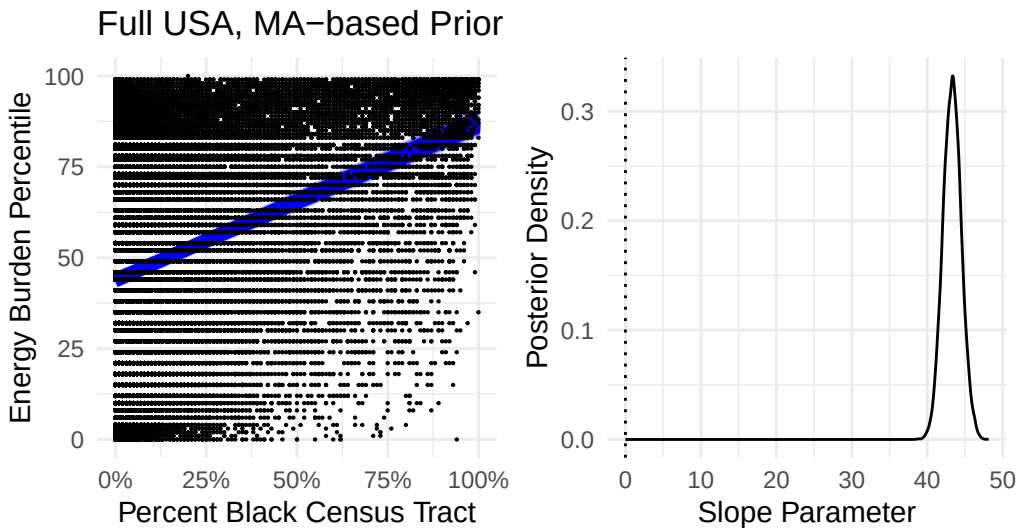
- With the MA-based prior, we would assign a high posterior probability to a positive slope
- With the MN-based prior, we would assign a high posterior probability to a *negative* slope
- With the NH-based prior, we would have an inconclusive analysis

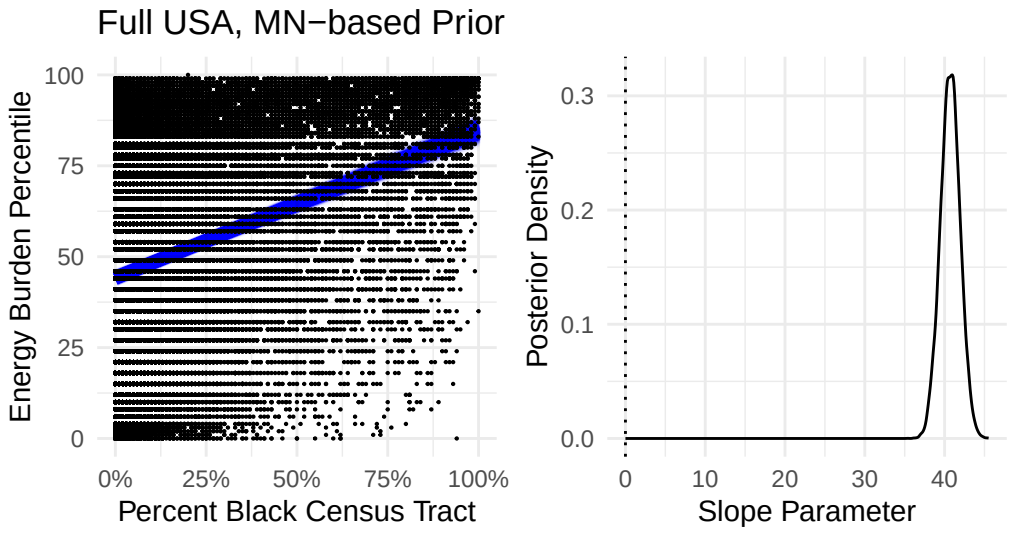
Full USA

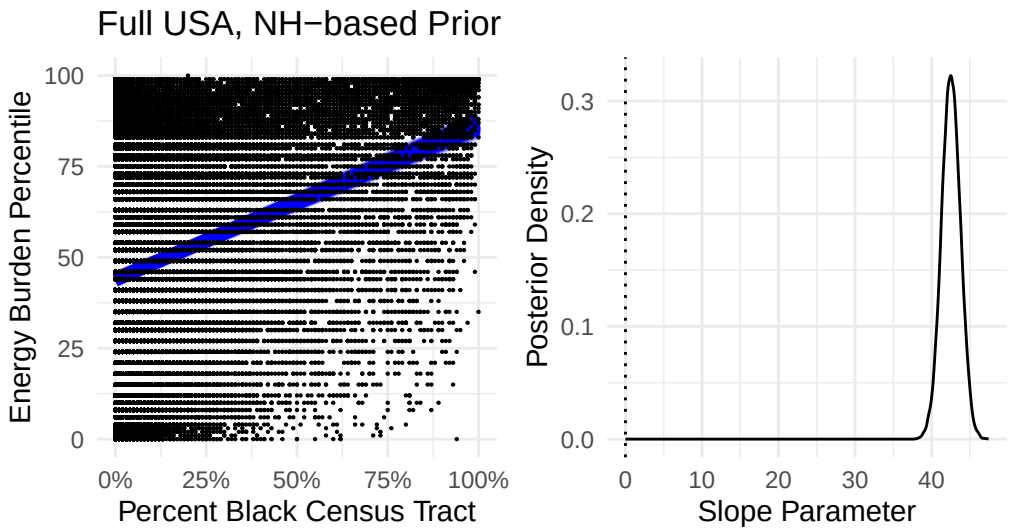
After waiting some time, we finally get access to the full U.S. CEJS dataset. With such a large dataset, we expect to see that the results will not depend so much on our choice of prior distribution.

(PREP) Print USA Results

Print the following three graphs and place them in envelopes labelled for the State-based prior.


Full USA, MA−based Prior
Energy Burden Percentile
100
75
50
25
0
0%
25%
50%
75%
100%
Percent Black Census Tract
Posterior Density
0.3
0.2
0.1
0.0
0
10
20
30
40
50
Slope Parameter


Full USA, MN−based Prior
Energy Burden Percentile
100
75
50
25
0
0%
25%
50%
75%
100%
Percent Black Census Tract
Posterior Density
0.3
0.2
0.1
0.0
0
10
20
30
40
Slope Parameter



MA-based Prior

MN-based Prior

NH-based Prior

Study the posterior predictions

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black across the whole U.S.?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

- With all three priors, we would assign a very high posterior probability to the slope being positive.