02 CEJST Activity

Instructor Note: Introduction

These are instructor notes; they will be removed from the student-facing assignment file. This is the **Frequentist** form of the activity.

Armed with all of that background knowledge on the CEJST 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.

Overview

Throughout this activity, you will be studying a statistical model fitted to data from the CEJST 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).

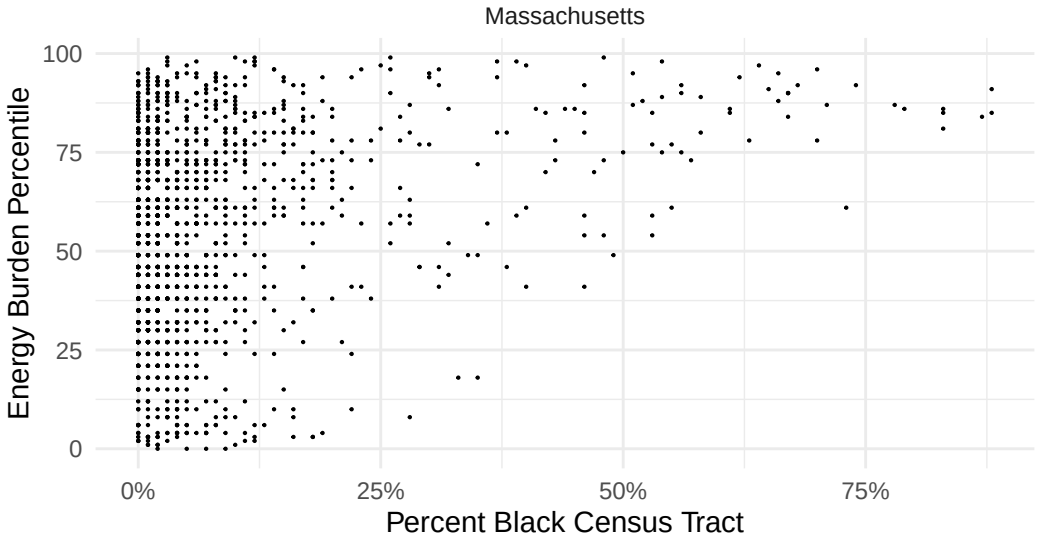

Figure 1: CEJST data for Massachusetts.

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 Frequentist Model

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

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

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.

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 enumerate 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 fixed but unknown constants.
  - This is a fundamental component of the Frequentist approach, and hence is directly related to the lesson's learning objectives.

Study the MLE estimates

Results from a Frequentist model take the form of a *maximum likelihood estimates* 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 Frequentist model to the Massachusetts dataset will result in point estimate ("best fit") values for the parameters and a *confidence interval* for each estimate. The point estimates and confidence intervals from fitting the Massachusetts data are shown below:

Term	Lower	Estimate	Upper
Intercept	49.4	50.8	52.2
Slope	43.1	52.5	61.8

A confidence interval 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 sets of point estimates and confidence intervals for the slope parameter m:

Case	**Lower**	**Estimate**	**Upper**
A	-10	60	100
B	-5	5	15
C	25	50	75

The most important thing to remember about confidence intervals is the *golden rule*....

Golden Rule for Confidence Intervals

When studying a confidence interval, we should assume the value we are trying to estimate could be **anywhere** inside the interval.

Model summaries

Questions for the Class

- Which case (A, B, or C) gives the *highest* point estimate?
  - (Write your response here):
- Which cases (A, B, or C) include 0 between their `Lower` and `Upper` values?
  - (Write your response here):
- Which case (A, B, or C) has the *narrowest* confidence interval? (i.e., the difference between `Upper - Lower` is smallest)
  - (Write your response here):
- Which case suggests the *highest confidence* that the slope parameter is positive? How can you tell?
  - (Write your response here):
- Which case (A, B, or C) gives the *highest* point estimate?
  - Case A
- Which cases (A, B, or C) include 0 between their `Lower` and `Upper` values?
  - Cases A and B
- Which case (A, B, or C) has the *narrowest* confidence interval? (i.e., the difference between `Upper - Lower` is smallest)
  - Case B
- Which case suggests the *highest confidence* that the slope parameter is positive? How can you tell?
  - Case C; it is the only CI that excludes zero.

General inference

Questions for the Class

- 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 agree`

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.

Term	Lower	Estimate	Upper
Intercept	49.4	50.8	52.2
Slope	43.1	52.5	61.8

Questions for the Class

As a reminder, our hypothesis is:

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

- For Massachusetts, does the fitted model support or contradict our hypothesis?
  - (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 confidence interval excludes zero, which suggests a high confidence in a positive slope. We can be confident at the level chosen; in our case 95%.

Study the prediction and confidence band

Frequentist analysis produces a "best fit" line, but the confidence intervals on the slope and intercept imply a *family* of lines. We call this the *confidence band* for the regression line. For

instance, the following visualizes the best fit line (solid blue) and confidence band (transparent blue) against the Massachusetts data. Practically, the confidence band tells us that any line we can draw within the bounds is *compatible* with the data we have.

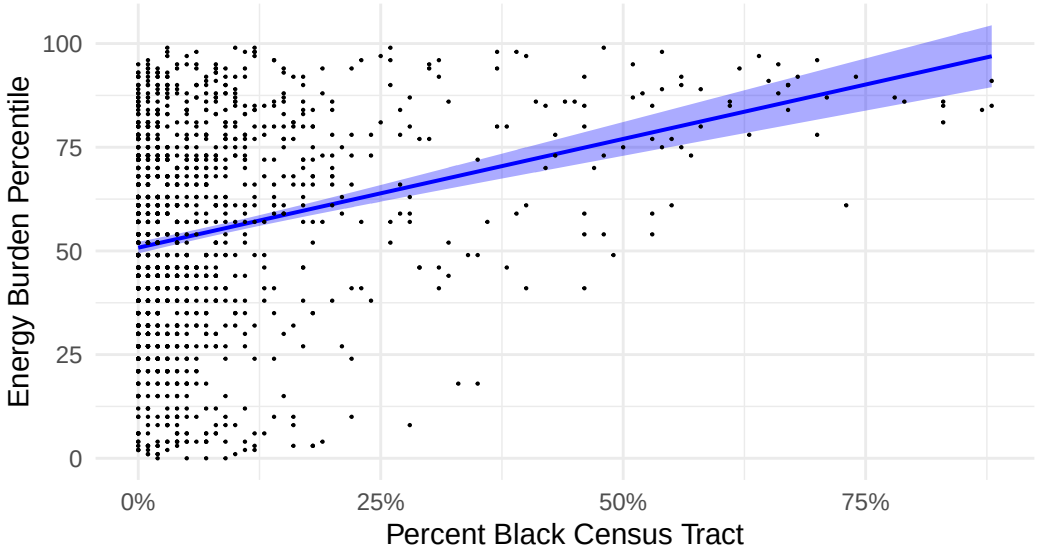

Figure 2: Best fit line and confidence band for Massachusetts data.

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

- Do all data points (black dots) land near the best fit line (solid blue line), or do some dots land far from the line?
  - (Write your response here):
- This model represents the *overall trend* in the data well. In your own words, describe how the model fits the overall trend in the data.
  - (Write your response here):
- Do all data points (black dots) land near the best fit line (solid blue line), or do some dots land far from the line?
  - No, there is variability in the data that remains unexplained by the model.
- This model represents the *overall trend* in the data well. In your own words, describe how the model fits the overall trend in the data.

- While the data points exhibit a great deal of variability, their average tends to increase from left to right. More formally, their mean conditional on the predictor tends to increase.

The model should fit the data reasonably well; otherwise, we should *distrust* its results. It doesn't matter if the MLE estimates agree with our hypothesis if the model fits the data poorly!

Data rollout

For the rest of the activity we'll consider a scenario where our access to the CEJST 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, this means we'll study some of the individual states before we study the full USA.

Armed with this fundamental understanding of statistical inference, we can now apply these ideas to study data from the other states!

Colorado

Study the results

Term	Lower	Estimate	Upper
Intercept	24.4	26.1	27.8
Slope	-48.9	-24.3	0.3

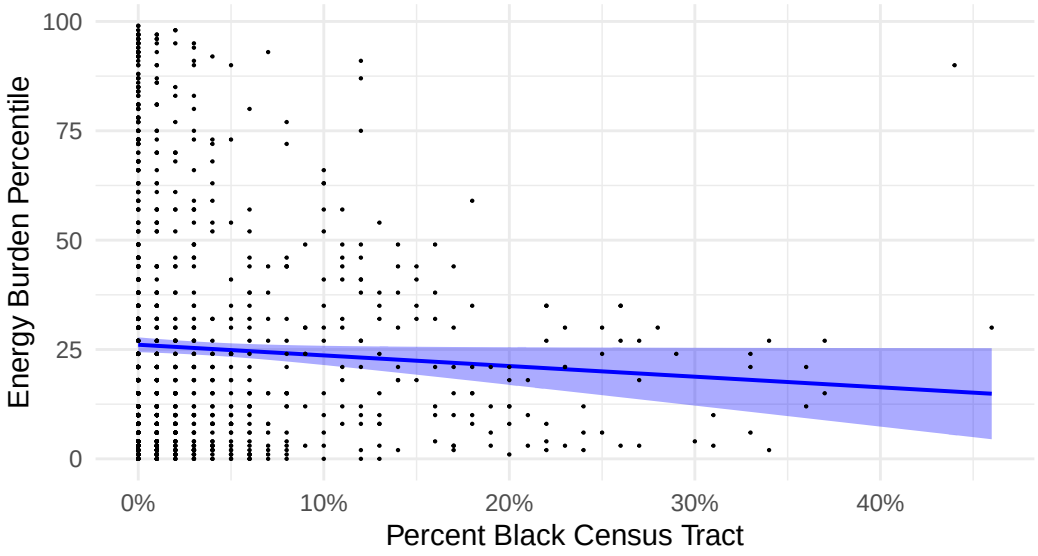

Figure 3: Best fit line and confidence band for Colorado data.

Studying Model Results

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

  - The model gives a negative MLE for the slope, but the CI includes zero.
- How well do the model predictions match the data?
  - There is much scatter around the line, but there does seem to be an overall trend downwards.
- How confident are you in your conclusion?
  - Not very confident, as the slope CI includes zero.

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:

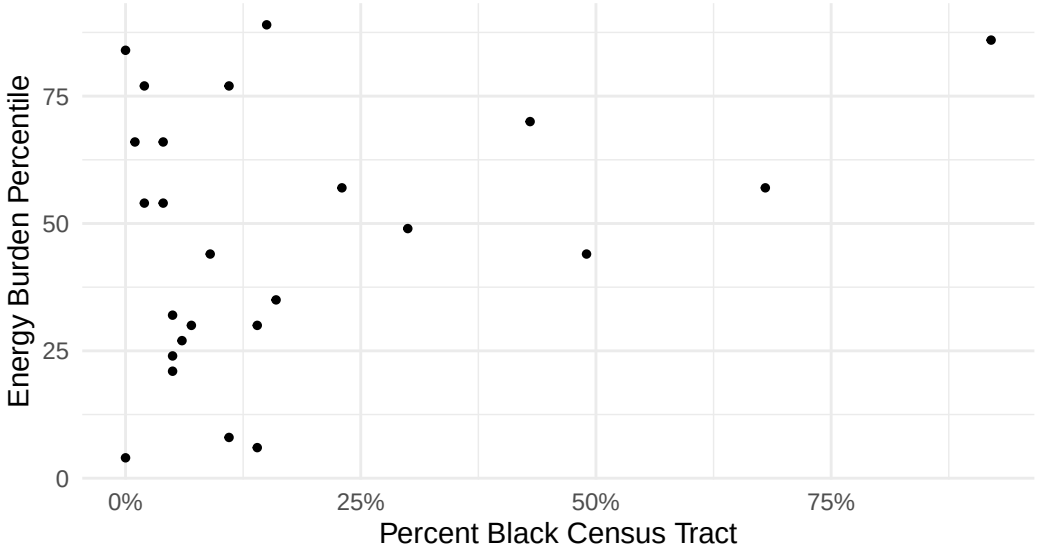

Figure 4: Limited data for Florida.

Given the limited data, we'd expect more uncertainty in our estimates. This will be reflected as a wider confidence interval.

Study the results

Term	Lower	Estimate	Upper
Intercept	29.1	41.7	54.3
Slope	-10.4	33.9	78.2

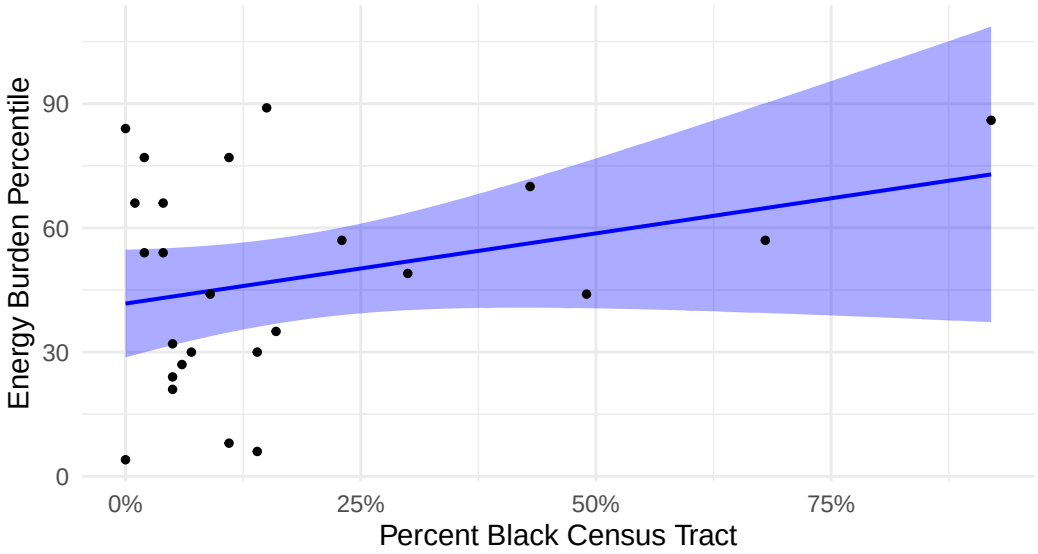

Figure 5: Best fit line and confidence band for Florida data.

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black in Florida?
  - (Write your response here):
- How well do the model predictions match the data?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):
- What does the model suggest about the trend of energy burden with percent Black in Colorado?
  - The model gives a positive MLE for the slope, but the CI is wide and includes zero.

- How well do the model predictions match the data?
  - There is much scatter around the line, but there does seem to be an overall trend downwards.
- How confident are you in your conclusion?
  - Not very confident, as the slope CI includes zero.

Full USA

After waiting some time, we finally get access to the full U.S. CEJST dataset.

Study the results

Term	Lower	Estimate	Upper
Intercept	43.4	44.0	44.7
Slope	41.0	43.5	46.0

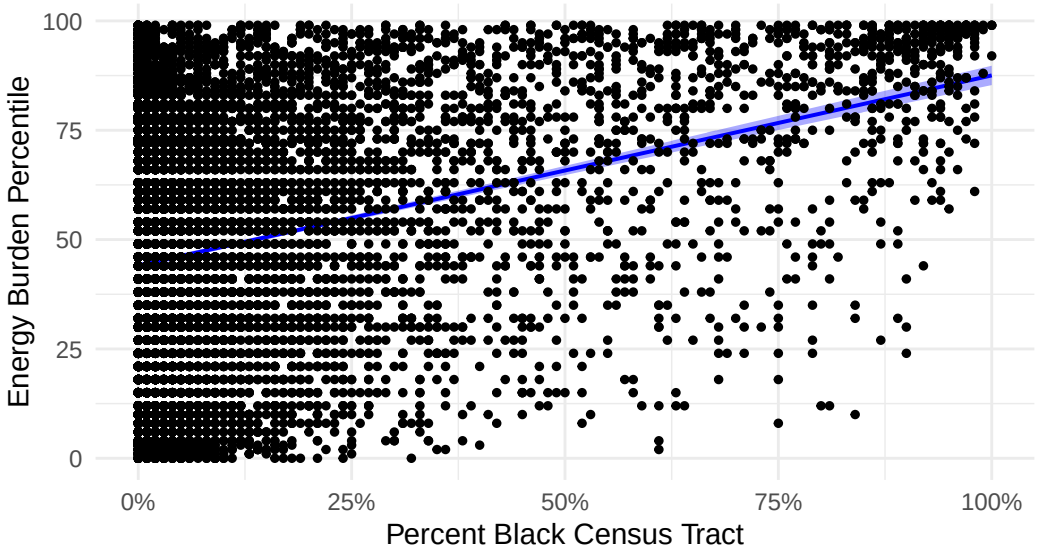

Figure 6: Best fit line and confidence band for full USA data.

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black in the full USA?
  - (Write your response here):
- How well do the model predictions match the data?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

- What does the model suggest about the trend of energy burden with percent Black in the full USA?
  - The model gives a positive MLE for the slope, the CI is narrow and excludes zero.
- How well do the model predictions match the data?
  - There is considerable scatter around the trend, much more than what we've seen before. However, there does appear to be a clear upward trend.
- How confident are you in your conclusion?
  - Quite confident, as the slope CI excludes zero.

Coming up next: When the instructor tells you, please start working through `04 Activity Wrap-up` with the other students at your table. **Please do not move on until your instructor tells you to.**

(This page intentionally left blank)