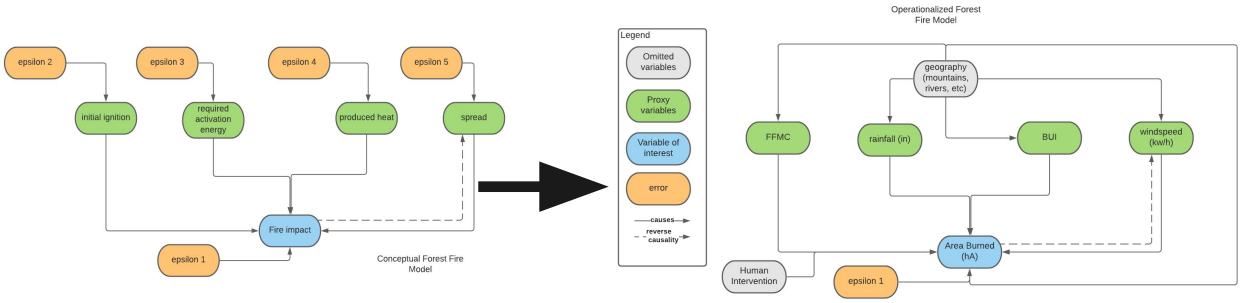
It's Lit! Students of Fire 🔥

Explanatory Research to Uncover the Variables that Cause More Severe Forest Fires

Team: Savita Chari, Denny Lehman, Tymon Silva

Executive summary:
We tried to explain the cause of forest fires. We found wind to be statistically significant in causing fires. Our dataset was very difficult to use for regression analysis, and the original researchers even call this out in the data dictionary attached to our dataset. We call into question our large model assumptions might not be appropriate and suggest alternatives to linear regression.

Background and Conceptualization



Denny - shoot for 1 min

In our study we attempted to explain the causes of forest fires. Our hypothesis was that atmospheric and forest floor conditions cause greater area burn. Using linear regression, we found that wind was statistically significant in explaining forest fire size. Our data was very challenging to use.

Forest fires in the pacific northwest have gotten out of control. The largest fires in centuries have all occurred in the last ten years. If we could better understand what causes fires, we could better prepare our communities and mitigate damage.
In this study, we successfully built an explanatory model of the concepts that cause forest. We will go through our research process in this presentation.

Our research started off with building a conceptual causal model. We theorized that the components of the combustion reaction, namely initial ignition, low activation energy of fuel, high produced heat from fuel, and conditions that cause fire spread, would cause larger fires. I'll let tymon tell you about how we operationalized these concepts and the data that we used.

Problem statement
Overview - we tried to explain what causes forest fires to spread. Our data was very challenging to work with. We found one variable that showed significance, which is

exciting!

Dataset and Methods

- Dataset
  - Montesinho National Park in Portugal
  - January 2000 to December 2003
  - 517 Records, 13 Columns
- Methodology
  - Linear regression

Concept	Proxy Variable	Units	Range	Description	Source
Fire impact	Area Burned	Hectares (ha)	0.00 to 1090.84	the burned area of the forest	Portugal Dataset
Initial ignition	Fine Fuel Moisture Code (FFMC)	Unitless index	18.7 to 96.2	Represents fuel moisture of forest litter fuels	Portugal Dataset
Required activation energy	Rainfall	millimeters per square meter (mm/m^2)	0.0 to 6.4	Current rainfall in the environment	Portugal Dataset
Produced heat	Build Up Index (BUI)	Unitless index	2.2 to 315.6	Represents the total amount of combustible fuel	Calculated Variable
Spread	Wind speed	Kilometer per hour (km/h)	0.40 - 9.40	Current wind speed	Portugal Dataset

Tymon

- The dataset came from
- It has X many records and Y columns
- Many of the variables are numeric

What models are we applying

- Causal approach to explain forest fires
- Linear regression
- Coefficients will explain the concepts

Modeling and Our Best Pick

- **Model 1 <- lm(log(area +1) ~ log(BUI))**
  - BUI represented a variable that satisfies causal theory
  - Showed No Significant coefficients
  - Failed to reject the null hypothesis
- **Model 2 <- lm(log(area +1) ~ log(BUI)) + wind + rain_binary + FFMC**
  - Included all of the variables that, we had theorized in the causal diagram
  - Only wind was a significant predictor
  - Failed to reject the null hypothesis
- **Model 3 <- lm(log(area +1) ~ wind)**
  - Learnings from the second model
  - With a p-values < 0.05 we reject the null hypothesis

	Dependent variable:		
	log(area + 1)		
	(1)	(2)	(3)
log(BUI)	0.024	0.095	
	(0.087)	(0.105)	
wind		0.096*	0.083*
		(0.042)	(0.041)
rain_binary		-0.637	
		(0.581)	
FFMC		-0.008	
		(0.016)	
Constant	0.947*	0.963	0.735***
	(0.416)	(1.267)	(0.178)
Observations	361	361	361
R2	0.0002	0.017	0.011
Adjusted R2	-0.003	0.006	0.008
Residual Std. Error	1.402 (df = 359)	1.396 (df = 356)	1.394 (df = 359)
F Statistic	0.080 (df = 1; 359)	1.509 (df = 4; 356)	4.083* (df = 1; 359)
Note:			*p<0.05; **p<0.01; ***p<0.001

Savita

Revisiting the causal question: Our research is trying to find out, if a statistical significant relationship exists between certain environmental conditions on the size of forest area burnt.

As mentioned in the Data section, This dataset was not easy. we log transformed **area** and it is also our outcome variable.We followed the following guidelines while deciding on the data models

1. It should follow our causal diagram and insights from the EDA
2. We should start with a simple model and observe the findings.
3. Then add complexity, as guided by our causal diagram and the learnings from the previous models

Based on these principals we came up with the following 3 models

1. Our first model simply contains only one variable, BUI.

   Finding: Our null hypothesis is that log(BUI) has no effect on area burned. With a p-value > 0.05, we fail to reject the null hypothesis. We cannot conclude that log(BUI) causes forest fires, which is a serious blow to our theory.

2. For our second model we chose to include all of our variables that. With the additional variables, this model is more complex than the first.

   Finding: All variable coefficients except for wind had p values > 0.05. So overall we reject the null hypothesis.

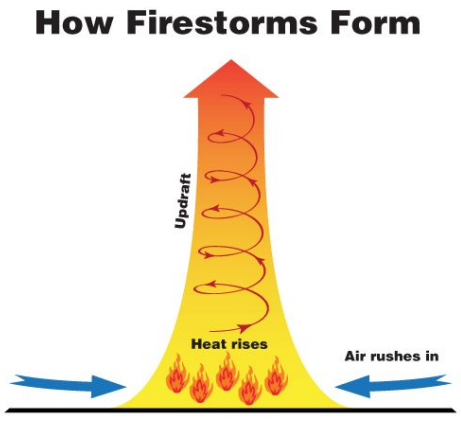
3. Our learnings from the second model led us to our third and best model. So for our third model, we tested if wind alone makes for a better model than the

1. others because wind was the only statistically significant coefficient in model 2 **Findings**: As expected, we reject the null hypothesis for this coefficient.Finally, to compare the three models, we use the f-test. Our f statistic will look at the change in variance as the models change the number of variables. Our null hypothesis is that the addition of variables will have no significant effect on the explained variance of the model.

Overall Finding:
Model 3 had an f-statistic of 4.083. When converted to a p-value, we find that it is statistically significant. With a p-value < 0.05, Therefore, we select this model as our best explanation of forest fire area burn.

Concerns and limitations

- Statistical Limitations
  - IID Violation
  - Unique BLP Exists Concerns
- Structural Limitations
  - Reverse Causality
  - Omitted Variable Bias
    - Human Intervention
    - Geography - Natural barriers or enhancers to fire spread
- Other Limitations
  - Dataset Problematic - suggestion to use log



In this section, I'll talk as much as I can about the concerns and limitations of the model until I run out of time. To conclude up to this point,
We found significance that wind causes forest fires. Our data was incredibly challenging to work with. There is some evidence that
Denny 3 min
Our data had concerns by the original researchers that collected it
Data was collected over 3 years from one national forest in portugal.
Therefore, there is concern about the independence of these data points.
Fires from March burn up all of the forest floor, affecting the size of a fire in April.

There could be geographic clustering of forest fires too.

BLP concerns
Both Rain, an input variable and area burned, our outcome variable were highly skewed. So while a unique BLP exists, the BLP is hyper sensitive to our outlier datapoints. When using the large model assumptions, we expect that data points over 100 will provide us enough variance in the features to make good models, but in our case,

Notes -
Explain that rainfall and FFMC are not related

Thank you for your time and attention.