# Lab 2: How to Predict the Size (in Hectares) of a Forest Fire

w203: Tymon Silva

November 28, 2021

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
library(ggplot2)
library(lmtest)
library(sandwich)
library(stargazer)
```

## Introduction

In the Pacific Northwest, drought and an increase in human water consumption have caused the most severe forest fires in centuries. For our group - this is personal - each one of us has been personally affected by recent fires. But not all forest fires are the same - atmospheric weather and moisture content in the forest floor will affect the size of the burn. In this causality project, we hope to uncover the conditions that limit the severity of forest fires with regression modeling. If we could predict the conditions that cause large fires we can better prepare our communities.

"Your introduction should present a research question and explain the concept that you're attempting to measure and how it will be operationalized. This section should pave the way for the body of the report, preparing the reader to understand why the models are constructed the way that they are. It is not enough to simply say,"We are looking for product features that enhance product success." Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also a good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data."

## Research Question

How might we predict the size of a forest fire based on wind, temperature, humidity, and other variables.

## Research Design

After you have presented the introduction and the concepts that are under investigation, what data are you going to use to answer the questions? What type of research design are you using? What type of models are you going to estimate, and what goals do you have for these models?

1. *What do you want to measure*? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications. How are the variables that you will be modeling distributed? Provide enough context and information about your data for your audience to understand whatever model results you will eventually present.

2. What covariates help you achieve your modeling goals? Are there problematic covariates? either due to *collinearity*, or because they will absorb some of a causal effect you want to measure?
3. What *transformations*, if any, should you apply to each variable? These transformations might reveal linearities in the data, make our results relevant, or help us meet model assumptions.
4. Are your choices supported by exploratory data analysis (*EDA*)? You will likely start with some general EDA to *detect anomalies* (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to *guide* your decisions. You can also leverage statistical *tests* to help assess whether variables, or groups of variables, are improving model fit.

At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects.

At a minimum, you need to estimate at least three model specifications:

The first model you include should include *only the key variables* you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute minimum number of covariates (usually zero or one covariate that is so crucial it would be unreasonable to omit it).

Additional models should each be defensible, and should continue to tell the story of how product features contribute to product success. This might mean including additional right-hand side features to remove omitted variable bias identified by your casual theory; or, instead, it might mean estimating a model that examines a related concept of success, or a model that investigates a heterogeneous effect. These models, and your modeling process should be defensible, incremental, and clearly explained at all points.

Your goal is to choose models that encircle the space of reasonable modeling choices, and to give an overall understanding of how these choices impact results.

# Data

```
fire_data <- read_csv(file = '../src/data/forestfires.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X = col_double(),
##   Y = col_double(),
##   month = col_character(),
##   day = col_character(),
##   FFMC = col_double(),
##   DMC = col_double(),
##   DC = col_double(),
##   ISI = col_double(),
##   temp = col_double(),
##   RH = col_double(),
##   wind = col_double(),
##   rain = col_double(),
##   area = col_double()
## )
```

```
# str(fire_data)

# unique(fire_data$month)
# unique(fire_data$temp)
```

```
# unique(fire_data$wind)
# unique(fire_data$rain)
# unique(fire_data$area)

# subset(fire_data, area > 1000)
# subset(fire_data, area > 500)
# subset(fire_data, area < 20)
# subset(fire_data, area < 5)

# ggplot(fire_data, aes(x=reorder(month, month, function(x)-length(x)))) +
#   geom_bar(fill='red') +
#   labs(x='Team')
# hist(fire_data$temp)
# hist(fire_data$wind)
# hist(fire_data$rain)
# hist(fire_data$area)
```

```
# ggplot(d, aes(x = average_rating, y = views)) +
#   geom_point() +
#   geom_smooth() +
#   ggtitle("average rating vs views")
#
# ggplot(d, aes(x = average_rating, y = log(views) )) +
#   geom_point() +
#   geom_smooth() +
#   ggtitle("average rating vs log(views)")
#
# ggplot(d, aes(x = log_of_average_rating, y = log(views) )) +
#   geom_point() +
#   geom_smooth() +
#   ggtitle("average rating vs log(views)")
```

# 4. Results

You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table.

In your text, comment on both *statistical significance and practical significance*. You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

### 5. Limitations of your Model

**5a. Statistical limitations of your model** As a team, evaluate all of the large sample model assumptions. However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Note that you may need to change your model specifications in response to violations of the large sample

model.

**5b. Structural limitations of your model** What are the most important *omitted variables* that you were not able to measure and include in your analysis? For each variable you name, you should *reason about the direction of bias* caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting. What data could you collect that would resolve any omitted variables bias?

# 7. Conclusion

Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.