GLM Project Proposal: Bombs Dropped during World War Two

Connor Putnam

11/17/2020

Dataset Overview

For the final project I choose to address generalized linear models when it comes to working with Big Data. I found an interesting dataset online that covers all of the aerial bombings conducted during World War Two. The dataset contains 178,281 observations, each indicating a single bombing mission between the May 15, 1940 and May 2, 1945. This data only covers Allied operations an thus this is why the records begin approximately eight months after the German advancement into Poland. With final observations coming a few days before the German surrender. The data set also contains 46 variables.

Question of Interest:

What factors lead to the amount of bombs dropped on a mission in WW2?

I will set the response variable to be the Total Weight of bombs dropped during a single mission. Some of the variables are not relevant to answering this question so I will discard those. Additionally, many of the variables are categorial variables and have many different factor levels. In order for my analysis more interpretable I might have to perform some data cleaning and categorization.

Overall Process

Step 1, Wrangle Data:

This dataset has a lot of observations and with that comes a level of messiness that needs to be cleaned up.

Step 2, Selection:

Due to the size and nature of this data there is a lot of imputation that needs to be done. This will be performed using statistical methods as well some basic information on how WW2 was fought (This is important if trying to impute any categorical variables, not sure if that will be possible most of the time). Then I will need to pick variables that are of interest in explaining my research question.

Step 4, Model Selection:

Sense this is count data I will start by implementing a Poisson regression and then check the fit. If the fit is poor, I will try to remedy it and possibly use a different model.