GLM Project Report

Connor Putnam

12/6/2020

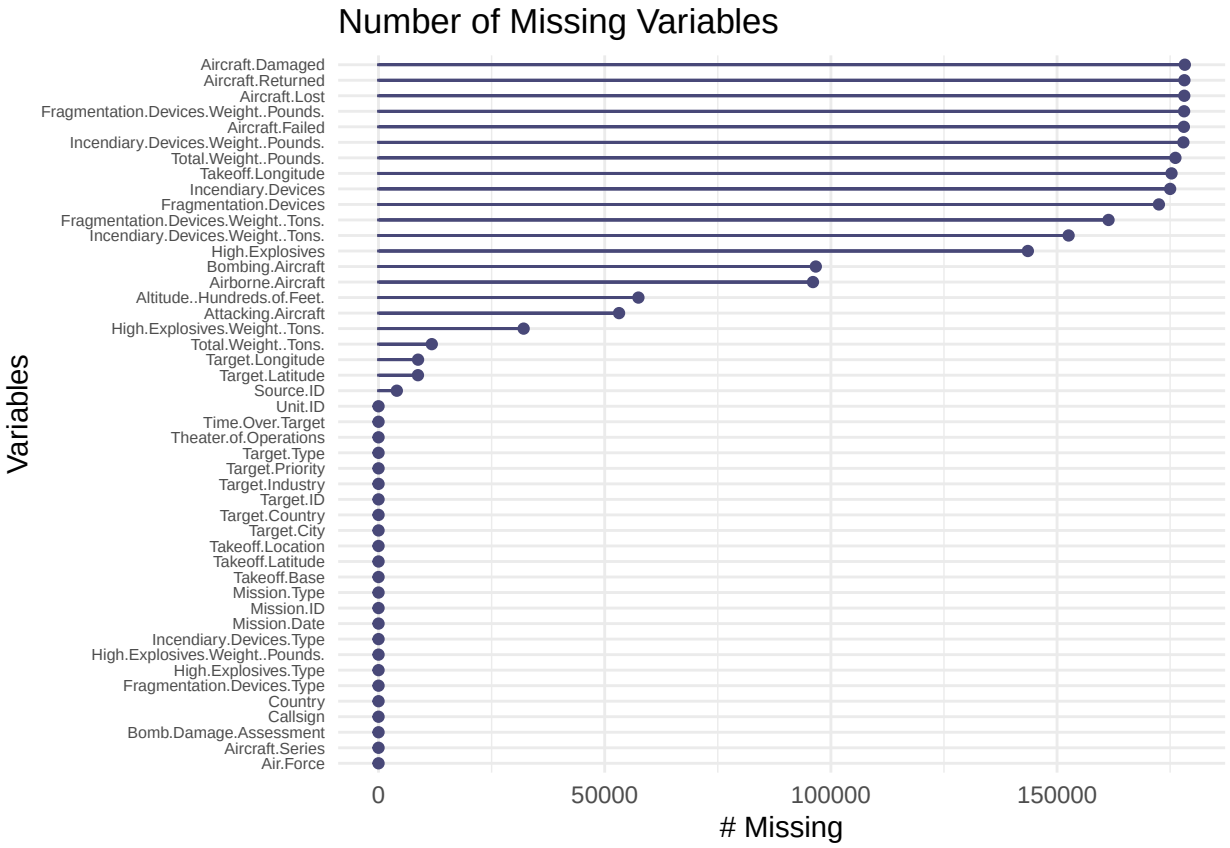
Overview

The goal of this project was to model the amount of bombs drops on a given mission during World War 2. When setting out to do this project I wanted to implimentent generalized linear models as per the assignemnt requirements, but I also what to explore between generalized linear models and big data. In order to get my toes wet with this I decided abone this this Arieal Bombing Dataset that can be found at https://www.kaggle.com/usaf/world-war-ii. 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. As with many large data sets there was alot of missing variable, this can also be explained in part becasue this is historical data during a time of war. The plot below shows that many of the variables have too many missing values and thus will have to be dropped from potential analysis.



After addressing this the next step was to impute and as well decide what variable to keep for analaysis. Based on the ability to impute as well as the relevence to myt quesiton of interest I decided on the following varibales for analysis. The table below list out variables being used after data wrangling/cleaning as well the first six obervations for those values.

Industry	Mission.ID	Date	Theater	Target_Country	Country
CITIES TOWNS AND URBAN AREAS	12	1943-08-15	ETO	GERMANY	GREAT BRITAIN
CITIES TOWNS AND URBAN AREAS	13	1943-08-15	ETO	GERMANY	GREAT BRITAIN
CITIES TOWNS AND URBAN AREAS	58	1943-08-15	ETO	GERMANY	GREAT BRITAIN
SHIPS	66	1943-08-15	MTO	ITALY	USA
SHIPS	67	1943-08-15	MTO	ITALY	USA
HIGHWAYS AND VEHICLES	68	1943-08-15	MTO	ITALY	USA

Total_Weight	Altitude	Attacking_Aircraft	Bombing_Aircraft	System_Time	Alliance	Industry_Type
1	250	11.462295	11.462295	-832550400	Axis	City
4	250	5.000000	5.000000	-832550400	Axis	City
87	135	11.462295	11.462295	-832550400	Axis	City
2	95	16.857143	16.857143	-832550400	Axis	Non-City
2	95	16.857143	16.857143	-832550400	Axis	Non-City
17	95	6.733871	6.733871	-832550400	Axis	Non-City

Variable_Names	Definition
Attacking_Aircraft	Number of Attacking Aircraft(Non-Bombers)
Bombing_Aircraft	Number of Bombing Aircraft
Altitude	How high the plans flew on a given mission, in hundreds of fee
System_Time	Date in which the mission occured converted into computer system time
Theater	Theater of warfare
Alliance	Clasifies a county's Alliance as either Axis(Germany, Italy, etc) or Non Axis(Allies:USA, Britian, USSR, etc
Industry_Type	Clasifies bombing targets as either city targets or industrial/warfare targets

Poisson Regression

After looking at the counts of bomb droped I devided to process with the following Poisson regression,

```
glm(Total_Weight ~ Attacking_Aircraft + Bombing_Aircraft + Altitude + System_Time +
Theater + Alliance, data = ww2, family = poisson(link = "log"))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	**11.1714**	**0.7074**	**15.7914**	**0.0000**
Attacking_Aircraft	**0.0782**	**0.0343**	**2.2789**	**0.0227**
Bombing_Aircraft	-0.0584	0.0343	-1.7020	0.0888
Altitude	**0.0005**	**0.0000**	**41.4953**	**0.0000**
System_Time	**0.0000**	**0.0000**	**532.1157**	**0.0000**
TheaterETO	**2.3432**	**0.7071**	**3.3138**	**0.0009**
TheaterMTO	**1.7116**	**0.7071**	**2.4205**	**0.0155**
AllianceNon-Axis	**0.1848**	**0.0018**	**100.5076**	**0.0000**

This resulted in some interesting findings regarding signifigance but ulitmately cannot be used becasue the deviance is clocking in at 3304409.942359 on 62334 degrees of freedom. So there is clearly an issue with overdispersion at play. The next time to is to take the quasi poisson but that also resulted in a similiar overdisprerion issue. Addtioanlly even with vearious transformations should as inverse and logarithmic is there was no meaningful improvement in this model.

Negative Binomial

Sense both the Poisson and the Quasi-Poisson reuslted in over disperion I decided to try my luck on a negative binomial model instead. The first negative binomial model I choose was the following:

```
glm.nb(Total_Weight ~ Attacking_Aircraft + Bombing_Aircraft + Altitude + System_Time +
Theater + Alliance + Industry_Type, data = ww2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	**14.4039**	**1.0506**	**13.7100**	**0.0000**
Attacking_Aircraft	0.0504	0.1039	0.4846	0.6279
Bombing_Aircraft	0.0269	0.1039	0.2585	0.7960
Altitude	**0.0011**	**0.0001**	**13.0162**	**0.0000**
System_Time	**0.0000**	**0.0000**	**134.2197**	**0.0000**
TheaterETO	1.0973	1.0450	1.0501	0.2937
TheaterMTO	0.6567	1.0450	0.6284	0.5298
AllianceNon-Axis	**0.4434**	**0.0120**	**36.9241**	**0.0000**
Industry_TypeNon-City	**-0.8505**	**0.0117**	**-72.4741**	**0.0000**

This resulted in a much better deviance value of 68687.55 on 62333 degrees of freedom. This level off overdispersion is much more managebale, esspeically considering the size of the data. The overdispersion

ration sits at approximaley 1.1 Now the next step was to to see if I could improve upon this in terms of AIC values and even less overdisperison.

The first proposed negative binomial model had an AIC value of 502674.41. I was able to come some imporvement based by trying out difference transformations. The best I was able to get was the following model:

```
glm.nb(Total_Weight + 1 ~ log(Attacking_Aircraft) + log(Bombing_Aircraft) + Altitude +
System_Time + Theater + Alliance + Industry_Type, data = ww2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	**11.9459**	**0.8183**	**14.5984**	**0.0000**
log(Attacking_Aircraft)	1.1973	1.4991	0.7986	0.4245
log(Bombing_Aircraft)	-0.3538	1.4991	-0.2360	0.8134
Altitude	**0.0007**	**0.0001**	**9.6967**	**0.0000**
System_Time	**0.0000**	**0.0000**	**121.2871**	**0.0000**
TheaterETO	-0.2023	0.8130	-0.2488	0.8035
TheaterMTO	-0.7031	0.8130	-0.8647	0.3872
AllianceNon-Axis	**0.3561**	**0.0102**	**34.9930**	**0.0000**
Industry_TypeNon-City	**-0.6557**	**0.0100**	**-65.7406**	**0.0000**

This dropped the AIC value down by 3144.91 to 499529.5. Additionly the devieince to degrees of freedom ratio was slightly reduced from 1.1 to 1.06. This transfomraiton is notworthy but the improvement is minimal and might just over complicate interation of the model for little reward in terms of predictability.

Conclusions

After performing this analysis I concluded that the best model was not a Poisson but instead a negative binomial. Through tranfomations I was also able to find a model that made modest improved both in terms of AIC and the overdispersion ratio. THe findings are that the Altitude the planes were flying at, when the bombing mission took place, the targets Alliance, and the industry type, were all very signifigant in terminaiting the amount of bombs dropped. The `System_time` variable signigicant was not surprising to me becasue I was aware before going into this that more bombs were dropped by the US and Great Britian later on in the war. Similiar a planes altiude is a huge determinate on how much wait it can hold, so that was not too surprising. I was surpried by the level of signigivan tin bothe `Alliance` and `Industry_Type`. I figure they would be improtant in predicting but not this important, my prior was that during world war two the lines became pretty blurred between what was an enemy target and what was a civilian target. THe `Theater`, `Bombing_Aircraft` and `Attacking_Aircraft` varibales were not signigiant. The `Theater` variables insigfigance did not surprised me much bit the other two did. I was expecting the ampoung of aircraft involved during a mission to be highly predictive in terms od the amount of bombs dropped during a mission. But this does not seem to be true.

Overall I found the findings to be interesting and I learned alot along the way in terms of applying GLM models to big data set. Hope the reader finds this report interesting as well. Thank you!