Final Report

Zihuan Qiao, Anna Cook, Yinfeng Zhou, Zixuan Liu, Jiaheng Li

3/25/2021

Introduction

Project Background

The client is researching rare disease advocacy organizations. In particular, she is interested in comparing organizations for rare vs. ultra-rare diseases along several different factors including funding, patient outreach, research, etc. in order to assess the organizations' need for support. The main research questions is, how do the prevalence of diseases (rare vs. ultra-rare) and the size and age of the organization, affect the outcomes of the advocacy organization? (priorities, funding, patient outreach, research, etc.)? For example, the client may expect to find that organizations for less rare diseases may have more resources, better patient outreach, etc. while more rare diseases may have fewer resources available and may need more support. In addition to this main research focus, the client is interested in exploring the data more generally and looking for patterns that may arise. The client is seeking advice from the MSSP team about how to effectively recode variables as necessary, conduct an initial exploratory data analysis, and determine an appropriate regression model in order to make the comparisons of interest.

Variable Description and Data Processing

The client has collected survey data from 217 different organizations' leaders or representatives, located in various locations worldwide, with one response per organization. The survey includes questions related to demographic data, budget/funding, disease prevalence, research efforts, etc.

There are three independent and three dependent variables which our analyses will focus on. The independent variables are organization size, organization age, and disease frequency (1 case per x births). We will refer to these as Size, Age, and Frequency, respecively. The dependent variables are organizations' top priority, whether research efforts are handled internally or externally, and whether clinical trials are in progress for a therapy/treatment or not. We will refer to these as Priority, Research, and FDA Therapy, respectively.

The first step in processing the data was to bucket the independent variables into discrete levels based on the client's literature review and judgments. Size has three levels: Small (0-300 members), Medium (300-1000 members), and Large (1000+). Age has two levels: Older (10+ years) and Younger ($<$ 10 years). Frequency has three levels: Rare (more frequent than 1 in 200,000 births), Ultra-Rare (less frequent than 1 in 200,000 births), and Unknown, although we are only focused on the Rare and Ultra-Rare levels for the sake of the analysis, so any responses of "Unknown" were removed.

EDA

In order to get a general sense of the dataset, we began with an exploratory data analysis (EDA). First, we plotted histograms of the independent variables (Size, Age, and Frequency). The histograms are shown in Figure 1. From these plots, we can see that there is some imbalance in the data. First, it appears that the organizations for ultra-rare diseases tend to be smaller and younger, whereas the rare disease organizations are more uniformly represented across the different size and age categories. Second, ...

Next, we plotted histograms of the dependent variables of interest (Priority, Research, and FDA Therapy). From these plots, we can see further imbalance in the data (see Figures 2-4). When looking at the priority

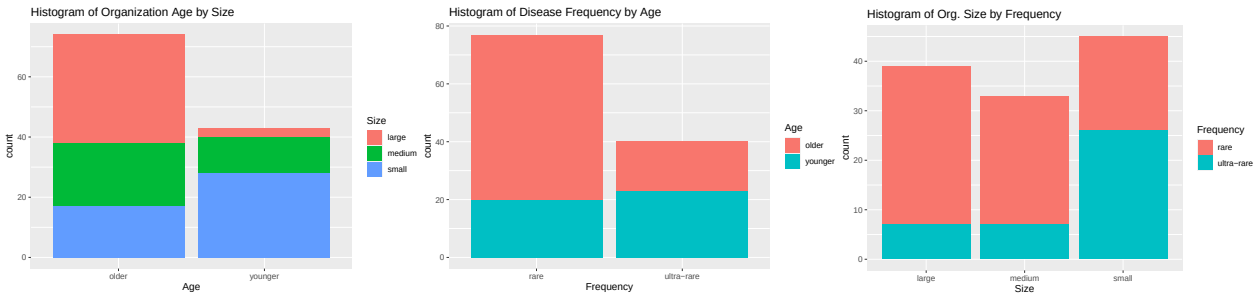


Figure 1: Histograms of each combination of independent variables.

variable (See Figure 4), we can see that many of the priorities are not equally represented by organizations in each of the frequency/age groups. The same holds true for the research and FDA therapy groups. This will be an important caveat to consider when interpreting results of the final analysis.

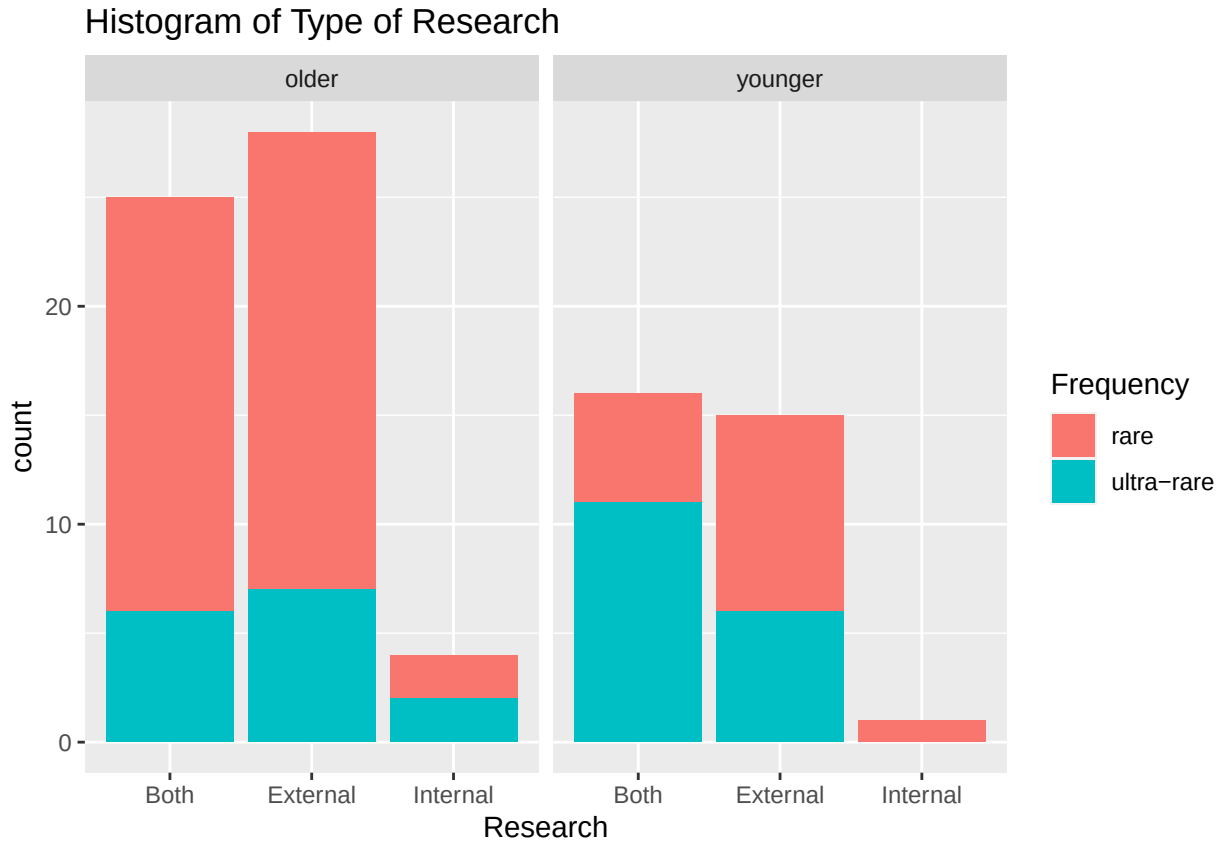


Figure 2: Histogram of internal research, external, or both, broken down by age and frequency.

As the last step in the EDA, we conducted a series of chi-squared tests to check for independence among the predictors of interest. The results of the tests are displayed in Figure ___. For all three pairs of predictors, the test shows p-values < 0.05, indicating that the variables are not independent at an alpha level of 0.05. Including correlated predictors in a regression model is problematic, so we must take this into account when fitting the models in the next step of the analysis.

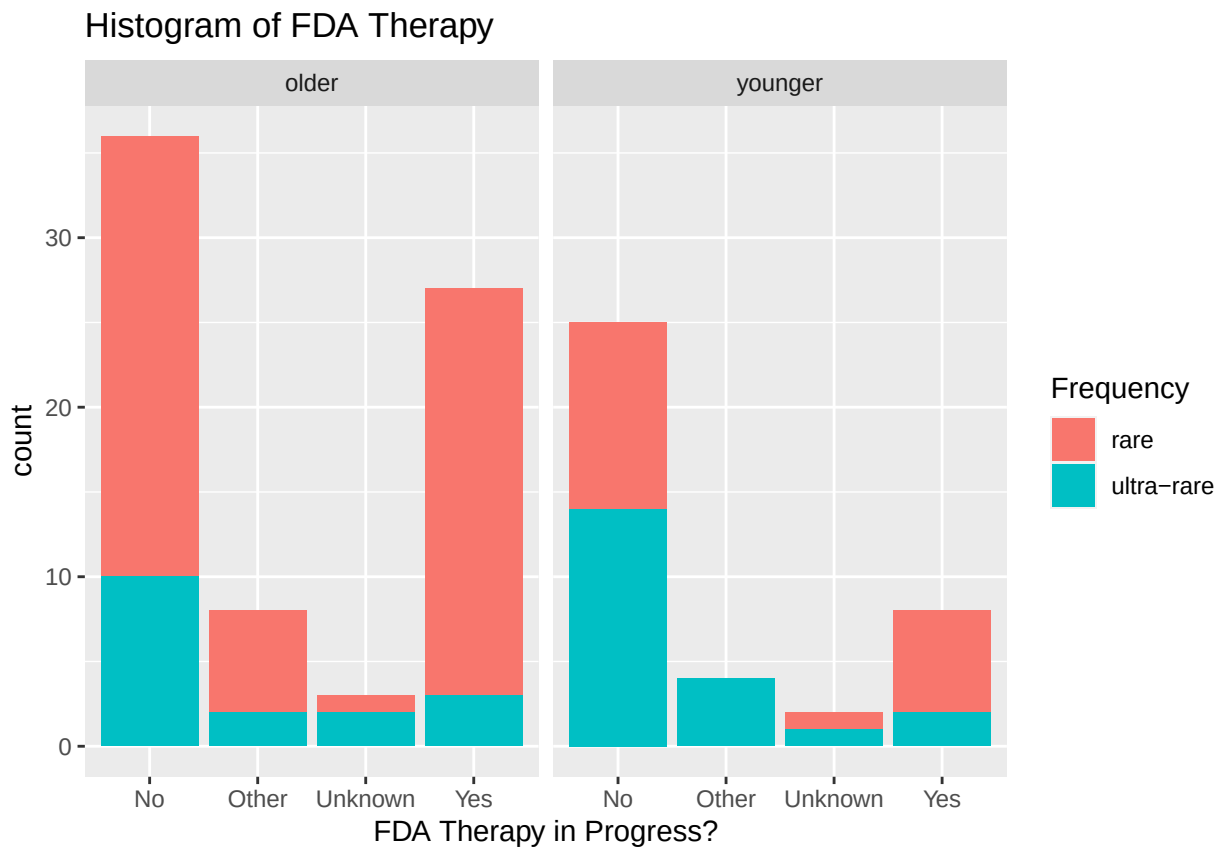


Figure 3: Histogram of whether FDA clinical trials are in progress or not, broken down by age and frequency.

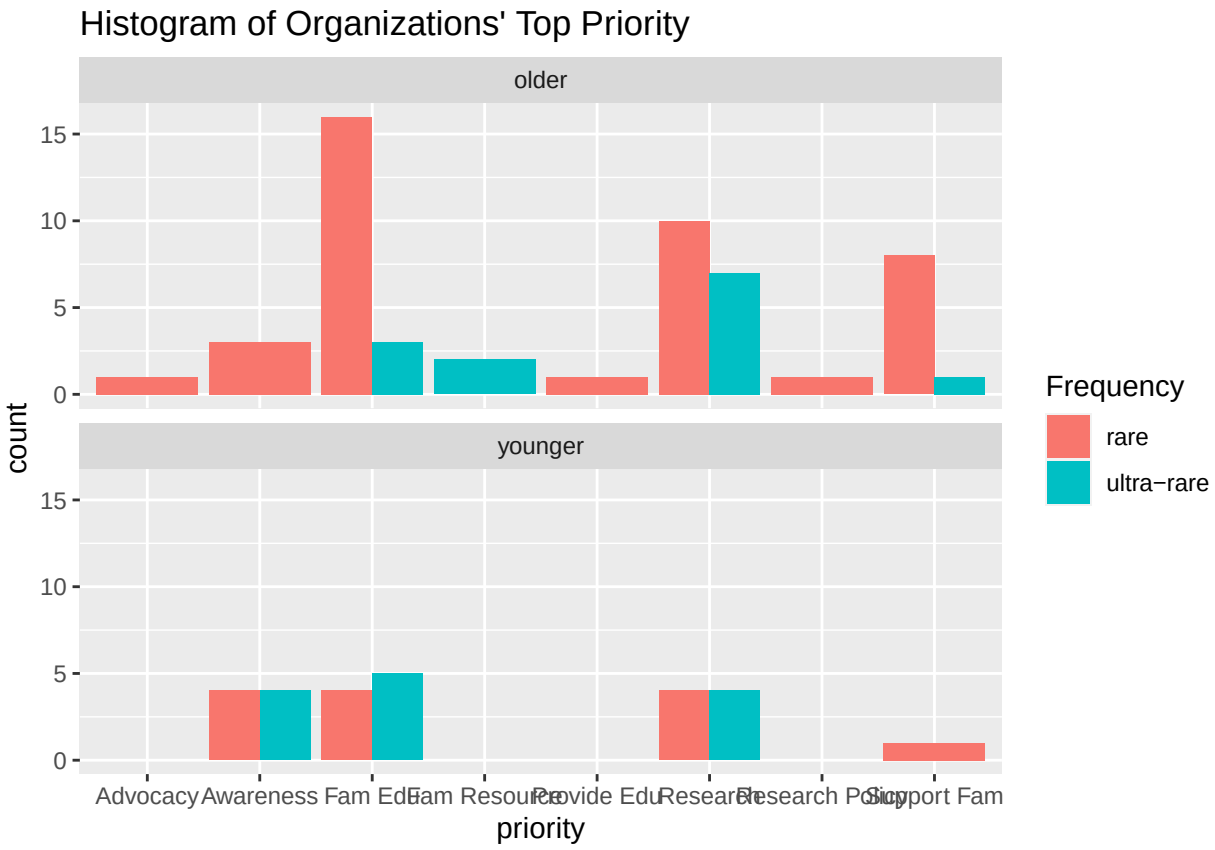


Figure 4: Histogram of organizations' top priorities, broken down by age and frequency.

Modeling

The client's variables of interest are categorical with more than two levels, so the most appropriate regression model to use for the analysis is a multinomial model. We fit a series of multinomial models, one for each response variable, using the "nnet" package in R.

To alleviate the problem of correlation between each pair of the predictors, we decided to remove some of them and to look for other useful predictors in the dataset. We found that Budget, a variable bucketed into four categories by the client, had no significant correlation with Frequency at the 0.05 alpha level, but significant correlation with Age and Size. Therefore, we added Budget as a predictor into the model along with Frequency, and removed Age and Size in order to eliminate the correlation between predictors.

Top Priorities

The first multinomial model we ran was using Priority as the response variable. As mentioned before, the predictors were Budget and Frequency. The result of the coefficient estimates shows that most of the combinations of predictor levels are significant at the 0.05 alpha level (see Figure ____ in the appendix for model output and residual plots). In addition, we did a point prediction to found out the most probable predicted categories under different conditions (see Figure 5). This shows that ultra-rare organizations are more likely to prioritize Research, while rare organizations are more likely to prioritize Family Education. Additionally, if the Budget is in the highest interval (> $500,000), the organization also has the highest probability of prioritizing Research.

```
## # weights:  48 (35 variable)
## initial  value 164.275882
## iter  10 value 106.693220
## iter  20 value 102.547632
## iter  30 value 101.769418
## iter  40 value 101.763571
## final  value 101.763565
## converged
```

Frequency	Budget	class
ultra-rare	Highest	Research
ultra-rare	Medium-High	Research
ultra-rare	Medium-Low	Fam Edu
ultra-rare	Lowest	Research
rare	Highest	Research
rare	Medium-High	Fam Edu
rare	Medium-Low	Fam Edu
rare	Lowest	Fam Edu

Research Internal vs. External

The second model we ran was using Research as the response variable, and Frequency and Budget as predictors. The model output and residual plots are shown in Figure ___. From the output, we can see that all of the predictors are non-significant at the 0.05 alpha level, suggesting that there is no relationship between Frequency and Budget and the whether the organizations' research is conducted internally, externally, or both.

FDA Therapy In Progress vs. Not In Progress

The last multinomial model we ran was using the FDA therapy variable as the response, and Frequency and Budget as predictors. The model output and residual plots are shown in Figure ___. From the output, we can see that...

Conclusion

There are several conclusions we can draw from the results explained above. First, different groups of organizations appear to have some significant differences in their top priorities. In particular, the ultra-rare disease and high budget organizations tend to be more likely to prioritize research, whereas the rare disease organizations tend to be more likely to prioritize family education. Second, we found no significant relationship between Frequency and Budget, and the response variable Research. Thus, it appears that there are no differences in the likelihood of having internal or external research for organizations for rare vs. ultra-rare diseases or orgaizations with different budgets. And finally, our model suggests that the ultra-rare groups tend to be less likely to have FDA clinical trials in progress.

There are several limitations to consider when interpreting these results. First, as mentioned previously, we found the predictor variables of interest (Age, Size, and Frequency) to be correlated with one another, so we were only able to use Frequency in the model with one other predictor, Budget, which was not correlated with Frequency but was correlated with Age and Size. Fitting a model with a choice of only two predictors is not ideal; had more predictors been available, including more than two in the model may have led to a better fit.

Second, the exploratory data analysis showed some imbalance in the dataset. For example, when looking at the histograms of the organizations' top priorities, we can see that not every priority was represented as a top priority for all of the predictor variable combinations. The issue of imbalance holds true for the other response variables as well. This makes it difficult for the models to accurately reflect patterns in the dataset.

Lastly, there is a concern that this analysis has fairly low power. The full dataset contains 217 responses, but based on the structure of some of the survey questions, some questions have even fewer responses. For example, the FDA therapy question only pertains to organizations who responded saying that there currently is no FDA approved therapy for that disease. Therefore, any organizations who do have a FDA therapy were not asked about whether clinical trials were in progress or not, making the sample size for this question much smaller than the original 217 respondents. Because of these three important limitations, the results must be taken and interpreted with caution.

References

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

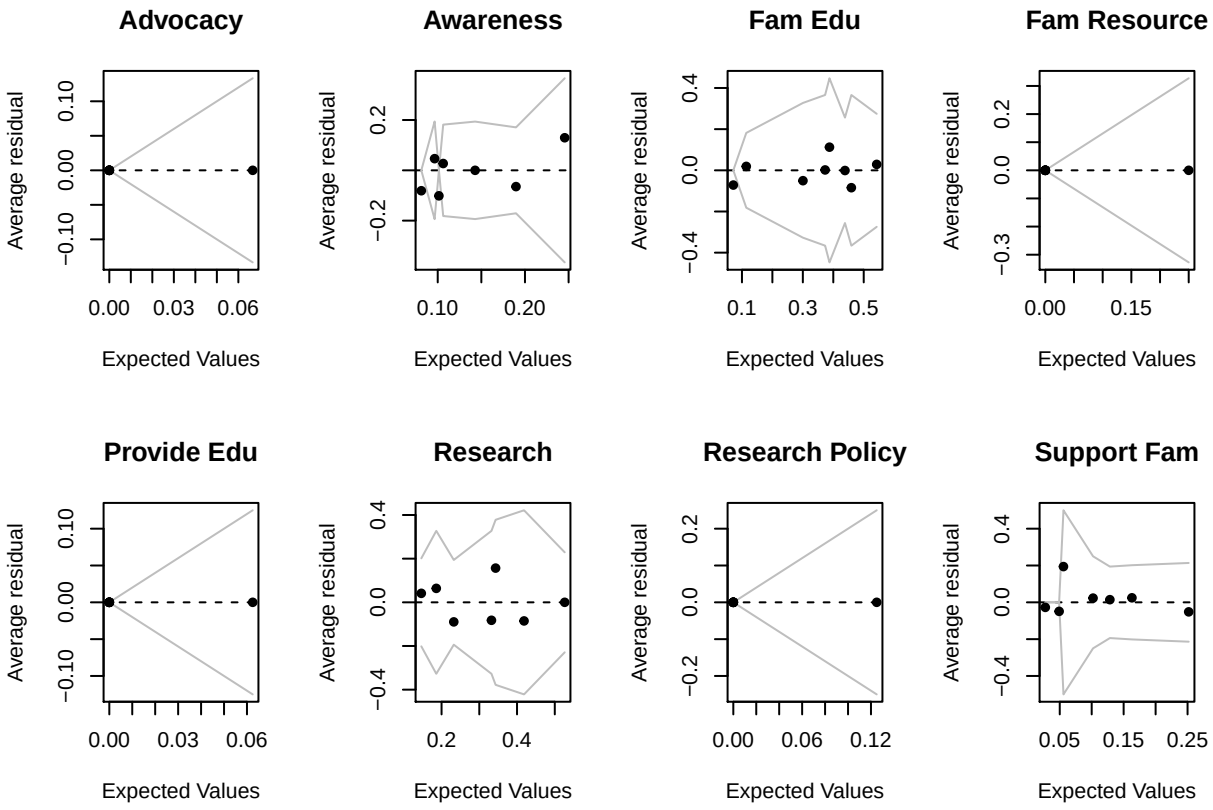
Stefan Milton Bache and Hadley Wickham (2020). magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. https://CRAN.R-project.org/package=magrittr

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Appendix

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Awareness	(Intercept)	0.47	1.24	3.800000e-01	0.71	-1.96	2.89
Awareness	Frequencyultra-rare	18.68	0.58	3.218000e+01	0.00	17.54	19.82
Awareness	BudgetMedium-High	16.12	0.83	1.943000e+01	0.00	14.50	17.75
Awareness	BudgetMedium-Low	17.25	0.73	2.364000e+01	0.00	15.82	18.68
Awareness	BudgetLowest	13.67	0.88	1.559000e+01	0.00	11.95	15.39
Fam Edu	(Intercept)	0.54	1.23	4.400000e-01	0.66	-1.87	2.95
Fam Edu	Frequencyultra-rare	18.26	0.47	3.905000e+01	0.00	17.35	19.18
Fam Edu	BudgetMedium-High	17.78	0.69	2.577000e+01	0.00	16.43	19.13

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Fam Edu	BudgetMedium-Low	18.02	0.69	2.627000e+01	0.00	16.68	19.36
Fam Edu	BudgetLowest	14.88	0.75	1.975000e+01	0.00	13.40	16.35
Fam Resource	(Intercept)	-61.00	0.39	-1.564000e+02	0.00	-61.76	-60.23
Fam Resource	Frequencyultra-rare	57.59	0.39	1.476600e+02	0.00	56.83	58.35
Fam Resource	BudgetMedium-High	39.81	0.39	1.020600e+02	0.00	39.04	40.57
Fam Resource	BudgetMedium-Low	-11.63	NaN	NaN	NaN	NaN	NaN
Fam Resource	BudgetLowest	-19.26	NaN	NaN	NaN	NaN	NaN
Provide Edu	(Intercept)	-28.35	0.66	-4.325000e+01	0.00	-29.63	-27.06
Provide Edu	Frequencyultra-rare	-29.62	NaN	NaN	NaN	NaN	NaN
Provide Edu	BudgetMedium-High	-2.23	0.00	-1.117018e+16	0.00	-2.23	-2.23
Provide Edu	BudgetMedium-Low	44.96	0.66	6.859000e+01	0.00	43.67	46.24
Provide Edu	BudgetLowest	-3.86	0.00	-1.109585e+20	0.00	-3.86	-3.86
Research	(Intercept)	1.93	1.07	1.810000e+00	0.07	-0.15	4.02
Research	Frequencyultra-rare	19.24	0.47	4.113000e+01	0.00	18.33	20.16
Research	BudgetMedium-High	15.54	0.56	2.789000e+01	0.00	14.44	16.63
Research	BudgetMedium-Low	15.53	0.58	2.684000e+01	0.00	14.39	16.66
Research	BudgetLowest	12.58	0.66	1.907000e+01	0.00	11.29	13.87
Research Policy	(Intercept)	-31.86	0.67	-4.748000e+01	0.00	-33.18	-30.54
Research Policy	Frequencyultra-rare	-37.48	NaN	NaN	NaN	NaN	NaN
Research Policy	BudgetMedium-High	-3.25	0.00	-2.574346e+22	0.00	-3.25	-3.25
Research Policy	BudgetMedium-Low	-1.40	0.00	-2.645756e+21	0.00	-1.40	-1.40
Research Policy	BudgetLowest	45.97	0.67	6.852000e+01	0.00	44.66	47.29
Support Fam	(Intercept)	1.33	1.12	1.190000e+00	0.24	-0.87	3.52
Support Fam	Frequencyultra-rare	17.22	0.85	2.031000e+01	0.00	15.56	18.89
Support Fam	BudgetMedium-High	15.55	0.76	2.059000e+01	0.00	14.07	17.03
Support Fam	BudgetMedium-Low	16.24	0.69	2.364000e+01	0.00	14.89	17.59
Support Fam	BudgetLowest	12.58	0.95	1.319000e+01	0.00	10.71	14.45


Advocacy
Awareness
Fam Edu
Fam Resource
Provide Edu
Research
Research Policy
Support Fam
Average residual
Expected Values