Written Report

Outliers: Matthew Hawkins, Aston Yong, Catherine Ning

November 11, 2021

1. Introduction and Data

For this project, we are looking into the Himalayan database, which is a compilation of all expeditions that have climbed in the Nepal Himalaya. The data cover all expeditions from 1905 through Spring 2019 to more than 465 significant peaks in Nepal. Also included are expeditions to both sides of border peaks such as Everest, Cho Oyu, Makalu and Kangchenjunga as well as to some smaller border peaks. Data on expeditions to trekking peaks are included for early attempts, first ascents and major accidents. The main research question we are hoping to solve involves looking into various different factors which can predict the safety and success of the expedition.

Many active climbers view the Himalayas as the proverbial "summit" of their climbing experience. These mountains provide a demanding and often dangerous task so, the most important goal for climbers is safety and success. We want to create a model that find the most significant factors for safe and successful expeditions so that future climbers can predict probabilities of safe, successful climbs. Climbing the Himalayas is still a dangerous and risky activity, and increased congestion is changing the dynamic of climbing (http://graphics.reuters.com/NEPAL-EVEREST/0100B4S22JR/index.html). Therefore, examining these dataset predictors and drawing insights from them will hopefully improve mountain climbing preparedness.

We wish to explore the following research question:

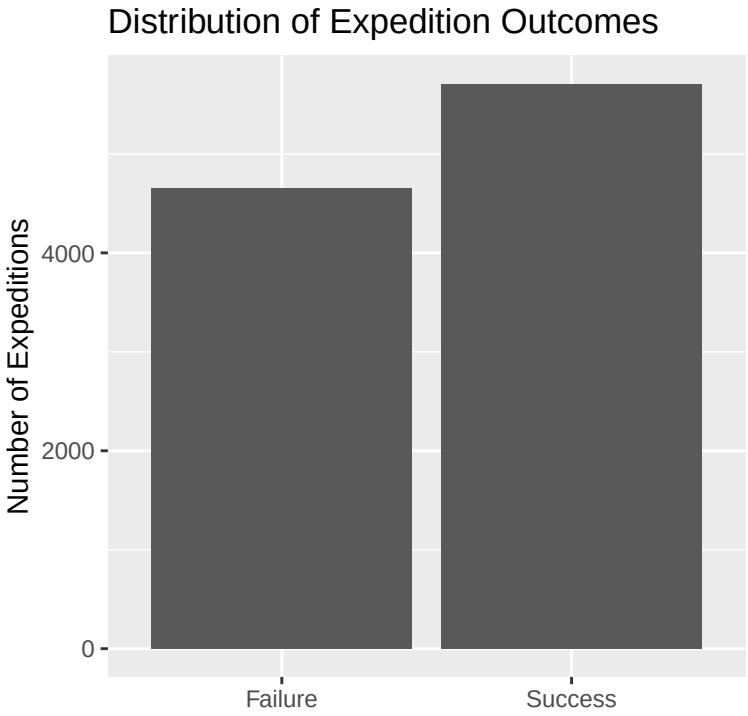
What expedition factors are most predictive of safe and successful expeditions and what is the probability of success given these predictors?

We hypothesize expeditions with conditions like lower peaks, historically climbed peaks, more members, warm seasons, younger average age, and more recent year to have greater probability of safe success.

The data cover all expeditions from 1905 through Spring 2019 to more than 465 significant peaks in Nepal. It is divided into 3 tables, one for peaks, expeditions, and members respectively. The peaks.csv file contains information about each Himalayan mountain with data on peak name, whether it has been climbed, height, and other variables. The expeditions.csv file contains data on individual expeditions, with variables like member count, season, and number of deaths. The members.csv file is the most narrow-scoped table with demographic and personal information for each member on an expedition. Multiple members can participate in one expedition, and multiple expeditions can climb one peak.

The database is based on the expedition archives of Elizabeth Hawley, a longtime journalist based in Kathmandu, and it is supplemented by information gathered from books, alpine journals and correspondence with Himalayan climbers. The database is updated bi-annuallly, and member information comes from submitted permit applications.

We created our response variable, sucess_expedition, defining an expedition as a success or a failure based on the reason for the expedition's termination (Success (main peak), Success (subpeak), accident, bad weather, etc). A successful expedition has the value "1" and an unsuccessful expedition has the value "0".



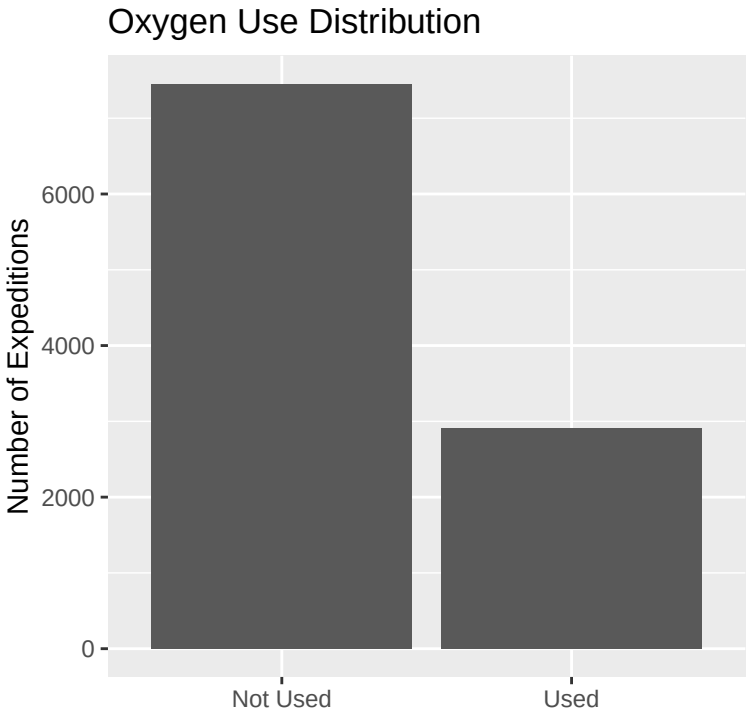
success_expedition	n
0	4657
1	5707

Based on the above count and calculation, 55.1% of the 10364 expeditions were classified as successful. On the other hand, 44.9 of the 10364 expeditions were classified as unsuccessful.

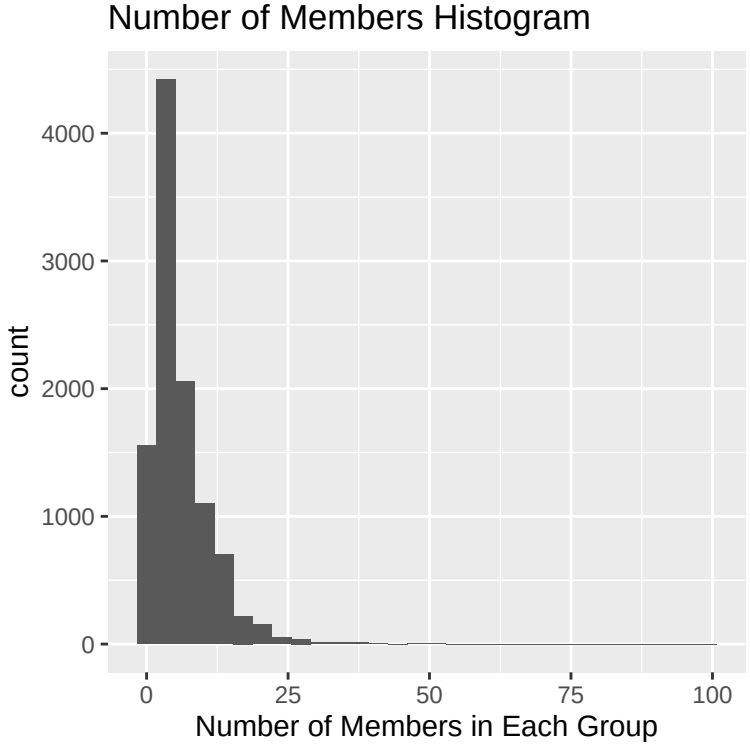
First of all, we join grouped members info and peak height into expeditions.

We selected all the possible predictor variables we are planning to use, including peak height, member count, number hired members, oxygen use, year, season, agency, member deaths, and hired deaths. We removed all NAs from trekking_agency. From the members.csv file, average member age on an expedition calculated from grouping by expedition id.

There are over 800 trekking agencies, so to simplify the data, we only kept the top 10 agencies and set trekking_agency to Other for any other agency. To make the model intercept interpretation meaningful, we also mean centered the continuous variables for mean age, mountain height, and member count.



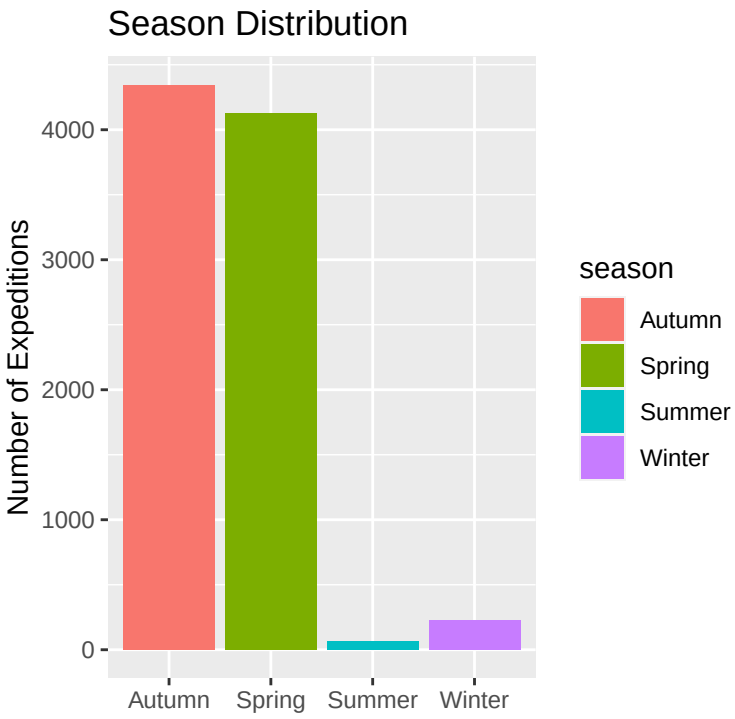
In the above bar graph, we can see the portion of expeditions in which oxygen was used, which we believe to be a potentially important predictor variable. As the graph shows, a little less than a third of expeditions used oxygen.



mean	sd	IQR
5.953	5.428	6

The above histogram shows the distribution of the number of members in each group, another variable which we believe may be a significant predictor of the odds of success of an expedition. The distribution is right-skewed and seems like it may have a few outlier expeditions that have greater than 25 members in the group. The distribution is centered with a mean of 5.953 members, and it has a spread characterized by a standard deviation of 5.428 and an interquartile range of 6.

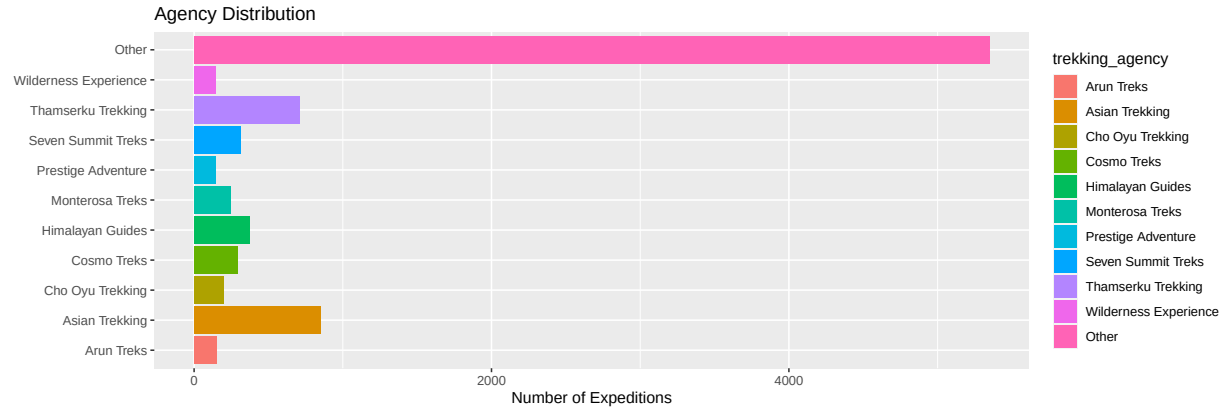
We also mapped out the distribution of season. We can notice that for spring and autumn, the numbers of expeditions are over 4,000 much higher than those of summer and winter, below 200.



trekking_agency	n
Arun Treks	148
Asian Trekking	853
Cho Oyu Trekking	199
Cosmo Treks	293
Himalayan Guides	374
Monterosa Treks	243
Prestige Adventure	147
Seven Summit Treks	313
Thamserku Trekking	707
Wilderness Experience	145
Other	5348

According to the graph, the “other” agency has the most trekking agencies, about 5500, while among the following trekking agencies, Thamserku and Asian trekking have relatively higher number of trekking agencies

compared to the rest of the agencies.



2. Methodology

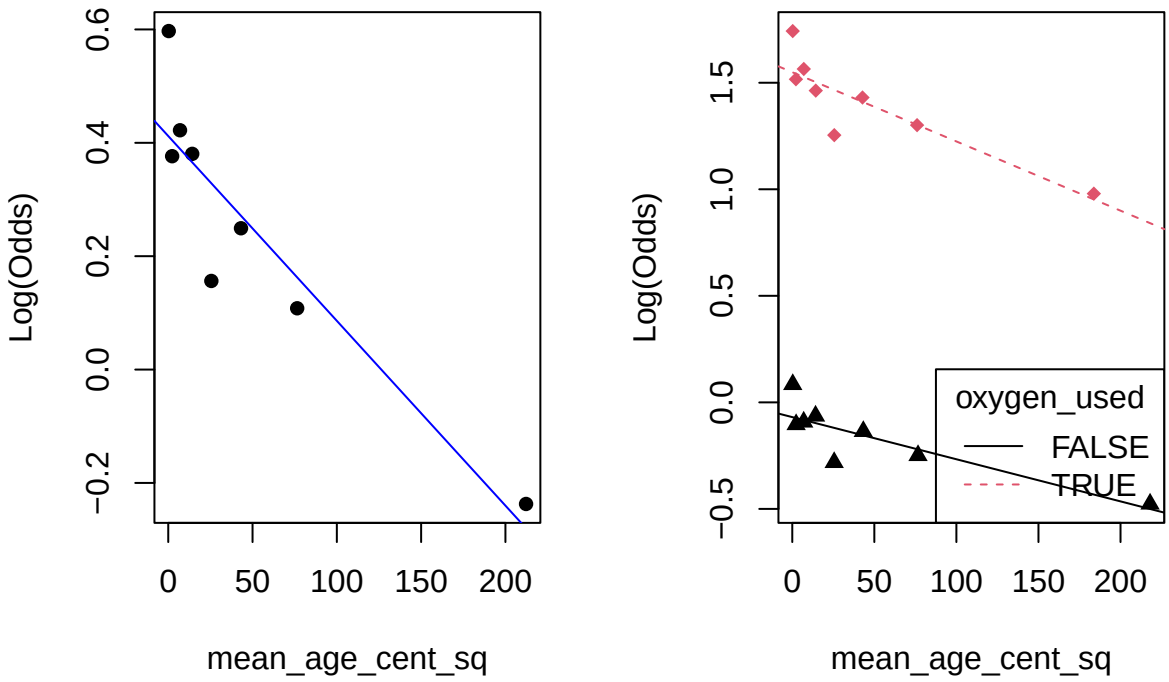
Because we want to predict success, a categorical response, we use logistic regression to model the log odds of expedition success versus failure. Probability of success must be between 0 and 1. We used forward selection and BIC to build our model and determine signficant predictor variables of odds of success. Forward selection model using BIC produces the following model:

term	estimate	std.error	statistic	p.value
(Intercept)	-21.35504	5.76027	-3.70730	0.00021
oxygen_usedTRUE	1.94305	0.06986	27.81445	0.00000
height_metres	-0.00052	0.00003	-15.28838	0.00000
member_count	0.06817	0.00506	13.48474	0.00000
mean_age	-0.03359	0.00361	-9.30802	0.00000
hired_staff_deaths	-0.92879	0.17038	-5.45115	0.00000
year	0.01298	0.00288	4.50529	0.00001

In the predictor variables' empirical logit plots (Appendix 5.1), we observe interaction effects between oxygen use and the quantitative predictors peak height and year. We therefore add interaction terms into our model. There is also a clear quadratic relationship between log odds of success and mean_age_cent, so we transformed mean_age_cent into a quadratic term. This is expected as too young or too old climbers will be less prepared or fit. Year also showed possible quadratic relationships, but squaring year did not change the empirical logit plot. We decided to keep year unchanged. The predictor year in our model with interaction terms and quadratic transformed mean age had a p-value greater than 0.05, so we performed a drop in deviance test. The test p-value is near zero, so we keep year in the model.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.12853	0.11233	-1.14423	0.25253	-0.34879	0.09158
oxygen_usedTRUE	-0.77096	0.25250	-3.05335	0.00226	-1.26556	-0.27532
height_metres_cent	-0.00063	0.00004	-17.59754	0.00000	-0.00071	-0.00056
member_count_cent	0.06468	0.00521	12.41366	0.00000	0.05456	0.07499
mean_age_cent	-0.02315	0.00388	-5.96951	0.00000	-0.03077	-0.01556
mean_age_cent_sq	-0.00219	0.00036	-6.10579	0.00000	-0.00290	-0.00150
hired_staff_deaths	-0.81198	0.16755	-4.84619	0.00000	-1.15379	-0.49903
year	-0.00155	0.00321	-0.48284	0.62921	-0.00785	0.00474
oxygen_usedTRUE:height_metres_cent	0.00098	0.00012	8.18924	0.00000	0.00075	0.00122
oxygen_usedTRUE:year	0.06035	0.00674	8.94971	0.00000	0.04715	0.07359

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
8762	10333.00	NA	NA	NA
8760	10238.65	2	94.352	0



Checking model conditions:

1. Linearity: The empirical logit plots (above and Appendix 5.1) show a general linear relationship between the log odds of success and year, peak height, and hired staff deaths. There is a strong linear relationship between log odds of success and member count and mean age squared.
2. Randomness: randomness is satisfied because the data is just a record of all Himalayan expeditions, so the data collection is representative of the population and is not a particular subset.
3. Independence: independence condition is satisfied because expeditions are separate and odds of success for one expedition do not tell anything about odds of success for another expedition.

Multicolinearity is not an issue, as the Variance Inflation Factor (VIF) is greater than 10 only for oxygen_used and its interaction with year. This is expected as an interaction term needs the associated main effects.

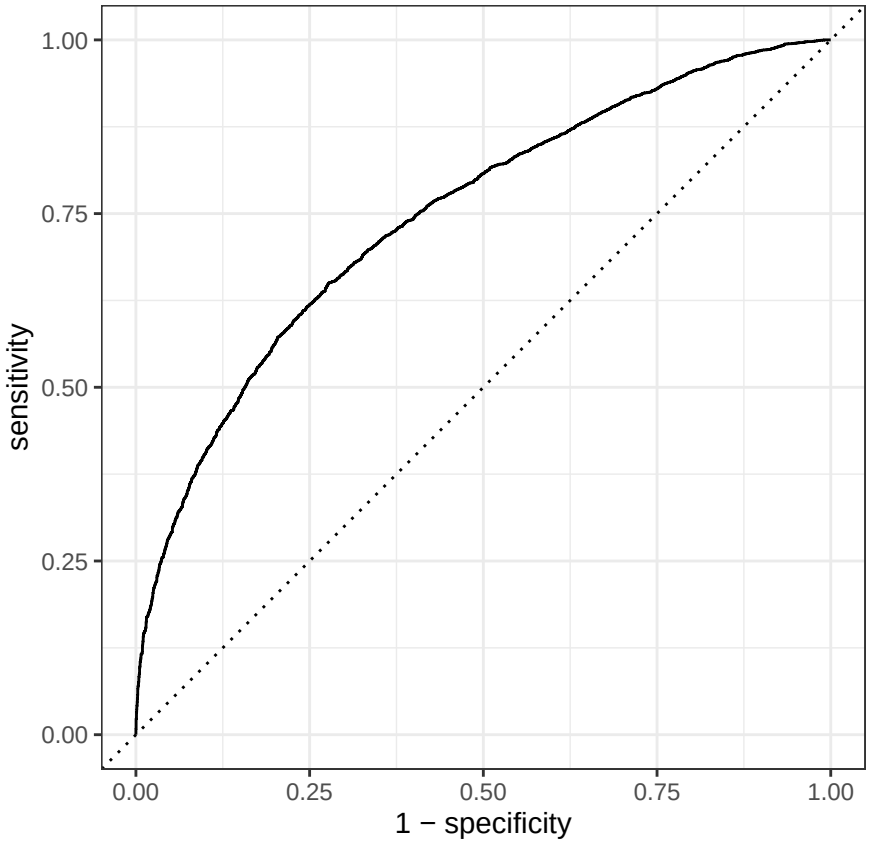
	x
oxygen_usedTRUE	19.10393
height_metres_cent	1.54860
member_count_cent	1.10563
mean_age_cent	1.19875
mean_age_cent_sq	1.12342
hired_staff_deaths	1.01715
year	1.42128

	x
oxygen_usedTRUE:height_metres_cent	3.07386
oxygen_usedTRUE:year	17.52701

3. Results

Our final model is displayed here:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.12853	0.11233	-1.14423	0.25253	-0.34879	0.09158
oxygen_usedTRUE	-0.77096	0.25250	-3.05335	0.00226	-1.26556	-0.27532
height_metres_cent	-0.00063	0.00004	-17.59754	0.00000	-0.00071	-0.00056
member_count_cent	0.06468	0.00521	12.41366	0.00000	0.05456	0.07499
mean_age_cent	-0.02315	0.00388	-5.96951	0.00000	-0.03077	-0.01556
mean_age_cent_sq	-0.00219	0.00036	-6.10579	0.00000	-0.00290	-0.00150
hired_staff_deaths	-0.81198	0.16755	-4.84619	0.00000	-1.15379	-0.49903
year	-0.00155	0.00321	-0.48284	0.62921	-0.00785	0.00474
oxygen_usedTRUE:height_metres_cent	0.00098	0.00012	8.18924	0.00000	0.00075	0.00122
oxygen_usedTRUE:year	0.06035	0.00674	8.94971	0.00000	0.04715	0.07359




```
## [1] 0.7474922
```

Interpretations:

The key predictors of expedition success are peak height, member count, mean member age, staff deaths,

year, and oxygen use.

Intercept: For an expedition in 1971 with mean peak height of 7855.83 meters, 6.989 group members, mean age of 38.506, 0 staff deaths, and no oxygen use, we would expect the odds of a successful expedition to be 0.879.

Height_metres_cent: For each additional meter in the height of the peak, we would expect, on average, that the odds of a successful expedition are multiplied by 0.999, holding all else constant.

Member_count_cent: For each additional member in the group, we would expect, on average, that the odds of a successful expedition are multiplied by 1.067, holding all else constant.

Mean_age_cent: Because of the quadratic term, mean_age_cent must be interpreted using the overall effect. The effect of mean_age_cent on log odds equals zero at mean centered mean_age of -5.285, which equals an expedition mean age of 33.221. The negative coefficient of the quadratic term means change in log odds is positive for mean_age_cent below -5.285, meaning mean_age between 0 and 33.221. In this interval, odds of expedition success is expected to multiply by a factor greater than 1 for a one year mean_age_cent increase. Change in log odds is negative for mean_age_cent above -5.285, or mean age above -5.285, so odds of expedition success is expected to multiply by a factor less than 1 for a one year mean_age_cent increase.

Hired_staff_deaths: For each additional death of hired staff, we would expect, on average, that the odds of a successful expedition are multiplied by 0.444, holding all else constant.

Year: For each additional year since 1971, we would expect, on average, that the odds of a successful expedition are multiplied by 0.998, holding all else constant.

Oxygen_usedTRUE: If oxygen were used, we would expect, on average, that the odds of a successful expedition are multiplied by 0.463, holding all else constant.

height_metres:oxygen_usedTRUE: If oxygen were used, we would expect, on average, that for every one meter increase in the height of the peak, the odds of a successful expedition are multiplied by 1.001, holding all else constant.

year:oxygen_usedTRUE: If oxygen were used, we would expect, on average, that for every additional year since 1971, the odds of a successful expedition are multiplied by 1.062, holding all else constant.

Findings

The AUC value of 0.747 indicates that our model has a good fit to the data and can predict expedition success outcomes with high sensitivity and low false positive rates. The key predictors of expedition success found are all reasonable. Climbers too young or too old would have a lower probability of success, explaining the quadratic relationship. Oxygen use may indicate higher elevation or health complications, negatively impactimpacting on log odds of success. The surprising finding that log odds of success declines with year may reflect the growing access to mountain climbing. More amateur climbers in the Himalayas translates into decreasing log odds of success.

4. Discussion and Conclusion

Through this research we learned that the most significant predictors of expedition success were peak height, member count, mean member age, staff deaths, year, and oxygen use. Higher peaks, staff, deaths, particularly young or old age, and use of oxygen reduced log odds of success. Expeditions further in the past and with more members had higher log odds of success. Our model with these predictors minimized BIC and created the simplest model for prediction. All the predictors identified were statistically significant, and the year variable, which initially had a high p-value was analyzed with a drop in deviance test, confirming the variable's significance in the model. As expected, interaction terms between oxygen_use and height indicate reduced oxygen at higher elevations. The interaction term with year may also suggest changing oxygen technology over time.

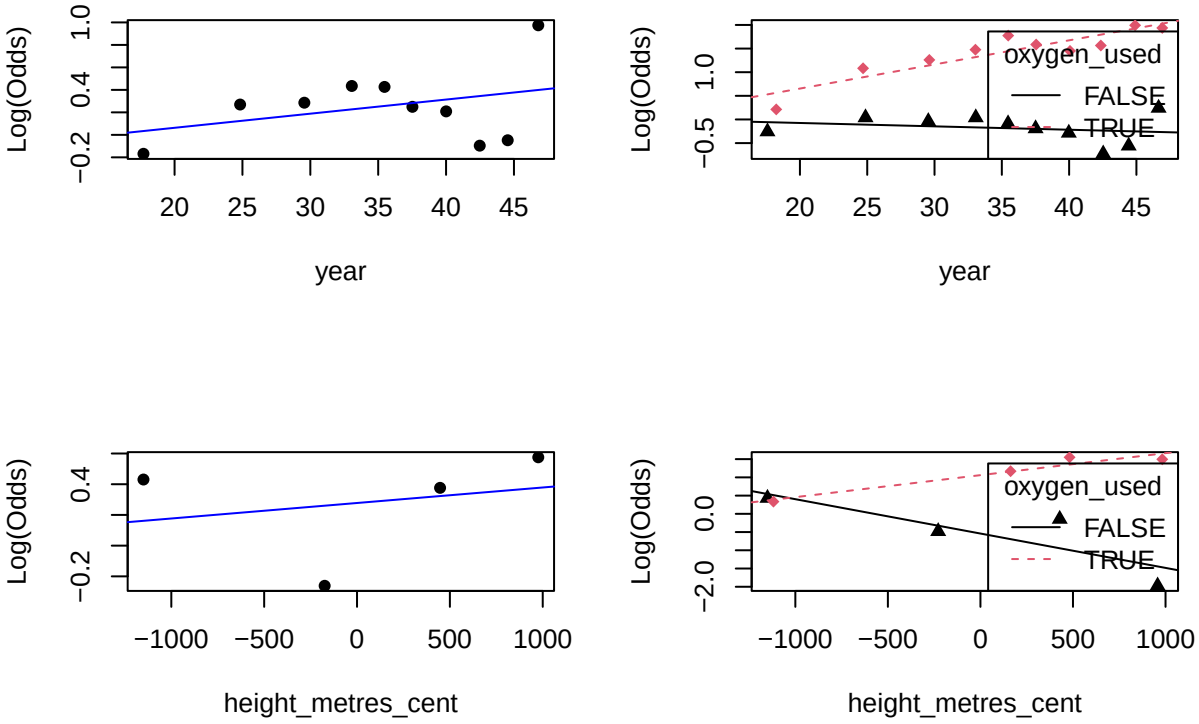
Limitations:

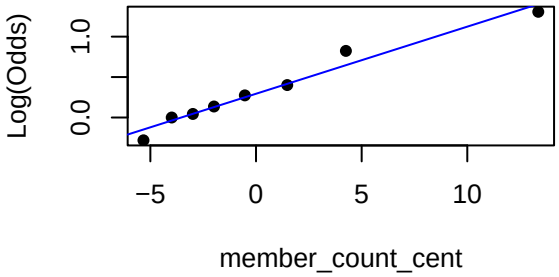
We did not examine all interaction terms between our predictor variables. These interaction terms might be impacting the accuracy of the model. Independence may also be influenced by particularly popular peaks. We remove all data entries with one or more NAs. These entries might influence the final model prediction. We did not explore other kinds of models other than logistic (after discovering linear models are probably not the best choice), but we fail to consider other possible models such as exponential and quadratic. We were unable to examine the distributions of all predictor variables to see their effects on the model I.e. There were likely more data from recent years

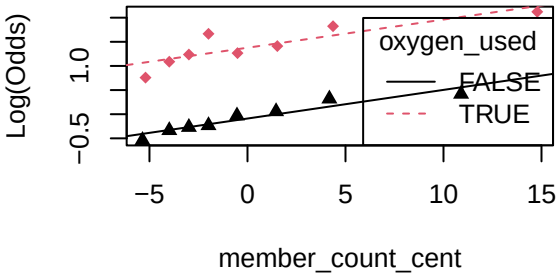
In the future, if we have more time to work on this project extensively, we might look at multiple interactions between the predictor variables we have selected. We can also look for alternative treatments of NA values, as well as comparing our model with other models. Determining if any independence issues are caused by popular peaks would be informative, as well.

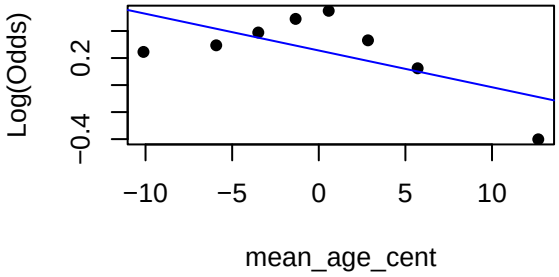
5. Appendix

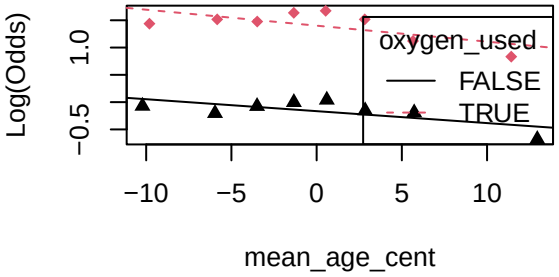
5.1 Empirical Logit Plots




Log(Odds)
0.0
1.0
−5
0
5
10
member_count_cent


Log(Odds)
−0.5
1.0
oxygen_used
FALSE
TRUE
−5
0
5
10
15
member_count_cent


Log(Odds)
−0.4
0.2
−10
−5
0
5
10
mean_age_cent


Log(Odds)
−0.5
1.0
oxygen_used
FALSE
TRUE
−10
−5
0
5
10
mean_age_cent


Log(Odds)
−0.5
−2.5
0
2
4
6
8
10
hired_staff_deaths