Himalayan Expeditions Success Analysis*

Analysis using data from Himalayan expeditions from 1905 through Spring 2019 leveraging Bayesian Logistic Regression

Kaavya Kalani

April 13, 2024

Abstract

Table of contents

*Code and data supporting this analysis is available at: https://github.com/kaavyakalani26/himalayan-expeditions-analysis

1 Introduction

Mountaineering, with its blend of adventure and challenge, has captivated explorers for generations. It represents a pinnacle of human endeavor, testing physical prowess, mental fortitude, and strategic planning against some of the most formidable natural landscapes on Earth. Amidst the allure of conquering these majestic peaks, lies a critical question: what factors contribute to the likelihood of a successful summit attempt? This paper delves into precisely this inquiry.

There are numerous mountain ranges around the world where people embark on expeditions. One of the most famous mountain ranges are the Himalayas, which also consist of the highest mountain peak in the world, Mt. Everest. Our paper uses the data from expeditions to the Himalayan mountain ranges in Nepal.

Our estimand is the relationship between different demographic, environmental and geographic factors (such as height of the peak, sex, age, season of expedition and if it was a solo ascent) and the successfulness of climbing a summit. Using our extensive analysis dataset, our goal is to identify trends and factors that influence a successful expedition and eventually conclude what factors help in a more successful attempt.

We use data from Alex Cookson's datasets (link) which are sourced from The Himalayan Expedition records (link), to understand these factors and trends. This is done by leveraging a Bayesian Logistic Regression model and then predicting the probability of a successful attempt over various demographic and environmental factors.

Our analysis led to us … (add what was found)

This research highlights the interplay between climber attributes and environmental conditions, offering insights into Himalayan expeditions. These findings have practical implications for expedition planning, risk management, and safety protocols. Ultimately, our study aims to improve decision-making in high-altitude mountaineering, making it safer and more informed in one of the world's most challenging environments.

The paper is further organized into four sections. Section 2 discusses how the dataset to be used for the analysis was obtained and pre-processed. I will explain the variables of interest in the dataset for the analysis. Section 3 describes the model being used for the analysis. Section 4 then highlights and discusses the trends and associations found during the analysis. Lastly, Section 5 talks about some interesting trends found in Section 4 in depth, link it to the real world and also highlight the weaknesses and future of my analysis.

2 Data

For this analysis, we have used combined three datasets into one, which is used for analysis. The datasets were cleaned and analysed using the statistical programming software `R` (R Core

Team 2023) along with the help `tidyverse` (Wickham et al. 2019), `knitr` (Xie 2014), `ggplot2` (Wickham 2016), `here` (Müller 2020), `dplyr` (Wickham et al. 2023), `rstanarm` (Goodrich et al. 2024), `broom.mixed` (Bolker and Robinson 2022), `modelsummary` (Arel-Bundock 2022) and `kableExtra` (Zhu 2024).

2.1 Analysis Dataset

The raw datasets were obtained from Alex Cookson's datasets (link). I chose the ones cleaned for Himalayan expeditions. Alex got his datasets from The Himalayan Database (link).

The Himalayan Database is a compilation of records for all expeditions that have climbed in the Nepal Himalaya. 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 original database currently covers all expeditions from 1905 through Spring-Summer 2023 to the most significant mountaineering 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 updates to this database are published bi-annually.

My dataset derived from Alex's contains the entries from 1905 through Spring 2019.

The three datasets I considered included information about all peaks, all expeditions on those peaks and all members on those expeditions. The data from these three datasets are combined to form our analysis dataset.

A person becomes an entry in my analysis dataset if, between 1905 and Spring 2019, they attempted to climb any one of the many Himalayan peaks in Nepal. It also included expeditions to both sides of border peaks as mentioned before.

Among the overall range of variables available, we chose the following to be included in our analysis dataset.

- `Height of the peak` in metres for the peak the person in the current entry is on an expedition for
- `Seasons` is the season the expedition is in. This takes on either of the 4 values: Autumn, Spring, Winter, Summer.
- `Sex` is the sex reported by the expedition member and it is either male or female.
- `Age` is the age of the expedition member at the time of the expedition. Depending on the best available data, this could be as of the summit date, the date of death, or the date of arrival at basecamp.
- `Success`represents whether the person was successful in summitting the goal peak.
- `Solo` represents whether the person attempted a solo ascent.
- `Died` represents whether the person died during the expedition.

(Plot the variables)

Out of these, we will be using `height of the peak`, `seasons`, `sex`, `age` and `solo` in our model as the independent variables and `success` as the dependent variable.

3 Model

I used a Bayesian Logistic Regression model to find the probability that someone will successfully summit the Himalayan peak they are on the expedition for. Logistic regression is a method used for binary classification to predict the probability of a categorical dependent variable.

For my analysis, a logistic regression model will be used to model if the person will be successful on their attempt to summit the particular Himalayan peak. The model will be based on five independent demographic variables: `height of the peak`, `sex`, `age`, `seasons` and `solo`.

The logistic regression model I will be using is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 \times \text{height} + \beta_2 \times \text{sex} + \beta_3 \times \text{age} + \beta_4 \times \text{seasons} + \beta_5 \times \text{solo} \tag{1}$$

$$\begin{aligned}
\beta_0 &\sim \text{Normal}(0, 2.5)\\
\beta_1 &\sim \text{Normal}(0, 2.5)\\
\beta_2 &\sim \text{Normal}(0, 2.5)\\
\beta_3 &\sim \text{Normal}(0, 2.5)\\
\beta_4 &\sim \text{Normal}(0, 2.5)\\
\beta_5 &\sim \text{Normal}(0, 2.5)
\end{aligned}$$

where,

- $\hat{p}$ represents the probability that someone will successfully complete the peak they are on the expedition for.
- β_0 represents the intercept term of this logistical regression. It is the probability that someone will successfully complete the peak they are on the expedition for if the predictors' values are zero
- β_1 is the coefficient corresponding to height of the peak
- β_2 is the coefficient corresponding to sex of the person
- β_3 is the coefficients corresponding to age of the person
- β_4 is the coefficients corresponding to seasons of the person
- β_5 is the coefficients corresponding to if the person is attempting the summit alone

In my model, normal priors with a mean of 0 and a standard deviation of 2.5 are used for both the coefficients and the intercept. Setting the mean of the priors to 0 implies that there is no expectation of a particular direction or magnitude for the coefficients or intercept. I chose this as I have no expectation of the same. The standard deviation of 2.5 reflects the uncertainty or variability in the prior beliefs. I chose a moderately wide prior to allow for a reasonable amount of uncertainty.

The chosen priors allow the data to largely determine the posterior distribution as they are relatively non-informative. They don't heavily influence the results unless the data provide strong evidence to the contrary.

The use of moderately wide priors can also help regularize the model, preventing overfitting and providing more stable estimates, particularly when dealing with limited data.

Table 1: Summary of the model

term	estimate	std.error	conf.low	conf.high
(Intercept)	-0.80	0.03	-0.84	-0.76
sexM	0.25	0.03	0.21	0.30
seasonsSpring	0.23	0.02	0.21	0.26
seasonsSummer	-0.44	0.09	-0.59	-0.29
seasonsWinter	-0.65	0.06	-0.74	-0.56

Table 1 shows the coefficients for my Bayesian model along with the standard error and the 95% credible interval. The standard error (SE) is a measure of the precision with which a sample statistic estimates a population parameter. It quantifies the variability of sample statistics around the population parameter. A 95% credible interval means that there is a 95% probability that the true parameter lies within the interval, given the observed data and the model assumptions.

Figure 1 illustrates the coefficients and their associated 95% credible intervals for the predictor variables in the Bayesian model. Each point represents the estimated coefficient for a predictor variable, while the horizontal lines depict the credible interval around the estimate. Variables with coefficients to the right of zero indicate a positive association with the outcome variable, suggesting that an increase in the predictor variable corresponds to an increase in the outcome variable. Conversely, coefficients to the left of zero indicate a negative association, implying that an increase in the predictor variable is associated with a decrease in the outcome variable. These insights can help in understanding the direction and magnitude of the relationships between predictor variables and the outcome in the Bayesian model.

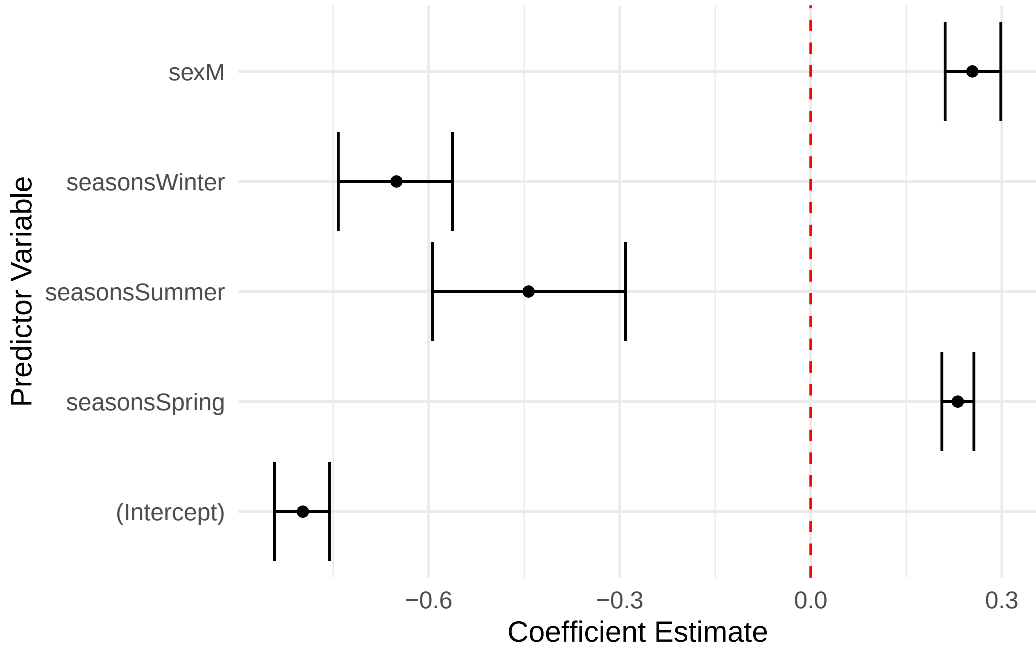

Figure 1: Coefficients of the model

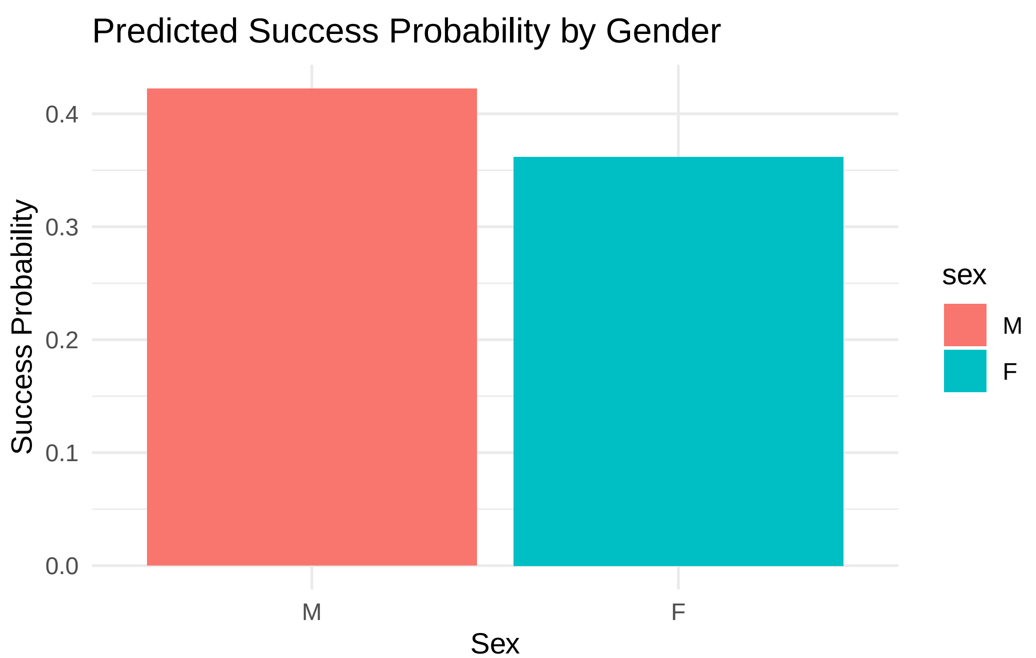

Figure 2: Results1

4 Results

5 Discussion

1) What is done in this paper?
2) What is something that we learn about the world?

5.1 Weaknesses and Limitations

5.2 Future Directions

6 Appendix

6.1 Cleaning

For the analysis data, the cleaning steps we took were:

1. Initial merging: I first merge the raw data from the `expeditions` and `members` datasets based on a common identifier, `expedition_id`, to consolidate information about expedition participants.
2. Column selection and renaming: Irrelevant columns are removed from the merged dataset, and the remaining columns (`peak_id.x`, `season.x`, `sex`, `age`, `success`, `solo`, `died`) are selected for further analysis. Additionally, column `peak_id.x` is renamed to `peak_id`.
3. Secondary merging: The cleaned `expeditions` dataset is merged with the `peaks` dataset based on a common identifier, `peak_id`, to incorporate information about the height of each peak climbed during expeditions.
4. Filtering out incomplete data: Rows with missing values for key variables such as `sex` or `age` are filtered out to ensure the integrity of the dataset.
5. Final dataset creation: The resulting dataset is further refined to include only relevant columns (`peak_id`, `height_metres`, `season.x`, `sex`, `age`, `success`, `solo`, `died`). Some column names are adjusted for clarity, such as `season.x` being renamed to `seasons`, and `height_metres` being renamed to `height`.

6.2 Analysis dataset

Here is a glimpse of the dataset used for analysis

Table 2: Analysis dataset

peak_id	height_range	seasons	sex	age	success	solo	died
ACHN	5750 - 6099	Autumn	M	25	TRUE	FALSE	FALSE
ACHN	5750 - 6099	Autumn	M	23	TRUE	FALSE	FALSE
ACHN	5750 - 6099	Autumn	M	19	TRUE	FALSE	FALSE
ACHN	5750 - 6099	Autumn	M	22	TRUE	FALSE	FALSE
ACHN	5750 - 6099	Autumn	M	29	TRUE	FALSE	FALSE
ACHN	5750 - 6099	Autumn	M	60	TRUE	FALSE	FALSE

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.