Himalayan Expeditions Success Analysis*

Analysis using data from XXX leveraging Bayesian Logistic Regression

Kaavya Kalani

April 13, 2024

Abstract

Table of contents

1 Introduction

1) broader context to motivate;

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

2) some detail about what the paper is about;
3) a clear gap that needs to be filled;
4) what was done;
5) what was found;
6) why it is important;
7) Estimand

(Likely 3 or 4 paragraphs, or 10 per cent of total.)

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 analysis data is from XXXX. This is how it was collected.

A person becomes an entry in this dataset if XXX. (Measurement)

The data from datasets for XXX are combined to form our anaysis dataset. Among the overall range of variables available, we chose XXX to be included in our analysis dataset.

- Describe the variables

Out of these, we will be using XXX in our model as the independent variables and XXX as the dependent variable.

3 Model

I used a Bayesian Logistic Regression model to do XXX. 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 first used to model XXX. The model will be based on XXX independent demographic variables: XXX.

The logistic regression model I will be using is:

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

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

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
- β_3 is the coefficients corresponding to age
- β_4 is the coefficients corresponding to seasons
- β_5 is the coefficients corresponding to solo

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	-0.75
sexM	0.25	0.03	0.21	0.30
seasonsSpring	0.23	0.02	0.20	0.25
seasonsSummer	-0.44	0.09	-0.59	-0.29
seasonsWinter	-0.65	0.05	-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.

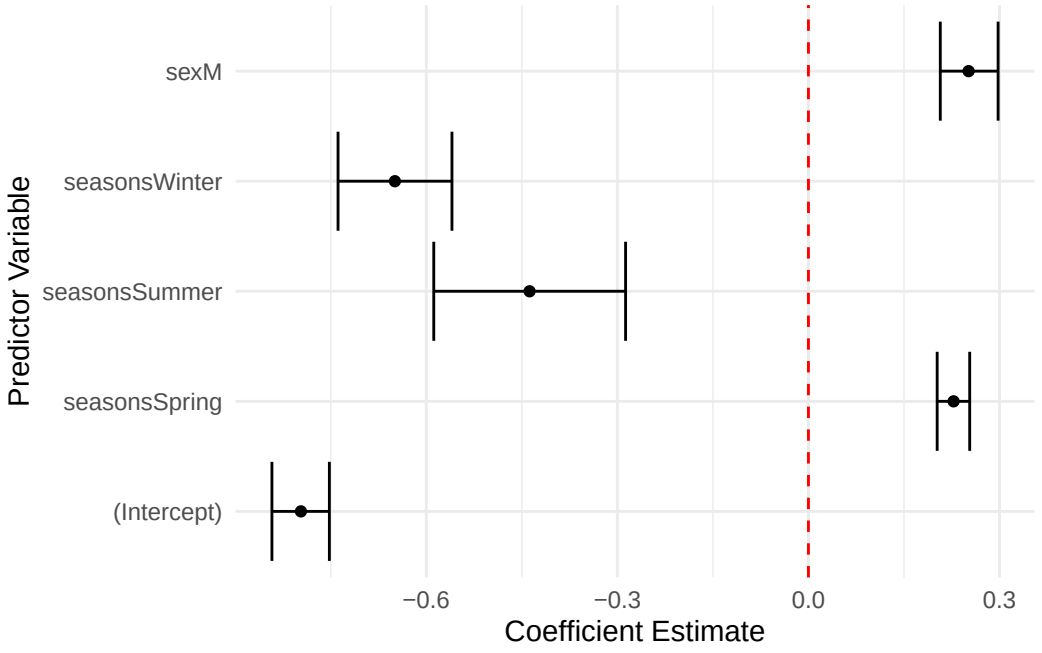


Figure 1: Coefficients of the model

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.

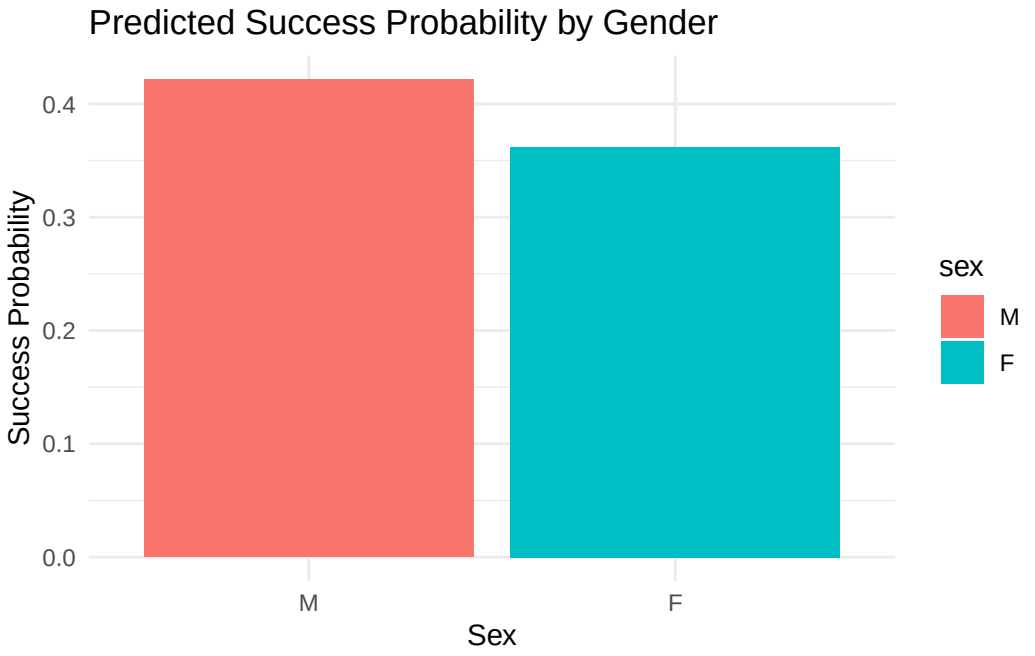
4 Results

```
# Create a data frame with all combinations of sex and seasons
sex_levels <- c('M', 'F')  # Replace with actual categories
seasons_levels <- unique(analysis_data$seasons)  # Extract unique seasons from your analys

new_data <- expand.grid(
  sex = sex_levels,
  seasons = seasons_levels
)

# Predicting success based on sex and seasons
predictions <- predict(expeditions_model, newdata = new_data, type = "response")

# Create a data frame to display the predictions
prediction_df <- data.frame(
  sex = new_data$sex,
  seasons = new_data$seasons,
  success_probability = predictions
)

# Print the predictions
print(prediction_df)
```

```
  sex seasons success_probability
1   M  Autumn           0.3670160
2   F  Autumn           0.3107092
3   M  Spring           0.4213689
4   F  Spring           0.3614783
5   M  Winter           0.2325879
6   F  Winter           0.1907263
7   M  Summer           0.2726392
```

```
8   F  Summer           0.2257269
```

```r
ggplot(prediction_df, aes(x = sex, y = success_probability, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Predicted Success Probability by Gender",
       x = "Sex",
       y = "Success Probability") +
  theme_minimal()
```



5 Discussion

1) What is done in this paper?
2) What is something that we learn about the world?
3) What are some weaknesses of what was done? What is left to learn or how should we proceed in the future?

5.1 Weaknesses and Limitations

5.2 Future Directions

6 Appendix

6.1 Analysis dataset

Here is a glimpse of the dataset used for analysis

Table 2: Analysis dataset

peak_id	height	seasons	sex	age	success	solo	died	members
ACHN	6055	Autumn	M	25	TRUE	FALSE	FALSE	5
ACHN	6055	Autumn	M	23	TRUE	FALSE	FALSE	5
ACHN	6055	Autumn	M	19	TRUE	FALSE	FALSE	5
ACHN	6055	Autumn	M	22	TRUE	FALSE	FALSE	5
ACHN	6055	Autumn	M	29	TRUE	FALSE	FALSE	9
ACHN	6055	Autumn	M	60	TRUE	FALSE	FALSE	9

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.