Final Project

Blue Team: Kevin Shah, Eddie Kong, Kelyce Allen, and Aram Lindroth

11/15/2020

Introduction

Data

This data was retrieved from the TidyTuesday GitHub repository called Himalayan Climbing Expeditions. It was originally taken from the Himalayan Database, which collects data from all expeditions that have been conducted in the Himalayas. According to the Himalayan Database website, much of the data was originally collected by an archivist named Elizabeth Hawley through her interviews with expedition teams both before and after they traveled. The comprehensive notes compiled from all of her interviews were "supplemented by information gathered from books, alpine journals and correspondence with Himalayan climbers" (The Himalayan Database). The dataset's earliest expedition was in 1905, and its most recent expeditions were conducted in 2019.

(Information collected from TidyTuesday Himalayan Climbing Expeditions, and the Himalayan Database history page. The links may not function in Gradescope.)

There are three different `.csv` files in this repository which we would like to gather data from. The `peaks.csv` file contains information on the different mountain peaks. Relevant variables include:

- `peak_name`: Common name for peak
- `height_metres`: Height of peak in metres
- `climbing_status`: Whether the peak has been climbed

The `expeditions.csv` file contains many variables surrounding aspects of the expedition, such as expedition start and terminations dates, the years and seasons of expeditions, and more. Relevant variables include:

- `year`: Year of expedition
- `members`: Number of expedition members. For expeditions in Nepal, this is usually the number of foreigners listed on the expedition permit. For expeditions in China, this is usually the number of non-hired members.
- `season`: Season of year when the expedition was conducted
- `hired_staff`: Number of hired expedition members
- `basecamp_date`: Date of expedition arrival at basecamp
- `trekking_agency`: Name of the trekking agency, or `NA` if no agency was involved
- `highpoint_date`: Date of expedition summiting the peak for the first time or, if peak wasn't reached, date of reaching its highpoint
- `member_deaths`: Number of expeditions members who died
- `hired_staff_deaths`: Number of hired staff (e.g. from a trekking agency) who died

- `termination_reason`: Reason for which the expedition was terminated (e.g. successfully completed expedition, injury, death, accident, poor weather, etc.)

The `members.csv` file includes information about characteristics of the individuals completing the expeditions. Relevant variables in this data set include:

- `died`: Whether the person died
- `injured`: Whether the person was injured
- `age`: Age of the person
- `sex`: Sex of the person
- `hired`: Whether the person was hired

Although we did not list them in the list of relevant variables above, in the course of our analysis, we may also use the `_id` variables to merge datasets.

Research

The broad goal of our research is to determine which factors most strongly correlate with climbers getting injured or dying during an expedition. We plan to perform analyses both on the `members` dataset (to analyze how demographic characteristics may correlate with injury and death) as well as on the `expeditions` dataset (to understand non-demographic factors that may correlate with the occurrence of accidents).

Some specific relationships we plan to investigate include the following:

- Does the height of a peak correlate with how dangerous it is? Our hypothesis is that higher peaks are more dangerous, i.e. a higher proportion of expeditions are fatal on higher peaks than on lower peaks.
- Can individual characteristics like sex, age, and whether an individual was hired or not predict the probability of death or injury among climbers? Our hypothesis is that all of these factors influence the probability of death or injury. This is a very important and informative part of our study, as it is important to learn how much individual characteristics can account for probability of death or injury. Notably, for this analysis, we focus only on demographic factors, not on factors that may be within an individual's control.
- How do factors like the season of an expedition, the year of the expedition, the number of members on the expedition, and whether the expedition used hired guides affect the probability of the expedition being fatal (i.e. at least one individual, either a member or hired member, dying on the expedition)? Our hypothesis is that the probability of fatality will decrease with year, but will increase with number of members and whether the expedition used hired guides, because these variables may correlate with other factors that more strongly affect odds of fatality. For example, it stands to reason that expeditions on more dangerous mountains will be more likely to use hired guides. We also hypothesize that expeditions in the winter will be the most dangerous. (Note that unlike in the previous question, this question will involve analyzing non-demographic factors.)

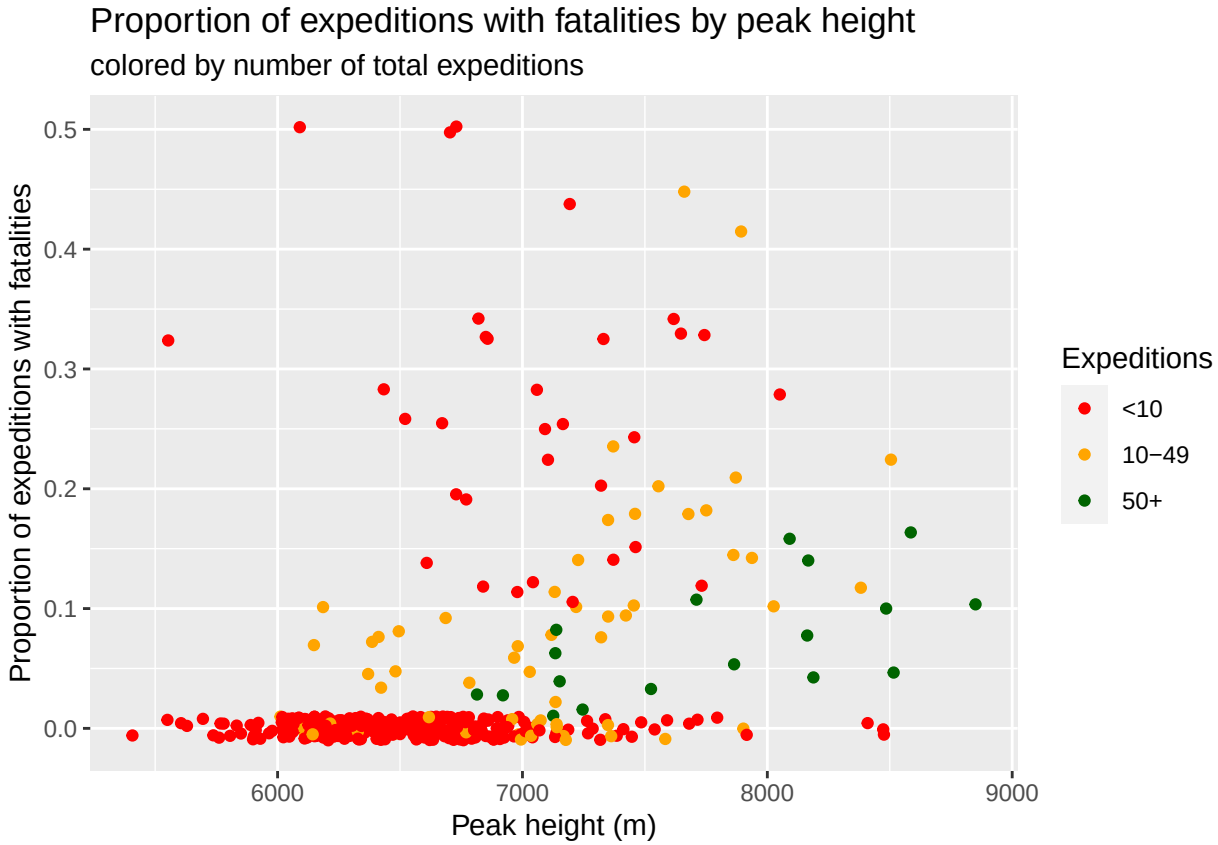
Methodology

For this project, we used a few different statistical tests. Our first relationship, testing whether the height of a peak correlates with how dangerous the peak is, initially used correlation coefficient R to find the strength of this relationship. Then, we fitted a linear model to predict the relationship between height and the proportion of expeditions that had at least on fatal accident. Our second relationship, testing whether individual characteristics (sex, age, hired/not hired) influenced the probability of death or injury among climbers, used two logistic models with injury/no injury and death/no death being the dependent binary variables. Our last relationship, testing whether expedition-level characteristics (year of expedition, season, hired guides or not) influenced the probability of the expedition being fatal, used logistic models similarly to the previous relationship. However, we are now using expedition-level characteristics instead of demographics.

Diagnostic plots for the linear models and logistics models can be found in the appendix. Relevant summary statistics and graphs for each relationship are provided in the corresponding sections below.

Peak Height

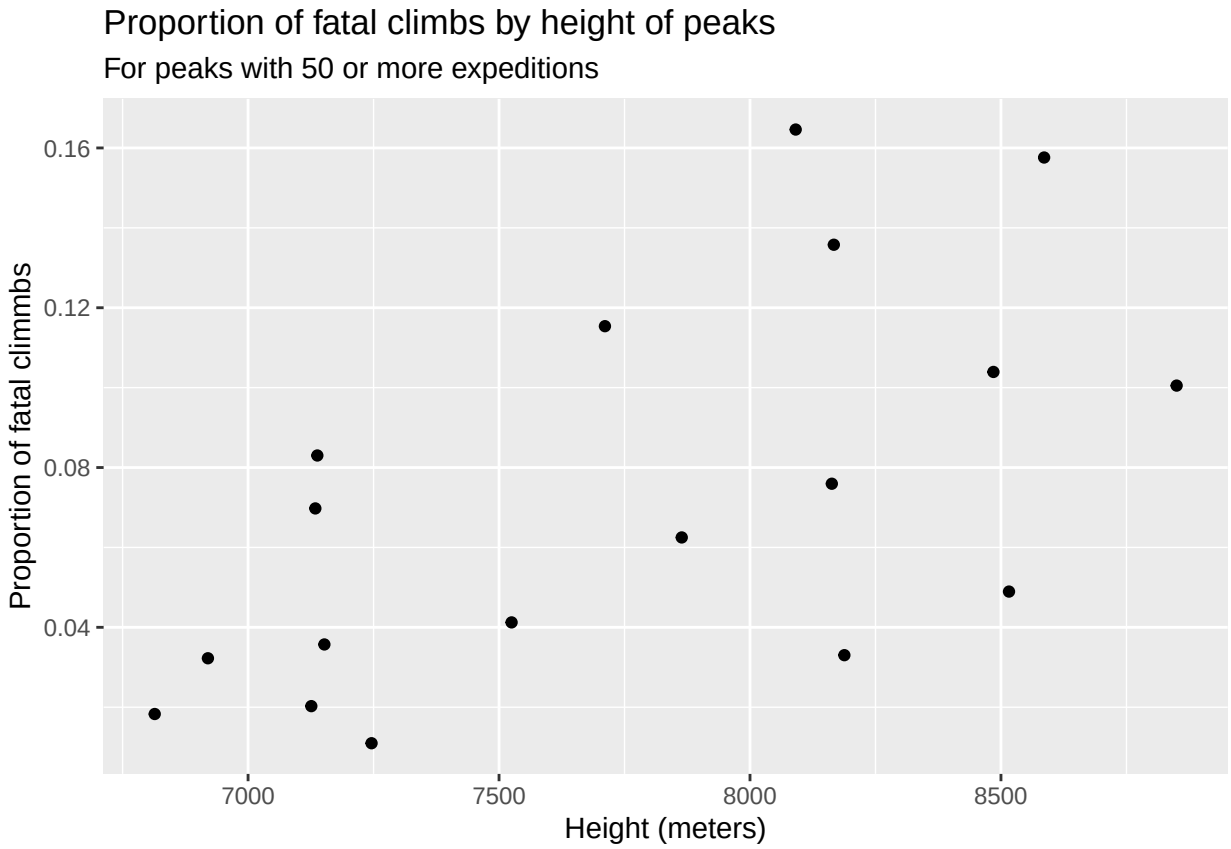
In this analysis, we are trying to determine whether there is a correlation between peak height and how dangerous a peak is to climb. Our hypothesis is that a higher proportion of expeditions are fatal on higher peaks than on lower peaks.



There appears to be some mild correlation between the height of a peak and the proportion of expeditions on that peak that were fatal (i.e. involved at least one death, either of a party member or hired staff). The point estimate of correlation is 0.369. Notably, analyzing only peaks with larger numbers of expeditions *increases* the point estimate of correlation between height and proportion of fatal expeditions: considering only peaks with more than 20 expeditions, the correlation is 0.462, and considering only peaks with more than 50 expeditions, the correlation is 0.605. This filtering does eliminate significant amounts of "noise" introduced by peaks with unreasonably high proportions of fatal expeditions simply due to the low number of total expeditions on that peak, however, it also eliminates a significant majority of peaks in the dataset. Of 390 total peaks, only 18 have 50 or more recorded expeditions, and these peaks have a higher average height than peaks with lower numbers of expeditions. However, the average proportion of fatal expeditions among peaks with less than 50 expeditions is actually lower (0.035) than the same proportion among other peaks (0.073), suggesting that the filtering actually removes many "safe" peaks too. That is, this filtering, while it may increase the strength of the correlation, does not seem to be decreasing the reliability of the model by skewing the data. Instead, it "un-skews" the data by filtering out the peaks with high proportions of fatal climbs but very few total climbs and ensures that our model, which models proportion, is computed using reliable proportions.

We will analyze the correlation between height and proportion of fatal accidents by fitting a linear model. A linear model looks reasonable given the visualization of proportion of fatal expeditions against height,

including only peaks with at least 50 expeditions:



Individual characteristics

In this analysis, we are looking to see whether certain characteristics of individual climbers, such as sex, age, and whether they were hired to complete the expedition or not, have an influence on probabilities of death or injury when climbing. Our hypothesis is that these characteristics do have an effect, and we will test these claims by comparing the raw proportions and by using a series of logistic models.

Raw proportions comparisons:

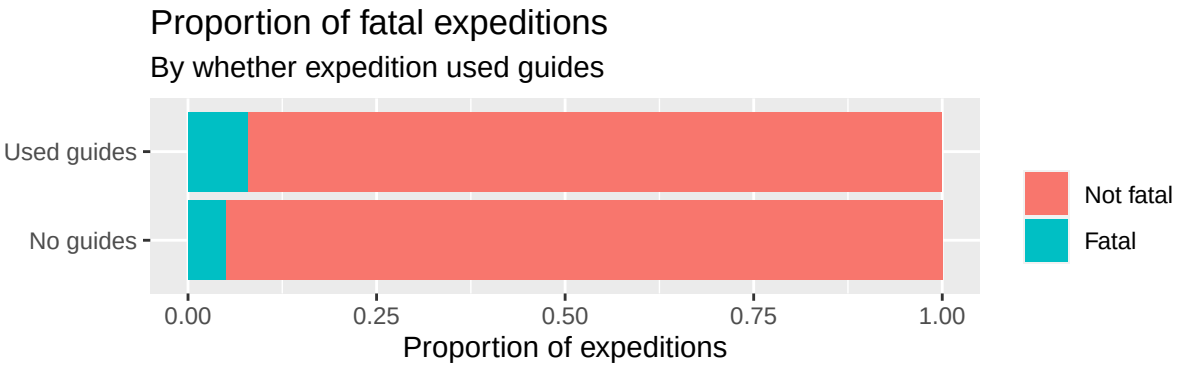
Surprisingly, climbers who died actually had a slightly lower mean age (36.03) than those who did not die (37.35). A higher proportion of male climbers died (.0152) compared to female climbers (.0071), even though almost ten times more males participated in climbs. Similarly, a higher proportion of hired climbers died (0.020) than non-hired climbers (0.013), even though non-hired climbers outnumbered hired climbers almost 4 to 1.

However, when we consider injury instead of death, different trends emerge. The disparity between the sexes decreases significantly: 2.1% of women were injured, compared to 2.2% of men. The disparity between hired and non-hired climbers also changes: 2.5% of non-hired climbers were injured, compared to only 0.9% of hired climbers.

Given the markedly different relationships between these individual characteristics and death or injury, we will create separate logistic models predicting the probability of death and the probability of injury, each of which will consider hired status, sex and age. We will not consider other factors, like season of expedition and the peak of the expedition, in this model, because these are factors within an individual's control, and we are primarily interested in the effects of demographic factors for this portion of our analysis. These factors will be considered in the next analysis, regarding modeling the probability of an expedition being successful.

Expedition Fatality

This analysis, as mentioned previously, centers on determining how certain expedition-level (as opposed to individual-level) factors influence the odds of an expedition being fatal. It is worth noting at the outset, however, that this is one particular analysis where it is essential to differentiate between correlation and causation. For instance, we see that a greater proportion of expeditions that used guides were fatal compared to expeditions that did not used guides:



This by no means implies that using guides would make an expedition more dangerous (that seems ludicrous). Rather, this is likely a consequence of the fact that guides are predominantly hired on expeditions to more dangerous or more challenging peaks.

Hence, when we proceed to create the logistic model in the next section, we must be particularly careful to remember this distinction, and to recognize that, while it may be more plausible to make informal inferences about the implications of correlations for other analyses (e.g. individual characteristics), doing so for this analysis will be far more difficult.

Results

Peak Height

The fitted linear model for proportion of fatal expeditions based on the height of the peak being climbed, using only peaks with greater than 50 expeditions recorded in the `expeditions` dataset, is:

$$\hat{p} = -0.28 + (4.50 \times 10^{-5} \times \text{Height})$$

This linear model suggests that we expect the proportion of fatal accidents on a peak of height 0 to be -0.28, and for each additional meter of height, we expect the proportion of fatal accidents to increase by, on average, `4.50e-5` ($p = .0078$). We cannot check the residuals for independence, because we can't meaningfully examine the residuals in order of collection since each residual corresponds to a long-term proportion, not a single expedition. The residuals do have equal variance and appear to be scattered around the predicted line. The distribution of residuals is not quite normal, but shows no significant patterned deviation from a normal distribution. Using a log transform on the proportion of fatal accidents is not helpful: the transformation does not improve the normality and increases the residuals by approximately a factor of 10. (See the appendix for the diagnostic plots.)

One important thing to note is that the adjusted R^2 for this model is only 0.326. This suggests that even if there is a relationship between height and fatal expeditions, this relationship is not the most important factor or indicator of whether an expedition will be fatal. There likely are other factors that have a larger effect.

Individual Characteristics

The resulting logistic model for probability of injury is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -3.58 + (-1.29 \times \text{Hired}) + (-0.01 \times \text{Age}) + (0.22 \times \text{Male})$$
$$\implies \frac{\hat{p}}{1-\hat{p}} = 0.03 \times 0.28^{\text{Hired}} \times 0.99^{\text{Age}} \times 1.25^{\text{Male}}$$

The logistic model for probability of death is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -4.47 + (-0.10 \times \text{Hired}) + (-0.01 \times \text{Age}) + (0.68 \times \text{Male})$$
$$\implies \frac{\hat{p}}{1-\hat{p}} = 0.01 \times 0.90^{\text{Hired}} \times 0.99^{\text{Age}} \times 1.97^{\text{Male}}$$

The results of these models are not particularly surprising: as expected, the odds of both injury and death are lower for hired climbers than for otherwise similar non-hired climbers, and the odds of both injury and death are higher for males than for females. Notably, the odds of both injury and death decrease with age, although the magnitude of the decrease is very small (1% decrease in odds for each additional year of age, holding all else constant). All p-values for both the injury and death models are below 0.012, except for the coefficient for hired status in the model for fatality, where the p-value is 0.23.

Comparing the two models, however, yields some interesting results. While the odds of both injury and death are lower for hired climbers, the magnitude of the effect of one's hired status differs between the two models. That is, being a hired climber decreases one's odds of death by a notably smaller amount than it decreases one's odds of injury. And conversely, while males have higher odds of both injury and death, the magnitude of the effect of one's sex is notably greater on one's odds of death than it is on one's odds of injury.

Diagnostic plots for both of these models indicate that they are generally accurate. However, these plots also highlight the fact that both of these models exclusively return fitted probabilities below 0.05. This is not necessarily a failing in the model; rather, this suggests that it is simply very unlikely for a climber to be injured or to die. The range of fitted values that the model does return, however, matches the observed probabilities quite well (although the model for death is slightly more accurate than the model for injury).

Probability of Expedition Fatality

We fitted a logistic model for probability of an expedition being fatal using year of the expedition, season of the expedition, number of climbers on the expedition, and whether the expedition used hired climbers as predictor variables. The resulting model is

$$\begin{aligned}\log\left(\frac{\hat{p}}{1-\hat{p}}\right) =& 44.83 + (-0.02 \times \text{Year}) + (0.28 \times \text{Spring}) + (-0.86 \times \text{Summer}) + (-0.48 \times \text{Winter}) + \\ & (0.07 \times \text{Members}) + (0.34 \times \text{Guides}) \\ \implies \frac{\hat{p}}{1-\hat{p}} =& 2.95 \times 10^{19} \times 0.98^{\text{Year}} \times 1.32^{\text{Spring}} \times 0.42^{\text{Summer}} \times 1.62^{\text{Winter}} \times \\ & 1.07^{\text{Members}} \times 1.40^{\text{Guides}}\end{aligned}$$

According to this model:

- Holding all else constant, for every one year increase in expedition year, we expect the odds of an expedition member dying to be multiplied by $e^{-0.02} = 0.98$ ($p = 2.5 \times 10^{-26}$).
- Holding all else constant, we expect the odds of a spring expedition being fatal to be $e^{0.28} = 1.32$ times the odds of a fall expedition being fatal. ($p = 6.9 \times 10^{-4}$).

- Holding all else constant, we expect the odds of a summer expedition being fatal to be $e^{-0.86} = 0.42$ times the odds of a fall expedition being fatal. ($p = 0.099$).
- Holding all else constant, we expect the odds of a winter expedition being fatal to be $e^{0.48} = 1.62$ times the odds of a fall expedition being fatal. ($p = 0.015$).
- Holding all else constant, we expect the odds of an expedition being fatal to be multiplied by $e^{0.07} = 1.07$ for each additional member in the expedition ($p = 4.3 \times 10^{-40}$).
- Holding all else constant, we expect the odds of fatality for expeditions that used hired guides to be $e^{0.34} = 1.40$ times the odds of fatality for expeditions that did not use hired guides ($p = 2.1 \times 10^{-4}$).
- We expect the odds of success for expeditions in the fall of 0 CE with 0 members that did not use any guides to be $e^{44.83} = 2.95 \times 10^{19}$ ($p = 6.9 \times 10^{-23}$), which clearly does not make sense in the context of the problem: we will never have expeditions with 0 members from 0 CE.

This model is quite accurate for relatively smaller predicted probabilities of fatality, but deviates significantly from actual probabilities for higher predicted probabilities. While this suggests that perhaps a logistic model is not entirely appropriate, the fact that the deviation occurs in a region that contains very few observations suggests that at least a significant portion of the model is accurate, and that, ultimately, it is difficult (at least, with a logistic model) to conclude that any expedition has a probability of fatality significantly greater than 0.5. This may be more a reflection of the ultimate unpredictability of mountain climbing and the fact that perhaps these measurable predictors included in our dataset do not fully explain the variation in probability of fatality. (See the appendix for the diagnostic plot.)

Discussion

Through our research, we discovered what makes expeditions in the Himalayas successful and safe by determining which factors correlate most strongly with climbers getting injured or dying during an expedition. We used the data originally taken from the Himalayan Database, a collection from all expeditions in the Himalayas starting from 1905 to 2019. Using this dataset, we were able to analyze how specific relationships such as height of the peak, individual characteristics of the climber, and other non-demographic factors of an expedition influence the proportion and probability of fatality or injury.

We began by investigating how the height of the peak being climbed correlates to the proportion of fatal expeditions. It is important to note that for this data set, our adjusted R^2 value was 0.33. This means that peak height is not the best indicator of the proportion of fatal expeditions, and that there are likely other more telling indicators of whether an expedition will result in fatality. After filtering peaks with a low number of expeditions, we saw an increased correlation point estimate of 0.605. We created a linear model which returned a height coefficient of 4.5×10^{-5} meaning the proportion of fatality increases with increase in the height of the peak. Although this number seems very low and close to 0, this is the increase in proportion given just 1 meter increase in height of the peak. While this finding supported our hypothesis that increased height of peak increases the proportion of fatality, it could have been easier to interpret if we measured the increase in proportion for every 100 meter increase in height. Our coefficient would then have stated that we expect around 0.5% increase in proportion of fatality per 100 meter increase in height on average. The resulting p-value for this height coefficient was 0.0078. If we were to perform a hypothesis test on this height coefficient with an alpha of 0.05, we would be able to reject the null hypothesis that there is no increase in proportion of fatality with an increase in height. However, given that our dataset is the population data, it does not make much sense to perform this hypothesis test. Furthermore, it is important to note that while the residuals for the linear model comfortably met the equal variance and linearity conditions, they were not quite normally distributed.

After learning that there is a mild correlation between the peak height and proportion of fatality, we investigated whether different individual characteristics (gender, age, and hire status) of a climber influenced the probability of fatality or injury on their expedition. Using logistic models, we found that hired climbers are less likely to be injured or die on climbs when all else is held constant. We also found that men are more likely to become injured or die compared to women, and younger people are more likely to be injured or

die compared to old people. By comparing the logistic model of death and the logistic model for injury, we were able to observe how the probability differed between the two. The differences between the two were surprising to us, as we hypothesized that the probability for reach would be relatively similar. Hired climbers, for example, are slightly less likely to die than non-hired climbers, but they are much less likely to be injured. As far as gender, men are more likely to get injured on a climb compared to women, and are much more likely to die on a climb. These findings are supported by diagnostic plots, which also remind us that the likelihood of becoming injured or dying during a climb is extremely low in the first place.

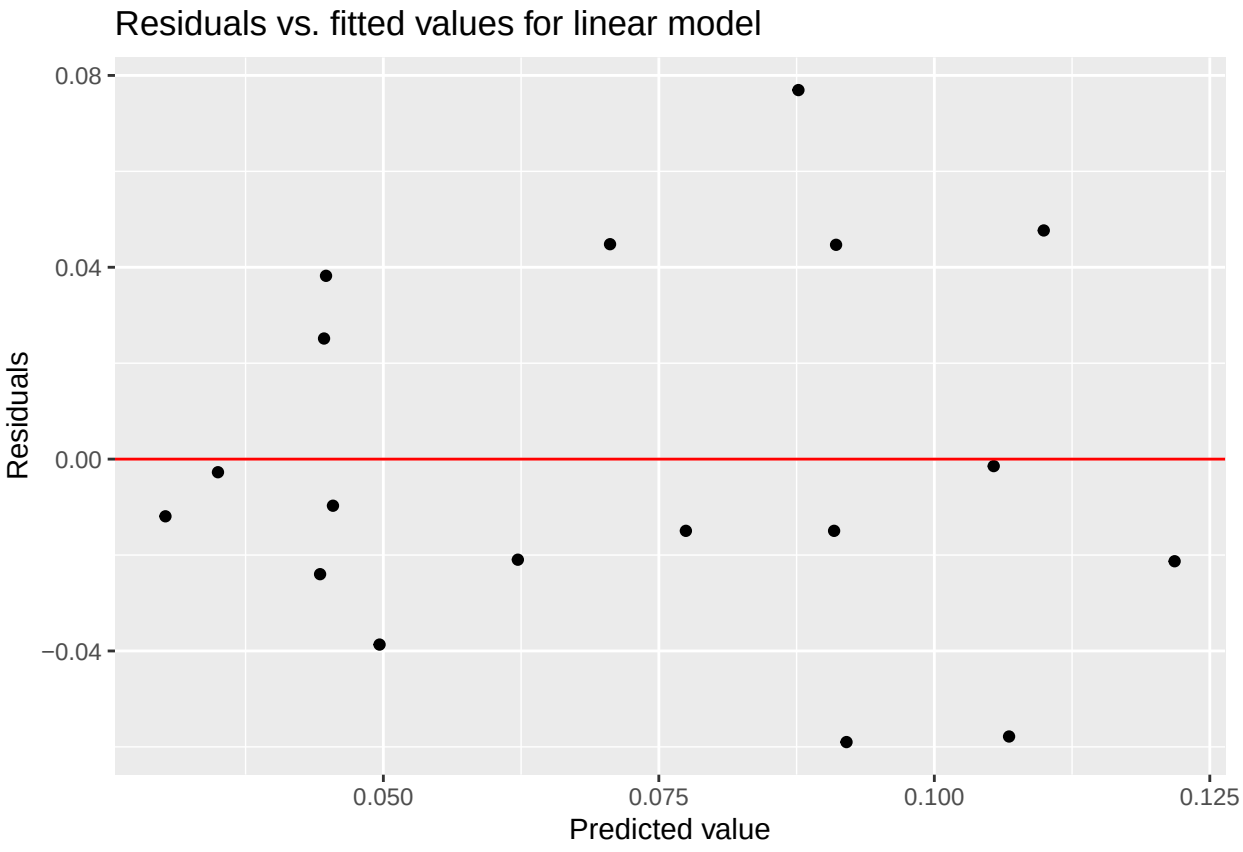
Finally, we analyzed whether there are non-demographic factors that affect the odds of an expedition being fatal. The non-demographic factors we analyzed included the year of expedition, season of expedition, number of climbers, and whether guides were present. Because we are using logistic models for this hypothesis, all of the following probabilities require us to hold all other variables constant. Under this assumption, the data suggests that probability of death on an expedition decreases very slightly each year, and, interestingly, the probability of fatality on an expedition varies by season. In order from highest to lowest probability of fatality (or most to least dangerous season), the order is winter, spring, fall, summer. It also suggests that the odds of an expedition being fatal increases slightly with each addition of a new climb member, and that the odds of fatality increase more significantly when an expedition is conducted with a hired guide.

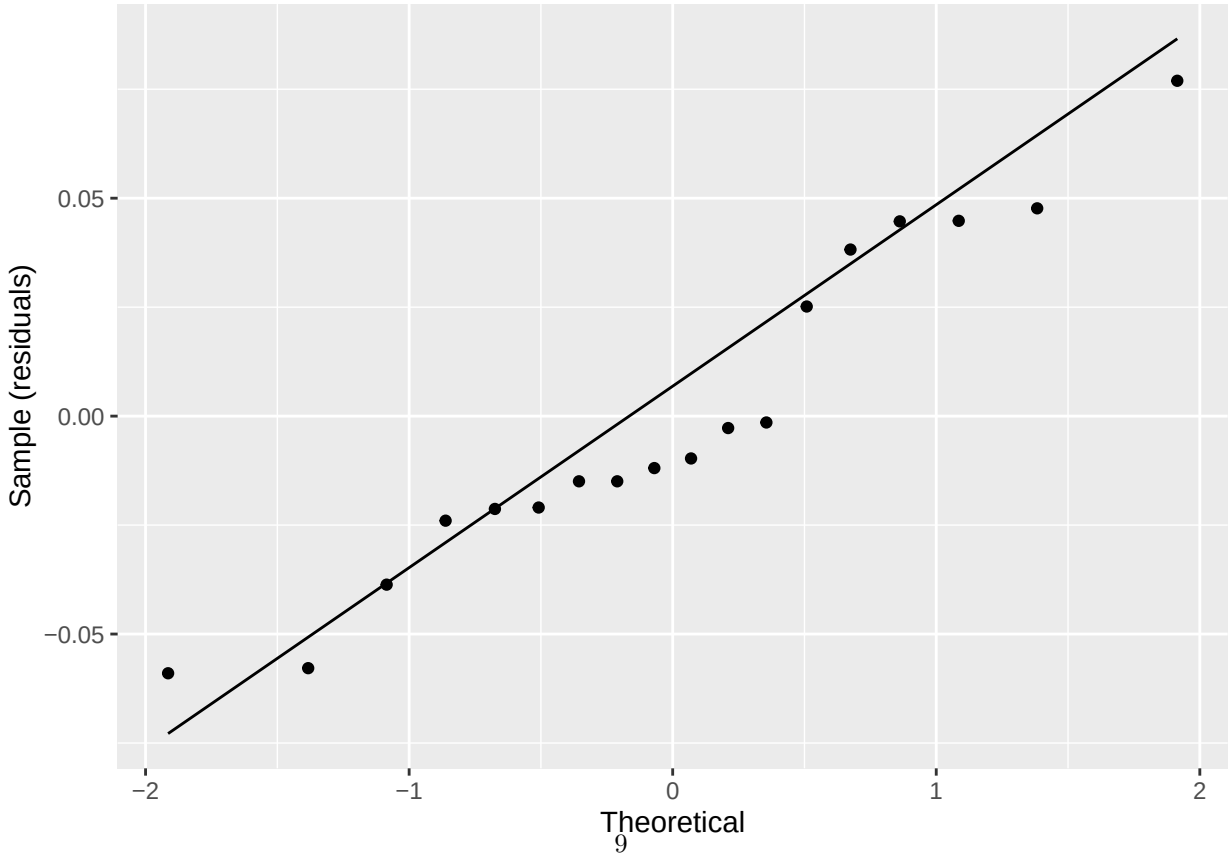
We must again note that the dataset used in this analysis contained data for the entire population of interest, according to the source of the data. (Whether this is actually the case is unclear, especially since the data collection methods described do not eliminate the possibility that expeditions, particularly ones from the early to mid 20th century, were not recorded. However, we operated under the assumption that the dataset did represent the entire population, given we had no concrete evidence to the contrary.) This means that hypothesis tests are not, strictly speaking, appropriate tools for analyzing this dataset because we do not need to use any statistical inference to make claims about population parameters—rather, we can directly compute population parameters from the data. Nonetheless, we include p-values as a means of trying to quantify the strength of our findings given observed variability in the data. That is, we expect the specific outcomes of certain expeditions (and hence values like proportions of expeditions that were fatal) to vary naturally, and the use of p-values allows us to in some way reflect the confidence we have in our findings relative to the natural expected variability of the parameter in question. However, ultimately, since these p-values are not essential to our statistical conclusions, it is not essential that our models meet all of the requirements needed to perform inference with the models; for instance, our linear model's non-normally distributed residuals do not invalidate the model, since we do not rely on using the model to make inferential conclusions.

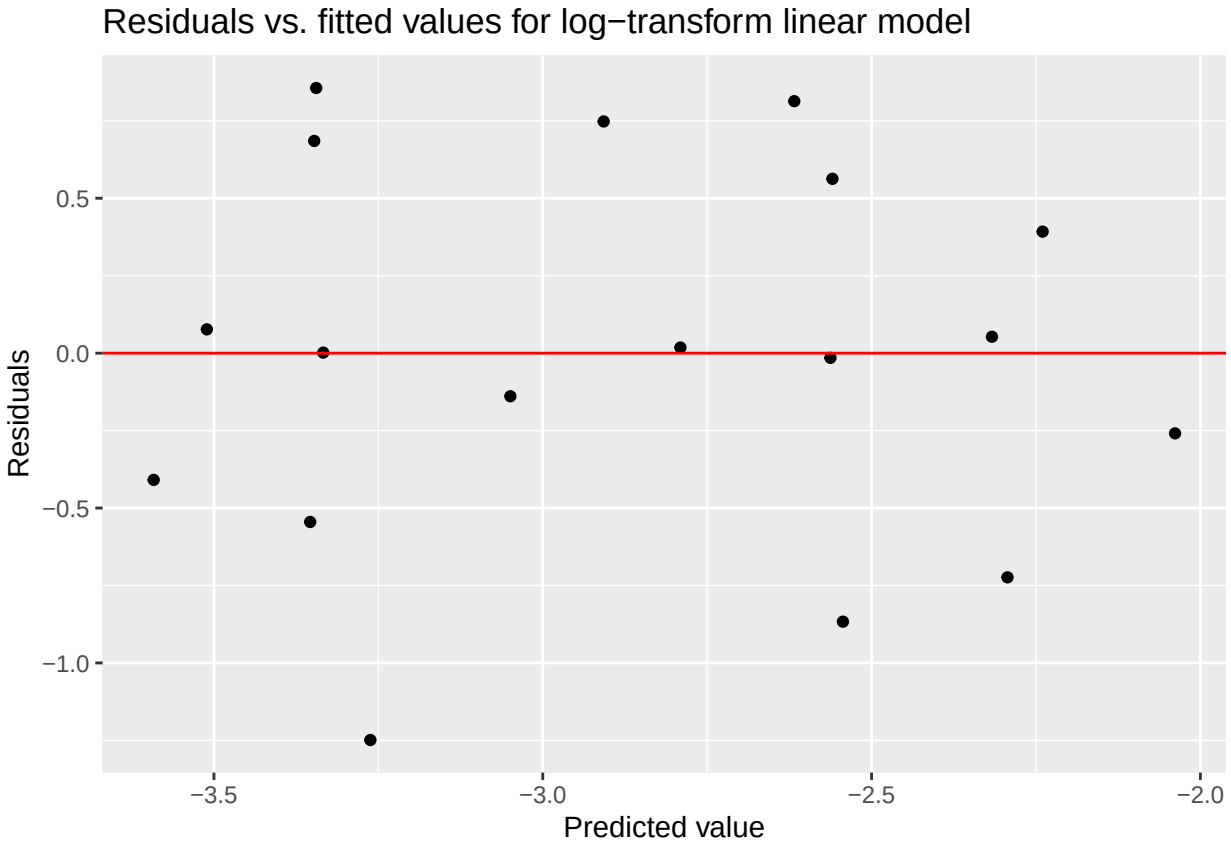
If we were able to do this project again, we would make more use of merging in order to associate individual members (in the `members` dataset) with the expeditions they participated in (in the `expeditions` dataset). This would allow us to associate injury data with expeditions, whereas the `expeditions` dataset on its own only includes data about deaths. We would also try to apply string manipulation tools to analyze some variables in the dataset that currently appear either as strings or as factors with many levels, like `termination_reason` in the `expeditions` dataset or `expedition_role` in the `members` dataset. Finally, we would try to incorporate more statistical tools beyond linear and logistic models; we might perform, for instance, a chi-square test to determine whether cause of death is independent of whether an expedition was a solo expedition.

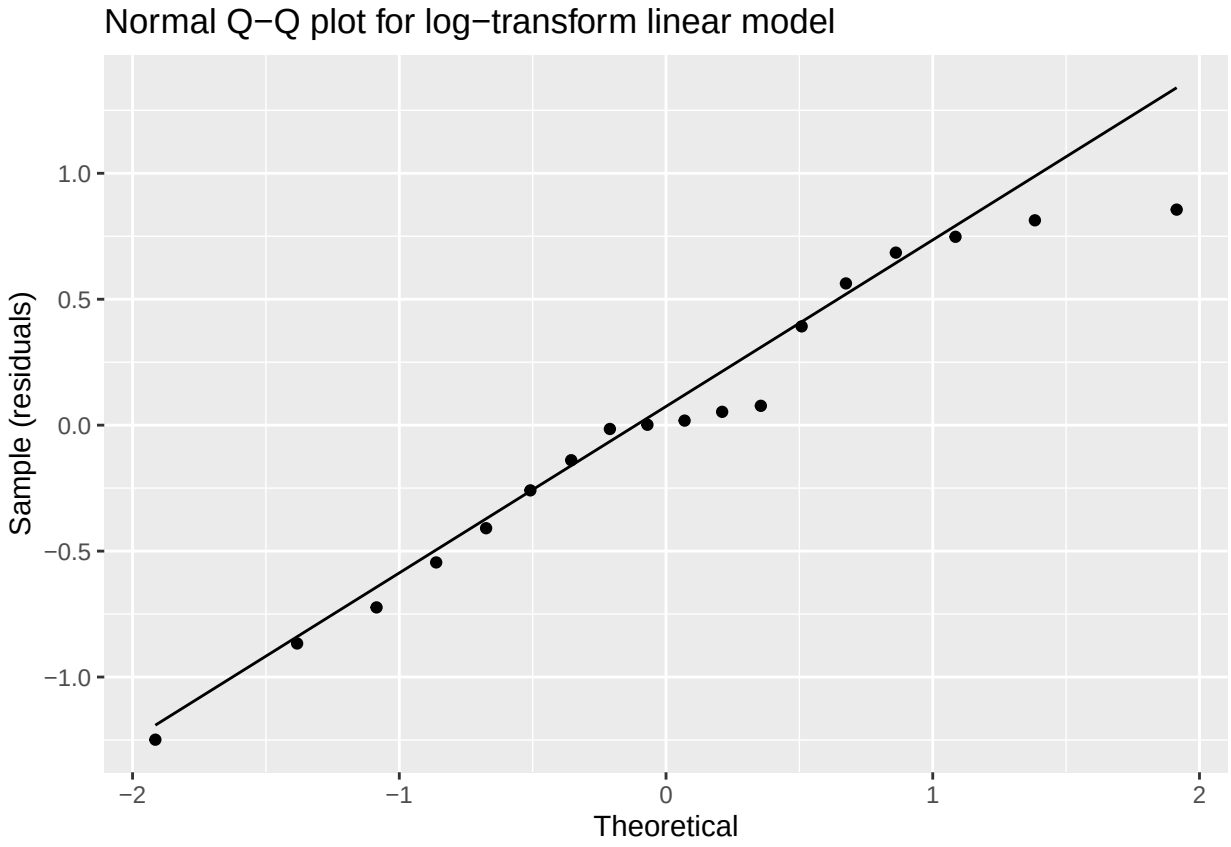
Appendix: Diagnostic Plots

Peak Height Linear Model



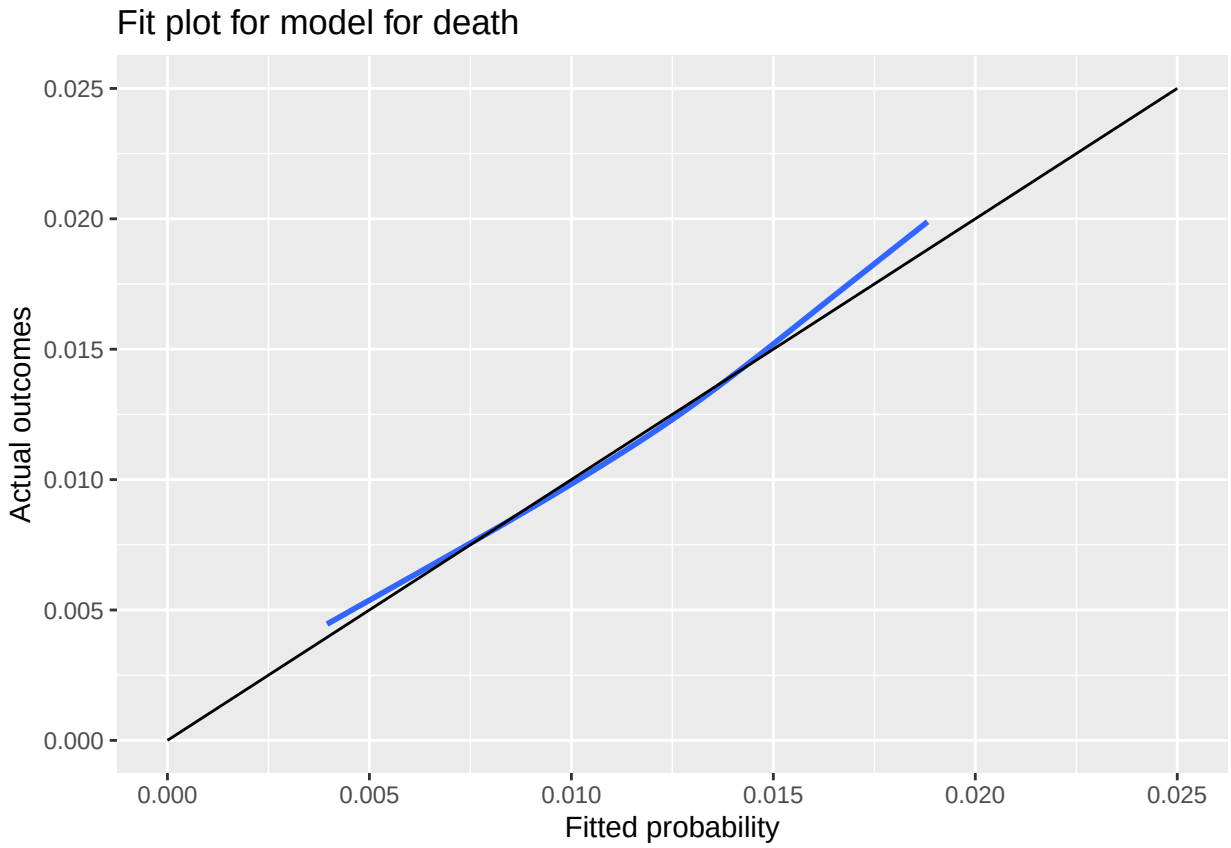



Residuals vs. fitted values for log−transform linear model
Residuals
0.5
0.0
−0.5
−1.0
−3.5
−3.0
−2.5
−2.0
Predicted value


Normal Q−Q plot for log−transform linear model
Sample (residuals)
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Theoretical

Individual Characteristics Models

Death



Injury


Fit plot for model for injury
Actual outcomes
Fitted probability
0.00
0.01
0.02
0.03
0.04
0.05

Expedition Fatality Logistic Model