BDA - project report

1. Introduction

1.1 Motivation

Climbing Everest is considered one of the toughest expeditions because of the various difficulties climbers face while trying to reach for the summit. In the past couple of years, the success of climbing over 8000 metre mountains depends on a various number of factors: supplemental oxygen usage, Sherpa support, equipment quality and logistics.

One of the most challenging parts of reaching the summit of such mountains is that between 7000 and 8000 metres, climbers enter the **death zone**. The death zone refers to altitudes above a certain point where the pressure of oxygen is insufficient to sustain human life for an extended time span. Many deaths in high-altitude mountaineering have been caused by the effects of the death zone, either directly by loss of vital functions or indirectly by wrong decisions made under stress, or physical weakening leading to accidents.

In our project, we analyse how supplemental oxygen and better alpinism equipment affect the success rate of climbing over 8000 metre mountains.

1.2 The problem

In the last couple of decades, very high altitude mountaineering has become more popular and more commercial. In 2019, the long queues near the Everest summit caused many deaths even though most climbers had Sherpa and oxygen support. This issue can be due to the fact that alpinists who don’t use supplemental oxygen support or run out of it end up spending too much time in the death zone because of the long queues.

1.3 Main modelling idea

We will analyse the success rates in reaching the Everest summit. For a range of 40 years, we are going to estimate the cumulative success rate depending on the number of successful expeditions out of the total number of attempts in the past. We assume that the supplemental oxygen is an aid for climbing the peak and that equipment and knowledge improve as time passes. Therefore, we have chosen oxygen support and the year of the expedition as the two features predicting the success rate.

2. Description of the data and the analysis problem

2.1. Dataset and preprocessing

Our dataset is named Himalayan climbing expeditions from Kaggle. Based on our knowledge, this dataset is not used for any kind of Bayesian analysis and we did not find any work related to this dataset. The dataset contains 10363 rows where each row is described by the following features:

expedition_id, peak_id, peak_name, year, season, basecamp_date, highpoint_date, termination_date, termination_reason, highpoint_metres, members, member_deaths, hired_staff, hired_staff_deaths, oxygen_used, trekking_agency

We needed to preprocess the data, as the initial dataset contains many variables which we did not plan to use in our model. The steps we took for preprocessing were the following:

- we only selected expeditions for the Everest peak, as it had the biggest number of expeditions during the time range we have chosen,
- for each year (1981 - 2019) we will count the number of attempts, and divide observations for that year into two parts: 1) the success rate for the attempts *that did not use* supplemental oxygen; 2) the success rate for the attempts *that did use* supplemental oxygen;
- for years that did not have any attempts or did not have any mountaineers that used supplemental oxygen, we inputted the 0 value for that year (number of attempts = 0, success rate = 0, oxygen = 0 or 1 depending on the missing value).

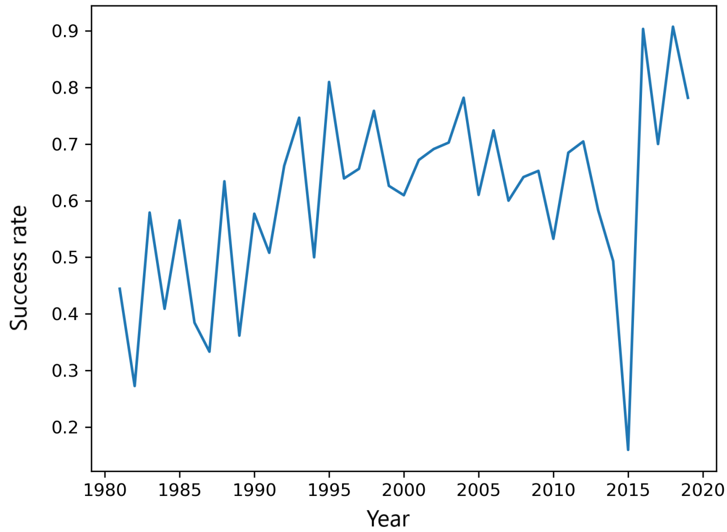
The resulting dataset we get after performing these steps has the following structure:

	year	oxygen_used	attempts	success_y
1	1981	0	2	0
2	1981	1	2	1
3	1982	0	2	0
4	1982	1	4	2
5	1983	0	4	1
6	1983	1	6	3
7	1984	0	3	1

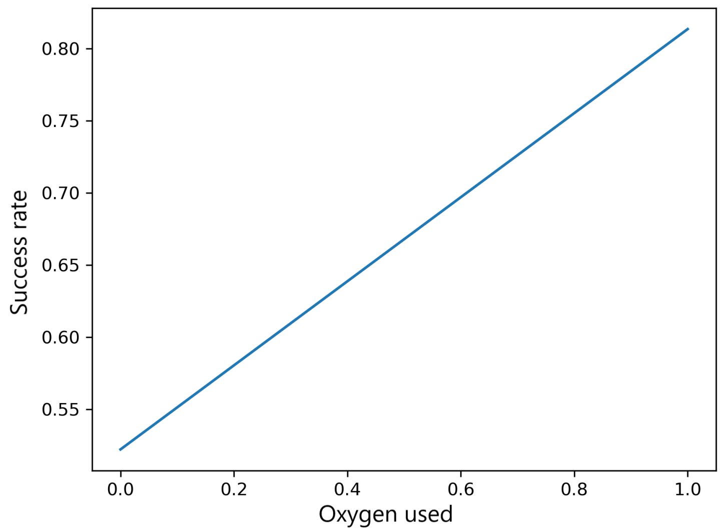
For each year, we have two lines describing

1. the number of oxygen supplemented expeditions (attempts) and the number of successful ones out of them
2. the number of not oxygen supplemented expeditions (attempts) and the number of successful ones out of them

The next step was to perform visual analysis on the raw data. Using the generated plots, we were able to reinforce our knowledge about the domain and check if the choice of our priors is reasonable. We have the following findings:



1. Success rate increases as years pass. This could be due to the fact that people benefit from the advancements in alpinism equipment and the increasing number of touristic climbing operators. While in the ‘80s the average success rate was around 30-40% we see that in the past decade it had risen up to around 80-90%.



2. Oxygen supplement is a key factor of the success in climbing the mountain. In this plot we can see that supplementary oxygen users have around 80% chance of success, while non-supplementary oxygen users have around 50% chance.

3. Description of the chosen models

For the model, it is natural to interpret each expedition as a Bernoulli trial with an unknown success probability θ, which we will be trying to infer through Bayesian analysis. Thus, it is straightforward to choose the binomial distribution to model our dataset.

Non Hierarchical Linear model (Bayesian linear regression)

We have first started with a linear model which expresses the success rate variable θ as a linear combination of the two factors considered: the year of the expedition and whether or not supplemental oxygen was used (and an intercept). However, since the success rate is a probability, its value must be between 0 and 1 and we have to account for this by regularising it through our choice of priors (this is discussed in greater detail in the next section).

We start just with the normal linear combination of the features:

$$\theta = a + w_1 \cdot year + w_2 \cdot oxygen$$

where year represents the year of the expedition and oxygen is a variable which takes the value 1, if additional oxygen was used and 0 otherwise. The intercept a can be seen as the success rate of an expedition taken at the beginning of our time range (1980) and without using additional oxygen.

However, to make the choice of the prior easier and more natural to interpret, we modified the formula above such that the value of three parameters: a, w_1 and w_2 can be seen as a percentage of its contribution to the success rate. For this we scaled the whole sum by 100 and converted the year to unit range ($year' = \frac{year - year_{min}}{year_{max} - year_{min}}$). Thus, we ended up with this formula:

$$\theta = \frac{a + w_1 \cdot \frac{year - year_{min}}{year_{max} - year_{min}} + w_2 \cdot oxygen}{100}$$

Hierarchical model using sigmoid function (Bayesian logistic regression)

For our choice of features, expressing the success rate as a linear combination has some drawbacks. Since the *oxygen* variable only takes value 0 or 1, its contribution to the model is basically just changing the intercept of the line (i.e. the rate of success at the beginning of our time range). Thus, this model will only obtain two parallel lines, as it can be seen from the plot in the section 7 - *Posterior Predictive Checks*. Therefore, we decided to improve this by considering a hierarchical model based on the fact that the oxygen was used or not. In consequence, depending if the oxygen was utilised, we used different priors for the intercept and year weight (w_1) variables, but with hyperparameters samples from common hyperdistribution.

Another change we did to improve the previous model was applying the sigmoid function to the linear combination of the features, to ensure that the success rate probability is always between 0 and 1 (thus avoiding the need of regularisation through priors). It is also expected for the sigmoid function to perform better for modelling the improvement of the success rate over the years, since it will be able to replicate the diminishing returns that this feature will have over a longer timespan.

Consequently, our formula for the success rate for this model is the following:

$$\theta_{oxygen} = sigmoid\left(\frac{a_{oxygen} + w_{oxygen} \cdot \frac{year - year_{min}}{year_{max} - year_{min}}}{100}\right)$$

, where the sigmoid function has the form $sigmoid(x) = \frac{1}{1+e^{-x}}$ and the index $oxygen$ has the role of suggesting that we use different variables for a and w depending if additional oxygen was utilised.

4. Informative or weakly informative priors chosen

Choosing a prior for Bayesian analysis is usually done by using expert knowledge on the subject, while also taking into account the specificities of the model. In our case, although our two models express the same events, our two models still have different specific requirements, which we need to account for in our choice of priors.

Non Hierarchical Linear model

Firstly, as mentioned before, we need to make sure that the weighted sum from our linear model cannot go under 0 or exceed 100. Taking this into consideration, we have chosen the following priors:

- intercept ~ N(30, 7)
- year_weight ~ N(7, 3)
- oxygen_weight ~ N(15, 5)

From consulting available evidence on the topic, we discovered that 30% success rate would be a good guess for the mean of the intercept (which can be interpreted as the probability of climbing Everest, without additional oxygen in the 1980). We used a standard deviation of 7, in order to not make the prior too informative, while also regularising the model to not get negative or unreasonably low values as a success rate.
Similarly, our research shows that the more recent climbs had better chances of climbing Everest than in the past and, therefore, we have used a normal prior with positive mean at 7 and 3 standard deviation, such that the values are not very likely to go negative. Applying a similar logic, we have chosen a normal distribution with the mean 15 and 5 standard deviation as the

prior for using oxygen. The only difference is that this prior is much more unlikely to allow negative values (than the one for the year weight), which makes sense since there is no reason why the success rate would decrease from the use of additional oxygen.
One thing to note is that despite the fact that these priors seem to be a lot more strongly informative than what we used in the assignments of this course, they are actually weakly informative since they cover a wide range of the possible success rate: between 5% and 95%. This allows enough freedom for the data to shape the posterior.

Hierarchical model

For the hierarchical model we had more freedom in choosing the priors since we didn't have the need to ensure such tight regularisation. On the other hand, we need to take into account that on the range [-6, 6], the sigmoid function covers (almost) all its domain [0, 1] and has the value 0.5 at x = 0. Hence, since we divide the weighted sum by 100, we chosen the priors bellow:

- hyper mean of the intercept ~ N(0, 100)
- hyper standard deviation of the intercept ~ N(70, 20)
- hyper mean of the year weight ~ N(5, 30)
- hyper standard deviation of the year weight ~ N(70, 20)
- intercept ~ N(hyper mean of the intercept, hyper standard deviation of the intercept)
- year weight ~ N(hyper mean of the year weight, hyper standard deviation of the year weight)

We chose the N(0, 100) for the hyperprior of the mean of the intercept, since $sigmoid(0) = 0.5$ and the high standard deviation allows enough variation to be weakly informative. The choice of N(5, 30) for the mean of the year weight needs to be interpreted as expecting a slight increase in the success over the year, but allowing a generous range from negative values to big positive ones.
We used the N(70, 20) for both the hyper-priors of the standard deviation of the intercept and the year weight, which is informative enough only to regularise it, by making sure it stays positive (standard deviation should always be positive since it is a squared value).
These priors ensure with large confidence that the sum ($\frac{a_{oxygen} + w_{oxygen} \cdot \frac{year - year_{min}}{year_{max} - year_{min}}}{100}$) covers the [-6, 6], without going outside much, which means that the whole range from 0 to 1 of the success rate is covered. On the other hand, it is clear that they are weakly informative enough to allow the data to affect the posterior distribution.

5. Stan code for the models and code explanation

i. Linear model:

The following R cell is used to read the data from the generated file (sec. 2.1) for the Linear model.

```r
everest_data = read.delim("out_everest.csv", header = TRUE, sep = ",")
stan_data <- list(
  N = length(everest_data$oxygen_used),
  success = everest_data$success,
  attempts = everest_data$attempts,
  year = everest_data$year,
  oxygen_used = everest_data$oxygen_used
)
```

Stan code:

```stan
data {
  int<lower=0> N;
  int success[N];
  int attempts[N];
  int year[N];
  int oxygen_used[N];
}

parameters {
  real intercept;
  real year_weight;
  real oxygen_weight;
}

transformed parameters {
  real theta[N];
  for(i in 1:N)
    theta[i] = (intercept + year_weight * (year[i] - min(year))/(max(year)
- min(year)) + oxygen_weight * oxygen_used[i])/100;
}

model {
  // priors
  intercept ~ normal(30, 7);
  year_weight ~ normal(7, 3);
  oxygen_weight ~  normal(15, 5);

  // likelihood
  success ~ binomial(attempts, theta);
```

```
}

generated quantities {
  vector[N] log_lik;
  int y_rep[N];

  for(i in 1:N) {
    y_rep[i] = binomial_rng(attempts[i], theta[i]);
    log_lik[i] = binomial_lpmf( success[i] | attempts[i], theta[i]);
  }
}
```

The following R cell is used to run the Linear model.

```
fit_result <- stan(file = "model_independent_features.stan", data =
stan_data)
```

ii. Hierarchical model:

The following R cell is used to read the data from the generated file (sec. 2.1) for the Hierarchical model.

```
everest_data = read.delim("out_everest.csv", header = TRUE, sep = ",")

success1 <- matrix(everest_data$success[seq(2,
length(everest_data$success), by = 2)])
success2 <- matrix(everest_data$success[seq(1,
length(everest_data$success), by = 2)])

attempts1 <- matrix(everest_data$attempts[seq(2,
length(everest_data$attempts), by = 2)])
attempts2 <- matrix(everest_data$attempts[seq(1,
length(everest_data$attempts), by = 2)])

year1 <- matrix(everest_data$year[seq(2, length(everest_data$year), by =
2)])
year2 <- matrix(everest_data$year[seq(1, length(everest_data$year), by =
2)])
```

```
success = cbind(success1, success2)
attempts = cbind(attempts1, attempts2)
year = cbind(year1, year2)
```

Stan code:

```
data {
  int<lower=0> N;
  int<lower=0> M;
  int success[N, M];
  int attempts[N, M];
  int year[N, M];
}

parameters {
  real hyper_intercept_mean;
  real hyper_intercept_sd;
  real hyper_year_weight_mean;
  real hyper_year_weight_sd;
  real intercept[M];
  real year_weight[M];
}

transformed parameters {
  real theta[N, M];

  for(i in 1:N)
    for(j in 1:M)
    {
      theta[i, j] = 1/(1+exp( - (intercept[j] + year_weight[j] * (year[i,
j] - min(year[, j]))/(max(year[, j]) - min(year[, j])) )/100 ));
    }

}


model {
  // hyper prior
  hyper_intercept_mean ~ normal(0, 100);
```

```
  hyper_intercept_sd ~ normal(70, 20);
  hyper_year_weight_mean ~ normal(5, 30);
  hyper_year_weight_sd ~ normal(70, 20);

  // priors
  for (i in 1:M){
    intercept[i] ~ normal(hyper_intercept_mean, hyper_intercept_sd);
    year_weight[i] ~ normal(hyper_year_weight_mean, hyper_year_weight_sd);
  }

  // likelihood
  for(i in 1:M)
    success[, i] ~ binomial(attempts[, i], theta[, i]);
}

generated quantities {
  matrix[N, M] log_lik;
  int y_rep[N, M];

  for(i in 1:N)
    for(j in 1:M){
      y_rep[i, j] = binomial_rng(attempts[i, j], theta[i, j]);
      log_lik[i, j] = binomial_lpmf( success[i, j] | attempts[i, j],
theta[i, j]);
    }
}
```

The following R cell is used to run the Hierarchical model:

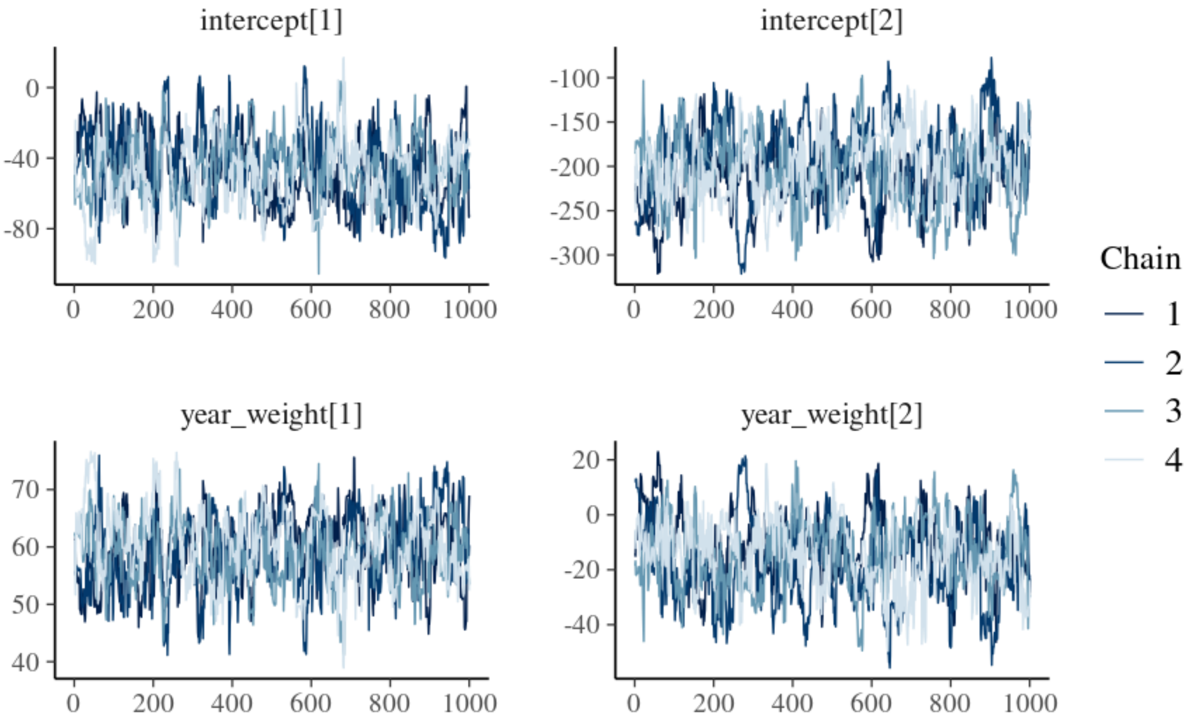
```
fit_result <- stan(file = "hierarchical_model_oxygen_logit.stan", data =
stan_data)
```

6. Convergence diagnostics

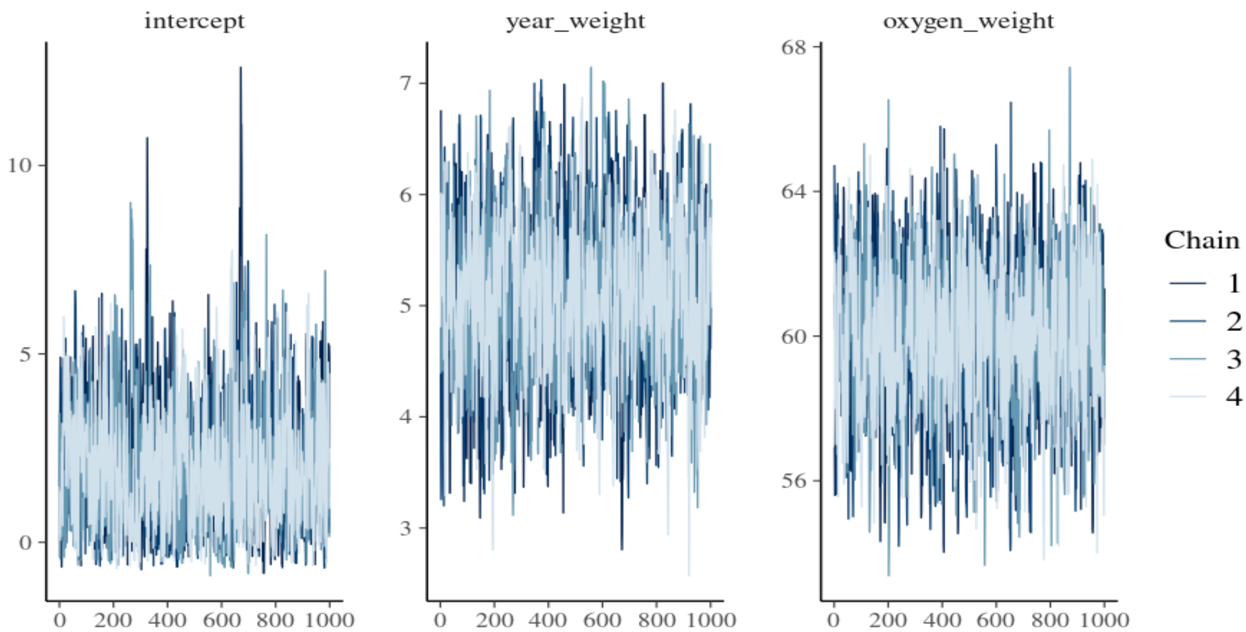
6.1 Convergence of the chains

We used a stan function to compute the MCMC. The following plots show the Monte Carlo chains for the two models we have used.

i. Linear model:



ii. Hierarchical model:



Checking the convergence of the chains is important because we have to somehow decide if our chains have managed to approximate the target distribution in a finite number of steps. In the plots above, the fact that the chain overlay and cover the same range is a good indicator for that. Nevertheless, it is sometimes hard to decide if the chains have converged just by looking at the plots. Therefore, we need numerical methods that can decide convergence. The summary method tells us that the Rhat values, which need to be below a threshold of 1.01 for convergence. Since Rhat should theoretically be above 1 (in practice it can go slightly under 1, due to computation inaccuracies), the good interval is [1.00, 1.01].
The Estimation of Effective Sample Size (ESS) is a metric of how many of the obtained samples are indeed effective in matching the desired posterior distribution.

We have the following results for our models:

i. Linear model:

- for the intercept: Rhat = **0.99,** ESS = **1714**
- for the year weight: Rhat = **1.00,** ESS = **1786**
- for the oxygen weight: Rhat = **1.00,** ESS = **1902**

ii. Hierarchical model:

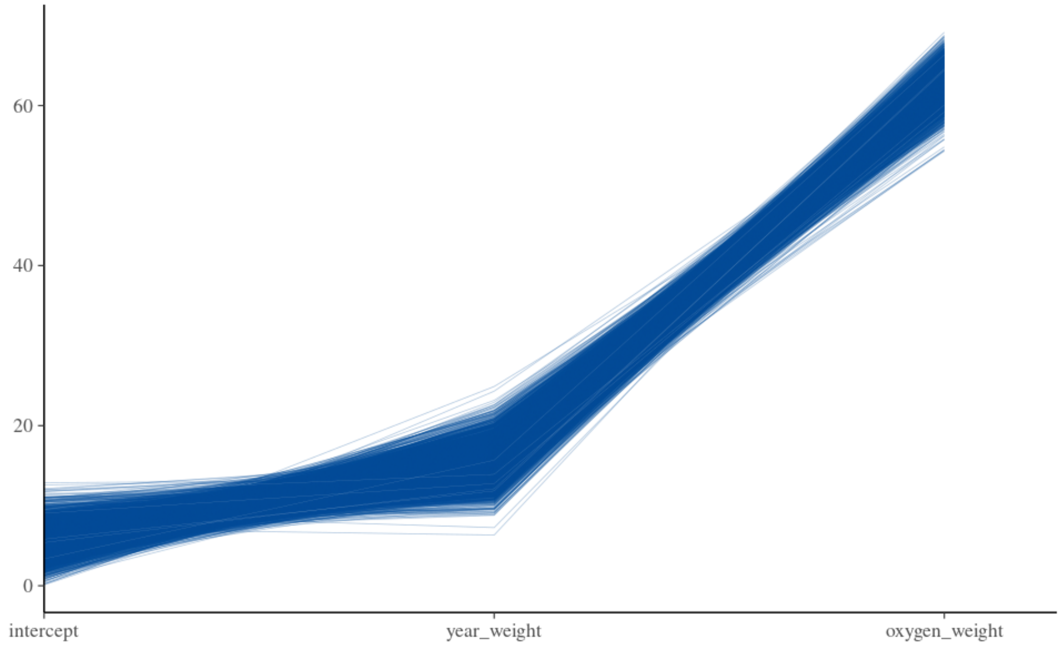
- for the intercept (with oxygen): Rhat = **0.99,** ESS = **2966**
- for the year weight (with oxygen): Rhat = **1.00,** ESS =**2749**
- for the intercept (without oxygen): Rhat = **0.99,** ESS = **2854**
- for the year weight (without oxygen): Rhat = **1.00,** ESS = **2796**
- hyper_intercept_mean: Rhat = **0.99,** ESS = **3860**
- hyper_intercept_sd: Rhat = **1.00,** ESS = **3428**
- hyper_year_weight_mean: Rhat = **0.99,** ESS = **4936**
- hyper_year_weight_sd: Rhat = **0.99,** ESS = **4110**

As it can be seen all our Rhat values are below the threshold of 1.01 and therefore we can conclude that both our models converged.

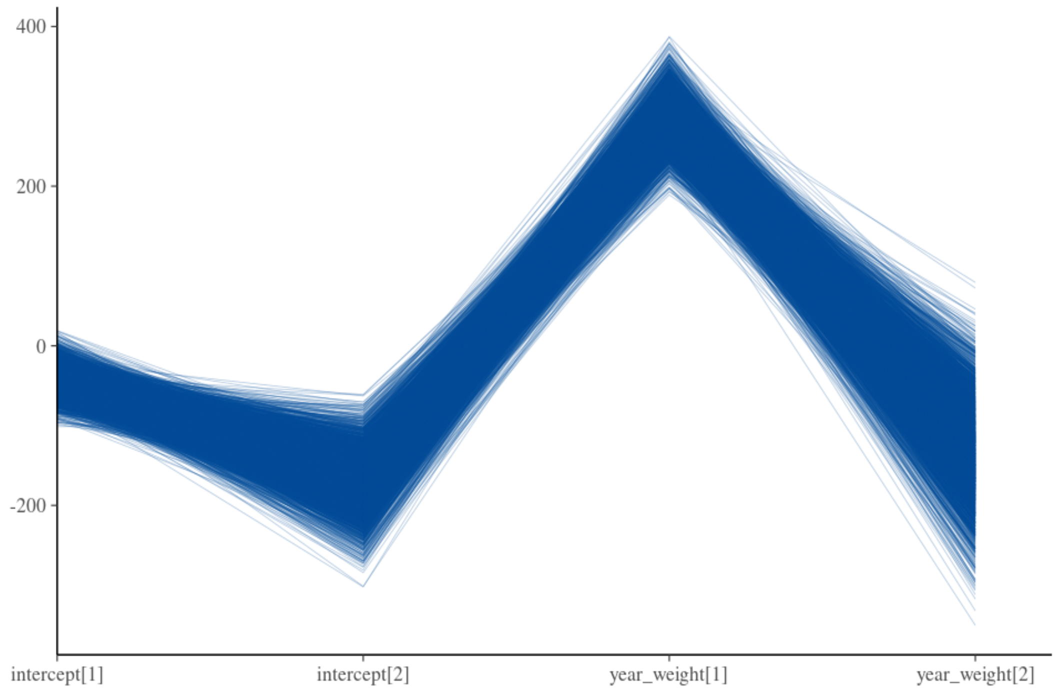
Similarly, it can be seen from our ESS values that for the *Linear model* the effective sample size for the parameters is in the range 1700-1900. That means that the results estimated through HMC are equivalent to using around 1700 to 1900 samples drawn from the real posterior.
For the *Hierarchical model*, the ESS for the parameters are around 2800 (and for the hyperparameters even higher). Thus in this case the estimates are equivalent to using that number of samples from the real posterior distribution. This means that the estimates for this model are more reliable and also that the HMC algorithm works better in approximating the posterior from this model.

6.2 Divergence assessment

i. Linear model:



ii. Hierarchical model:



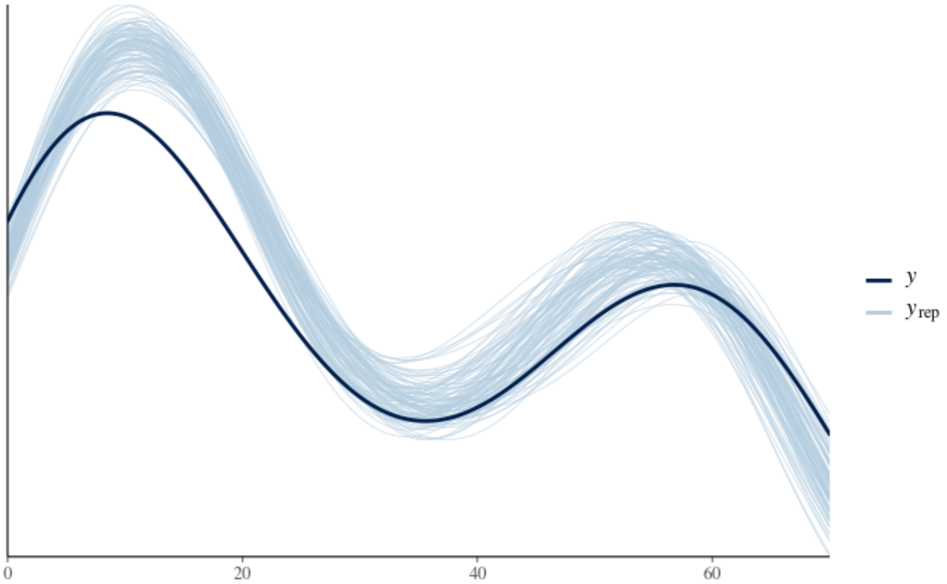
The two previous plots show the parallel values for each parameter in the model, where a line represents a draw. This is useful as a HMC diagnostic, since it is easy to spot the divergences: we have highlighted them in white. We can see that there are very few divergent values for the Linear model, and that the Hierarchical model has no divergences.

7. Posterior predictive checks

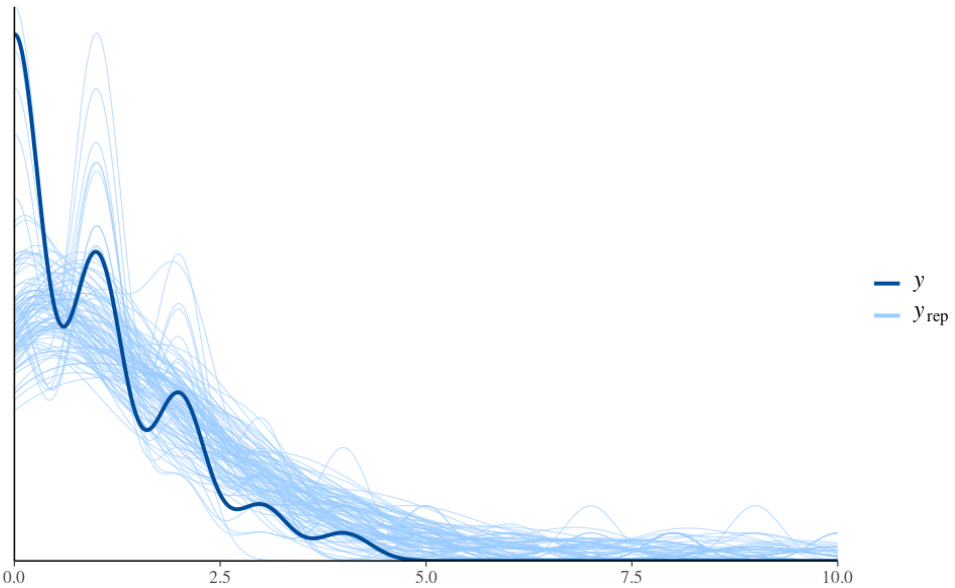
We can see that the posterior predictive distribution looks plausible according to our observed data. We have multiple plots to showcase this fact for the two models we use.

For the posterior predictive checks we plotted the distribution of the observed data y (with a dark line) and 100 lighter blue lines of the replications of the data y_rep sampled from the posterior predictive distribution. There are two plots for each model, depending if the oxygen was used or not.

i. Linear model:

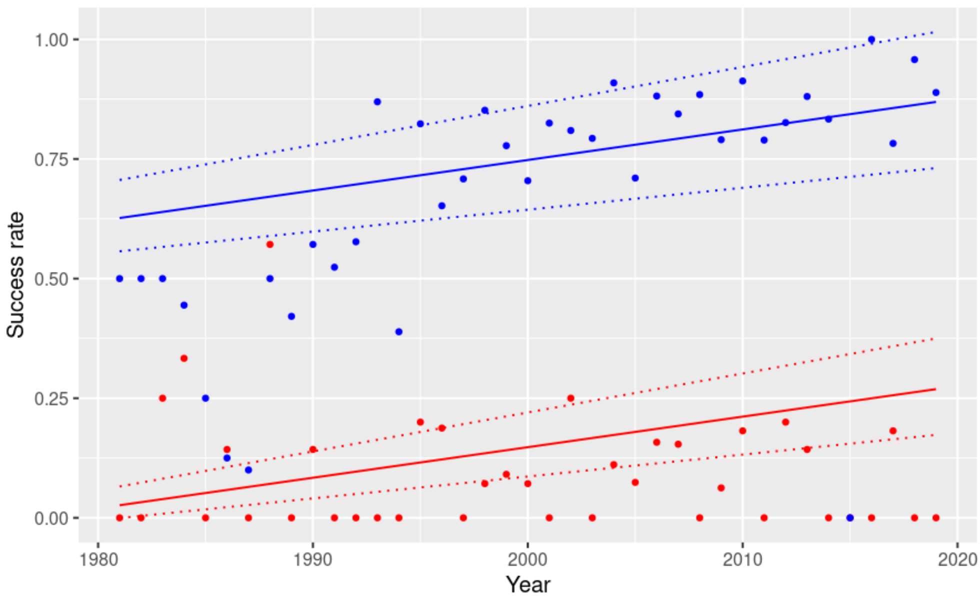


The above plot is the predictive check for the Linear model when oxygen was used. The plot makes clear the fact that the model overestimates the values of y since for most of the interval, all the lighter blue lines are above the dark blue. This trend is different at the two ends, where the model actually predicts fewer observations for those values.

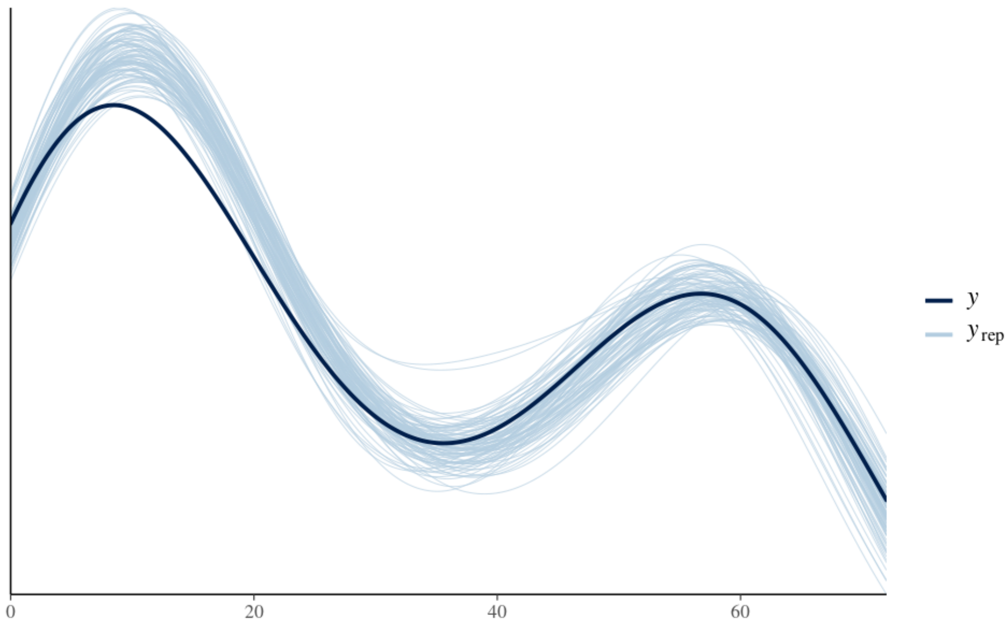


The above plot is the predictive check for the Linear model when oxygen was not used. This model predicts a lot fewer values for 0 than what we have observed (the dark blue line is much higher at 0 than the lighter blue lines).

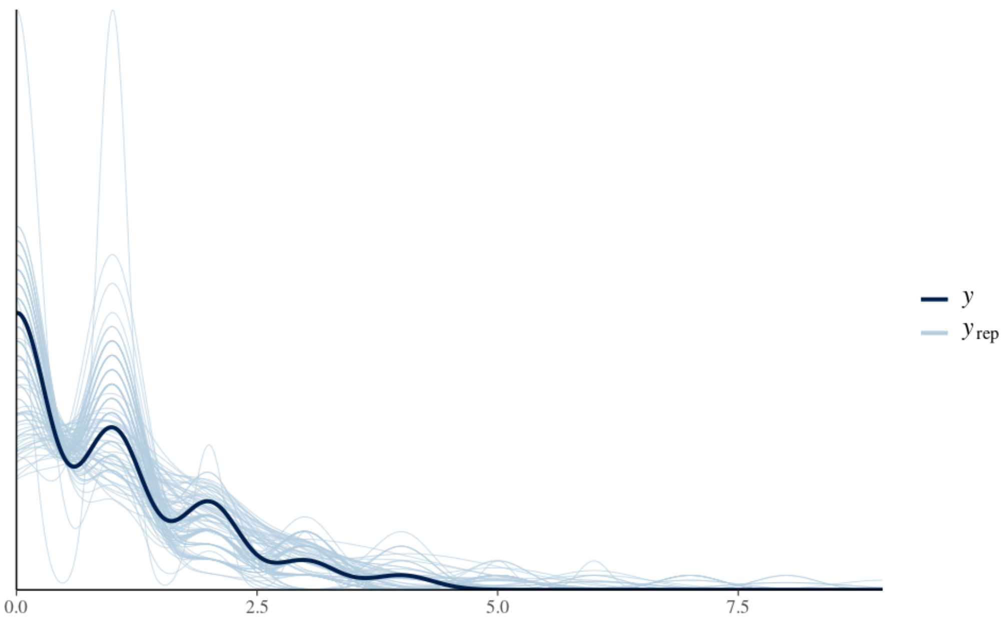
The following plot shows the fit of the model on the observed data. The blue and red lines are the estimations given by the posterior for two classes of our data: success rate of expeditions using supplemental oxygen (here the blue dots) and success rate of expeditions that do not use supplemental oxygen (here the red dots). The dashed lines are the corresponding 95% confidence intervals. We can observe the fit is weaker, missing many of the observations in our data.



ii. Hierarchical model:

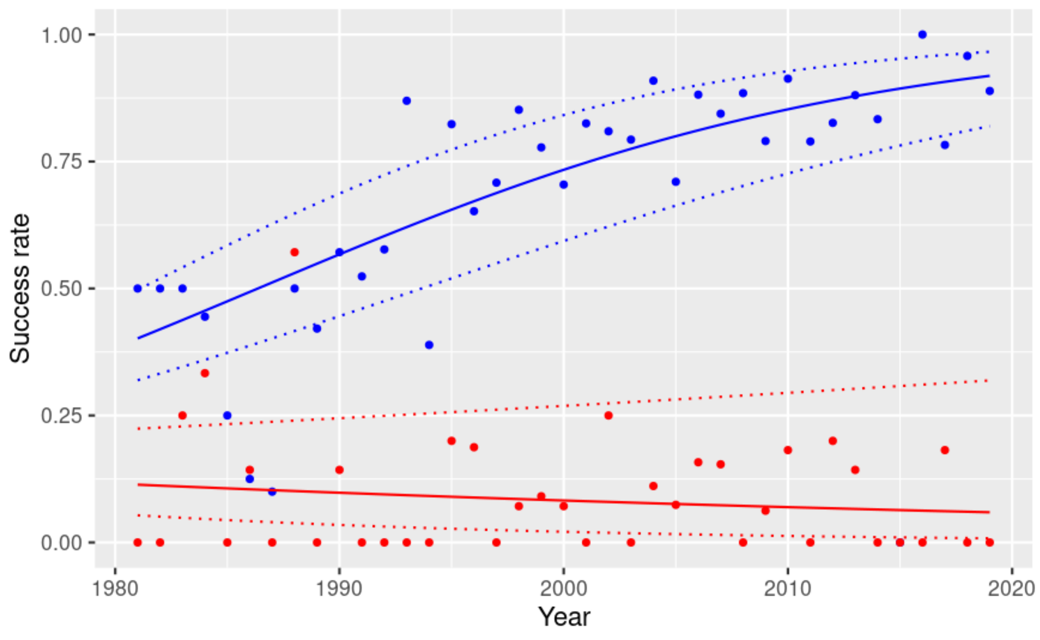


The above plot is the predictive check for the Hierarchical model when oxygen was used. It is straightforward to see that in general the model correctly accounts for all the values of y (the dark line is in the middle of lighter blue lines for the majority of the interval) with just a small overestimation for the interval [10, 20].



The above plot is the predictive check for the Hierarchical model when oxygen was not used. In general the model predicts correctly all values of y since the dark line is in between lighter lines representing *y_reps*.

The following plot shows the fit of the model on the observed data. The blue and red lines are the estimations given by the posterior for two classes of our data: success rate of expeditions using supplemental oxygen (here the blue dots) and success rate of expeditions that do not use supplemental oxygen (here the red dots). We can observe that the fit is sensible compared to the original distribution of the observed data.



8. Prior sensitivity analysis

For the prior sensitivity analysis part, we evaluated our model using different weakly informative and informative priors.

Some of the alternative priors used for the linear model:

- intercept ~ N(50, 10)
- year_weight ~ N(10, 5)
- oxygen_weight ~ N(20, 7)

Some of the alternative prior used for Hierarchical model:

- hyper mean of the intercept ~ N(0, 100)
- hyper standard deviation of the intercept ~ N(100, 30)
- hyper mean of the year weight ~ N(0, 100)
- hyper standard deviation of the year weight ~ N(100, 30)

Our results show that although the choice of prior has some impact on the model, the predictive posterior distributions that result after running the MCMC chains are very similar. This was also confirmed by the cross validation since the *elpd* values obtained with different priors are also comparable (with just very little improvement when the priors become more informative). The changes for the *elpd* from using alternative priors on the Linear model was up to 5% and around 1% for the Hierarchical model. Therefore we can state that the choice of priors *does not heavily affect* the stability so our model is robust and not sensitive to prior changes.

The facts stated here were made based both on visual comparison of the posterior distribution and analysing the change in *elpd* value.

9. Model comparison (LOO-CV)

9.1 ELPD values

ELPD is the theoretical expected log pointwise predictive density for a new dataset, which can be estimated using leave-one-out cross-validation. The values obtained for elpd-loo for the two models we have described are the following:

i. Linear model:

	Estimate <S3: AsIs>	SE <S3: AsIs>
elpd_loo	-205.9	22.2
p_loo	8.5	2.4
looic	411.8	44.3

ii. Hierarchical model:

	Estimate <S3: AsIs>	SE <S3: AsIs>
elpd_loo	-161.0	13.2
p_loo	7.9	1.7
looic	322.1	26.3

3 rows

The best model should be the one with the highest **elpd-loo**. Here, the best one is the Hierarchical Logistic model. However, the Linear model has an elpd-loo and Monte Carlo Standard error close in value to the Hierarchical one, which means both models perform reasonably well. We computed the LOO-CV function using the *loo()* function.

9.2 *Pareto-k diagnostic*

The Pareto k diagnostic estimates how far an individual leave-one-out distribution is from the full distribution. If leaving out an observation changes the posterior too much then importance sampling is not able to give a reliable estimate.

The results are the following:

i. Linear model:

```
Computed from 4000 by 78 log-likelihood matrix

------
Monte Carlo SE of elpd_loo is 0.1.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     77    98.7%   636
 (0.5, 0.7]   (ok)        1     1.3%   215
   (0.7, 1]   (bad)       0     0.0%   <NA>
   (1, Inf)   (very bad)  0     0.0%   <NA>

All Pareto k estimates are ok (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

ii. Hierarchical model:

```
Computed from 4000 by 78 log-likelihood matrix

------
Monte Carlo SE of elpd_loo is 0.2.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.
```

The Hierarchical model has all the k-estimates below 0.5, which is a sign that such a model is reliable. The next best model is the linear one which has only 1 k-estimate between 0.5 and 0.7.

The previous posterior plots from section 7 visually confirm what we have discovered here numerically: the hierarchical model is better than the linear one in estimating the success rate and predicting observations that match. This can be explained by the fact that the oxygen use doesn't influence the success rate by a constant factor on the entire time range.

10. Discussion of issues and potential improvements

As mentioned above, the first model created was a linear model. It was considered that as a preliminary step to creating a more complex model, it would be interesting to see how a simpler model would perform. However, there was an issue with the constraints of the parameter we were trying to estimate. Because the values of this variable (success rate) are in an interval [0,1], we need to ensure that through a tight regularisation of the priors. Furthermore, this caused a small number of divergences and made the model more prior sensitive. Therefore, for the next model, we decided to use a sigmoid function whose values fall within this interval and allows for more freedom in the prior choice and robustness.

Apart from this problem, which has been solved with the sigmoid function, there are other possible improvements that could be considered for future work. On the one hand, our main objective has been to analyse the cumulative success rate depending on the number of attempts, the number of that used supplemental oxygen and the number of non-supplemental oxygen used. Apart from these variables, other ones could be used to give the model more information and see if there is any improvement in this way. As an example, the variable "hired_staff" could be added, which can be of great help to the climbers.

Moreover, annual data have been used in each of the models developed. Perhaps another possible improvement could be to change the granularity of the time to monthly data. In this way, we could check if the model, with more information than before, generalises better and, in this way, obtains more accurate results. This has the potential to be a significant improvement since the seasonality has a big impact on the success of an expedition. However, if it is decided to do this, the form of the function will be completely different and therefore the priors will have to be changed.

11. Conclusions what was learned from the data analysis

The main conclusion that we can draw from this project in the process of analysing the data available to us, is the fact that oxygen is essential to reach the Everest summit. As we can see in the plot shown in section 7 (posterior predictive checks for hierarchical model), there is a notable difference between the success rate achieved by using and not using oxygen. Interestingly enough, the gap between the success rates has been widening over the years.

On the one hand, due to improvements in both oxygen masks and equipment used by the climbers in general, the success rate has improved considerably, reaching almost 100%.

On the other hand, over the years we have analysed, the success rate decreases for the expeditions which are not using oxygen. Moreover, if we look at the confidence intervals, the uncertainty increases over the years as well.

The decrease mentioned above was certainly an unexpected result for us. In doing some research into the reasons for this, we have found that over the last few years, the number of people climbing Everest has increased considerably, to the extent that there are queues at the peak. Due to this, people who don't use oxygen masks may die because the waiting time to reach the summit and turn at a safe altitude increases.

12. Self-reflection of what the group learned while making the project

This project allowed us to understand and learn several new concepts related to Bayesian Data Analysis which we hadn't worked with before. First of all, understanding the nature of the data has been essential in order to come up with different models that fit in our context. Later, in the elaboration of these models, one of the most important parts of the project was to try to obtain different priors that would adapt well to our data and, in this way, achieve accurate results.

Later on, this group has been able to learn new methodologies when creating new models, with special emphasis on the hierarchical model. Apart from devising priors that adapt to our data, in addition to the models created, one of the things that this group has internalised the most during the project has been the fact of carrying out of a diagnosis of the convergence of each of these models, making use of statistics that we were unaware of until now.

The most challenging part was to make sure that the good results the models get using various convergence diagnostics are not falsely correct. As a complement to the convergence diagnosis, it has also been very important to understand that different priors must be tested in order to check whether the results obtained are stable.