BDA - Project

Anonymous

Analysis of survival chance relating to size of the maligant melanoma tumor size

Introduction

This project analyses the chance to survive maligant melanoma with the thickness of melanoma tumor size. The data is a dataset collected at the Department of Plastic Surgery, University Hospital of Odense, Denmark from 1962 to 1977. Each patient had their tumour removed with surgical operation, with around 2,5cm of the surrounding skin. The dataset contains the survival time of each patient after the surgery, information about the patient (survived/died, sex, age), year of the operation, thickness of the tumour and if there were ulcers present in the tumour. The patients were followed to the end of 1977.

Analysis

Read data.

```
csv <- read.csv("../data/Melanoma.csv", header=TRUE)
data_melanoma <- as.data.frame(csv)
```

Quick glimpse inside the dataset

Now to give you a rough understanding of what the data set looks like:

```
data_melanoma$ulcer <- as.logical(data_melanoma$ulcer)
data_melanoma$sex <- as.factor(data_melanoma$sex)

summary(data_melanoma)
```

```
##        X            time          status     sex          age       
##  Min.   :  1   Min.   :  10   Min.   :1.00   0:126   Min.   : 4.00  
##  1st Qu.: 52   1st Qu.:1525   1st Qu.:1.00   1: 79   1st Qu.:42.00  
##  Median :103   Median :2005   Median :2.00           Median :54.00  
##  Mean   :103   Mean   :2153   Mean   :1.79           Mean   :52.46  
##  3rd Qu.:154   3rd Qu.:3042   3rd Qu.:2.00           3rd Qu.:65.00  
##  Max.   :205   Max.   :5565   Max.   :3.00           Max.   :95.00  
##       year        thickness       ulcer       
##  Min.   :1962   Min.   : 0.10   Mode :logical  
##  1st Qu.:1968   1st Qu.: 0.97   FALSE:115      
##  Median :1970   Median : 1.94   TRUE :90       
##  Mean   :1970   Mean   : 2.92                  
##  3rd Qu.:1972   3rd Qu.: 3.56                  
##  Max.   :1977   Max.   :17.42
```

```r
head(data_melanoma)
```

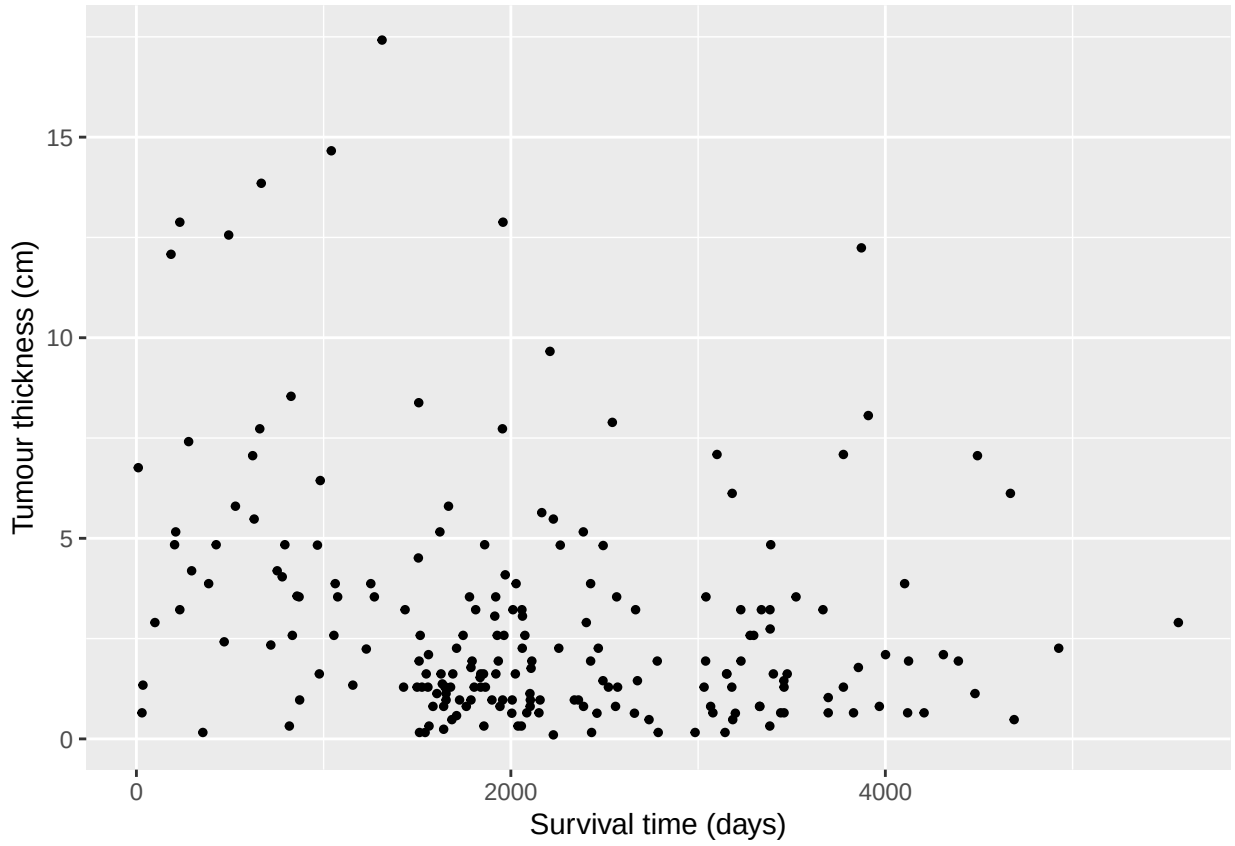
```
##   X time status sex age year thickness ulcer
## 1 1   10      3   1  76 1972      6.76  TRUE
## 2 2   30      3   1  56 1968      0.65 FALSE
## 3 3   35      2   1  41 1977      1.34 FALSE
## 4 4   99      3   0  71 1968      2.90 FALSE
## 5 5  185      1   1  52 1965     12.08  TRUE
## 6 6  204      1   1  28 1971      4.84  TRUE
```

Scatterplot

Plot scatterplot with survival time and tumour thickness.

```r
input <- list(x = data_melanoma$time,
              y = data_melanoma$thickness)

ggplot() +
  geom_point(aes(x, y), data = data.frame(input), size = 1) +
  labs(y = 'Tumour thickness (cm)', x = 'Survival time (days)') +
  guides(linetype = F)
```



Linear model

Data

```
writeLines(readLines("linear.stan"))
```

```
##
## data {
##   int<lower=0> N; // number of data points
##   vector[N] x; // survival time
##   vector[N] y; // size of the tumour
##   real xpred;  // prediction
## }
##
## parameters {
##   real alpha;
##   real beta;
##   real<lower=0> sigma;
## }
##
## model {
##   y ~ normal(alpha + beta * x, sigma);
## }
##
## generated quantities {
##   real ypred;
##   ypred = normal_rng(alpha + beta * xpred, sigma);
## }
```

```
input_linear <- list(N = nrow(data_melanoma),
                     x = data_melanoma$time,
                     y = data_melanoma$thickness,
                     xpred = 2000)
fit_linear <- stan(file='linear.stan', data=input_linear, seed=SEED)
print(fit_linear)
```

```
## Inference for Stan model: linear.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff
## alpha    4.24    0.01 0.44    3.42    3.94    4.25    4.55    5.12  1567
## beta     0.00    0.00 0.00    0.00    0.00    0.00    0.00    0.00  1855
## sigma    2.90    0.00 0.14    2.65    2.80    2.90    2.99    3.20  1540
## ypred    3.02    0.05 2.88   -2.68    1.11    2.97    4.91    8.76  4009
## lp__  -319.02    0.03 1.23 -322.07 -319.62 -318.70 -318.12 -317.65  1252
##       Rhat
## alpha    1
## beta     1
## sigma    1
## ypred    1
## lp__     1
```

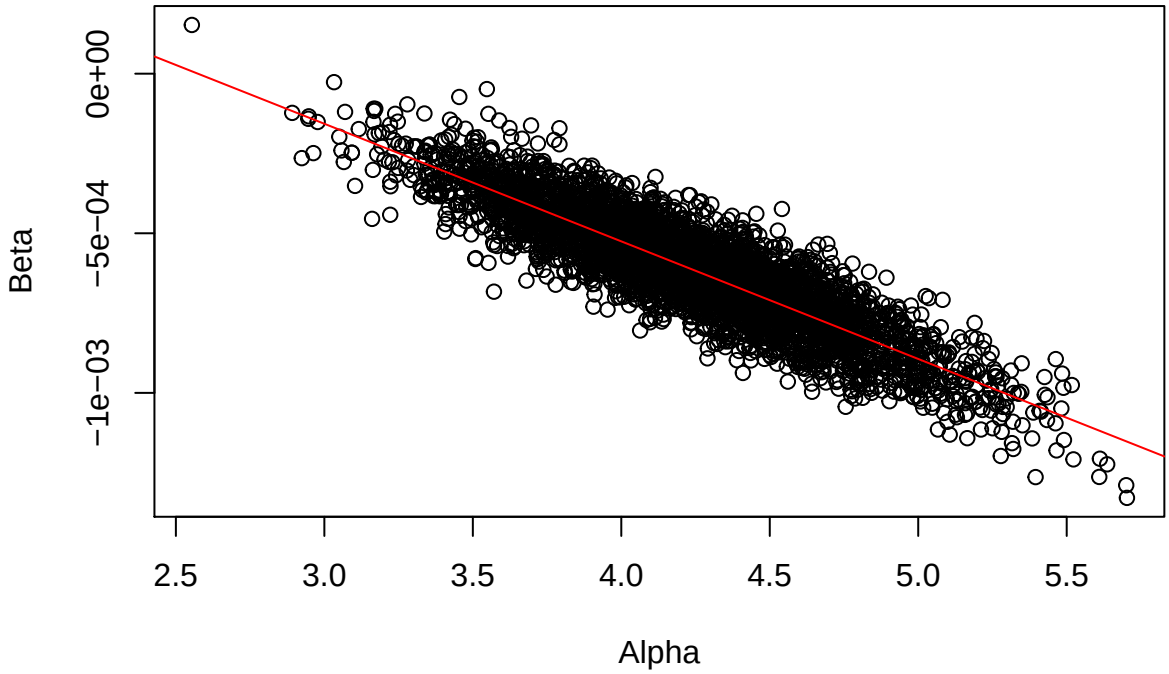
```
##
## Samples were drawn using NUTS(diag_e) at Sun Dec 08 23:46:10 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

```
data_extract <- as.data.frame(fit_linear)
```

Plot

Plotting to see if linear correlation can be found:

```
plot(data_extract$alpha, data_extract$beta, xlab = "Alpha", ylab = "Beta")
abline(lm(data_extract$beta ~ data_extract$alpha), col = "red")
```

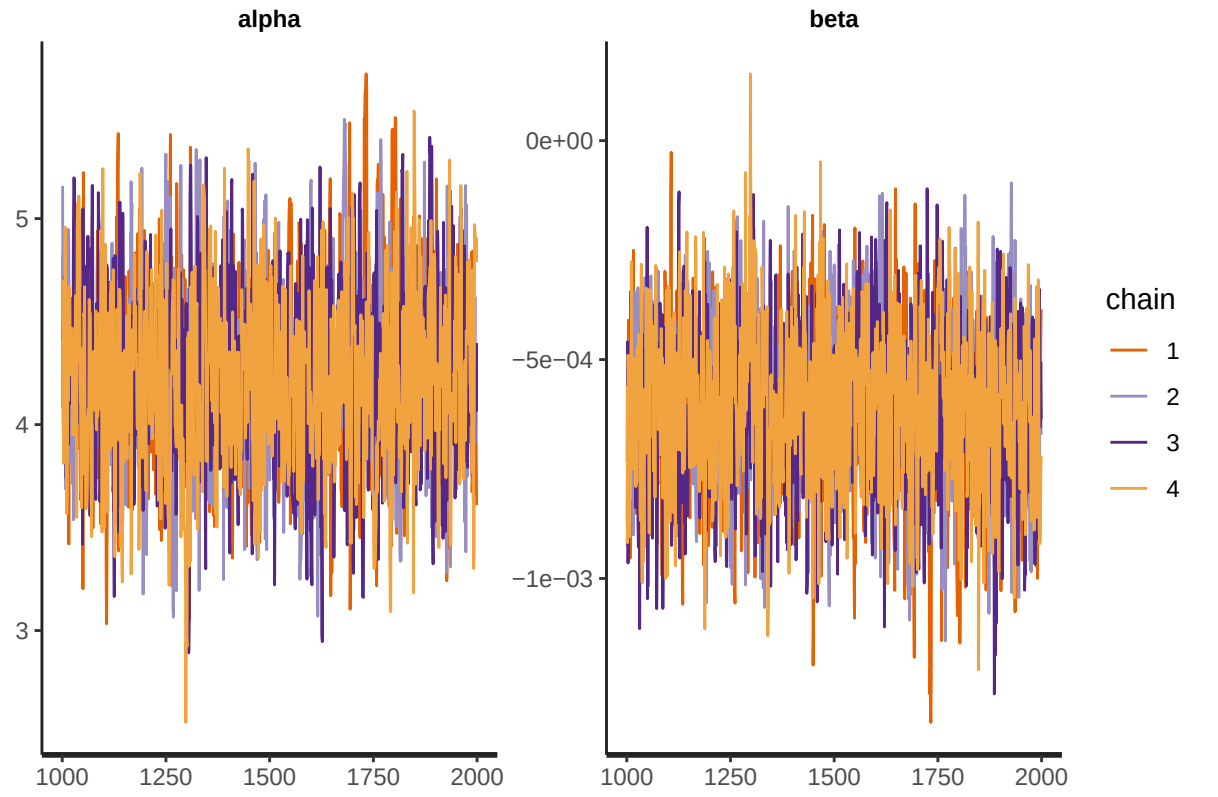


In the plot Alpha represents the time survived after the melanoma operation and Beta the thickness of the ulcer. The values seem to have a clear linear correlation in the way that those with smaller ulcers have lived longer after the operation.

Chain converging

Check convergence of chains:

```
rstan::traceplot(fit_linear, pars=c('alpha','beta'))
```



Although the scales are rather differing, the forms of the chains do behave a lot like one another.

HMC diagnostics

HMC diagnostics:

```
check_hmc_diagnostics(fit_linear)
```

```
##
## Divergences:

## 0 of 4000 iterations ended with a divergence.

##
## Tree depth:

## 0 of 4000 iterations saturated the maximum tree depth of 10.

##
## Energy:

## E-BFMI indicated no pathological behavior.
```

None of the HMC-test output values give warnings, which can be considered a positive trait.

Rhat

Rhat diagnostics:

```
rhat(fit_linear)
```

```
##     alpha      beta     sigma     ypred      lp__
## 1.0033176 1.0029187 1.0009496 0.9994646 1.0026805
```

The alpha and beta values are remarcably close to one another, meaning the chains are well converged.

ESS -values

From monitor, the effective sample size ESS can be seen:

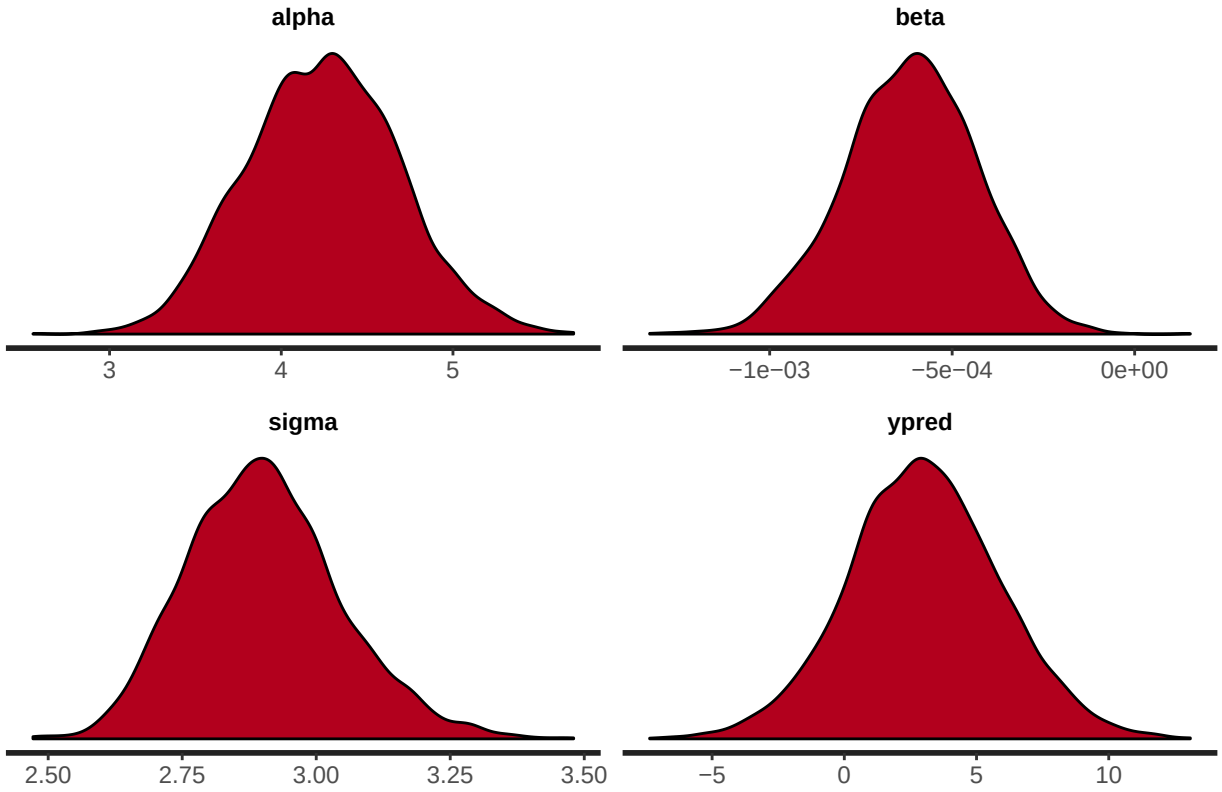
```
monitor(fit_linear)
```

```
## Inference for the input samples (4 chains: each with iter = 2000; warmup = 0):
##
##           Q5    Q50    Q95   Mean  SD  Rhat Bulk_ESS Tail_ESS
## alpha    3.5    4.2    5.0    4.2 0.4     1     1580     1314
## beta     0.0    0.0    0.0    0.0 0.0     1     1886     1917
## sigma    2.7    2.9    3.2    2.9 0.1     1     1588     1258
## ypred   -1.6    3.0    7.9    3.0 2.9     1     4032     3631
## lp__  -321.3 -318.7 -317.7 -319.0 1.2     1     1424     1644
##
## For each parameter, Bulk_ESS and Tail_ESS are crude measures of
## effective sample size for bulk and tail quantities respectively (an ESS > 100
## per chain is considered good), and Rhat is the potential scale reduction
## factor on rank normalized split chains (at convergence, Rhat <= 1.05).
```

Posterior densities

Plot posterior densities:

```
stan_dens(fit_linear)
```



Posterior predictive checks

In the figure above is plotted density of ypred where is the distribution of possible tumour sizes when the predicted survival time is 2000 days. The expected tumour size is about three centimeters, which is very close to the mean of tumour sizes.

Bernoulli model

Purpose of this model is to get information and predict the survival chance from melanoma, when the tumour thickness is known.

Clean data, 1 represents individuals that died because of melanoma, 0 represents individuals that are alive or died from other causes.

```
data_melanoma[data_melanoma$status == 2,]$status <- 0
data_melanoma[data_melanoma$status == 3,]$status <- 0
```

Weakly informative priors were chosen for alpha and beta, both are normal(0, 10). The selected prior represents the thickness pretty well.

```
writeLines(readLines("bernoulli.stan"))
```

```
##
## data {
```

```
##   int<lower=0> N;
##   vector[N] x;               //thickness
##   int<lower=0,upper=1> y[N]; //status
##   real xpred;                //thickness prediction
## }
##
## parameters {
##   real alpha;
##   real beta;
## }
##
## model {
##   alpha ~ normal(0, 10);
##   beta ~ normal(0, 10);
##   for (n in 1:N)
##     y[n] ~ bernoulli_logit(alpha + beta * x[n]);
## }
##
## generated quantities {
##   real ypred;
##   ypred = bernoulli_logit_rng(alpha + beta*xpred);
## }
```

Create input list and run the stan model.

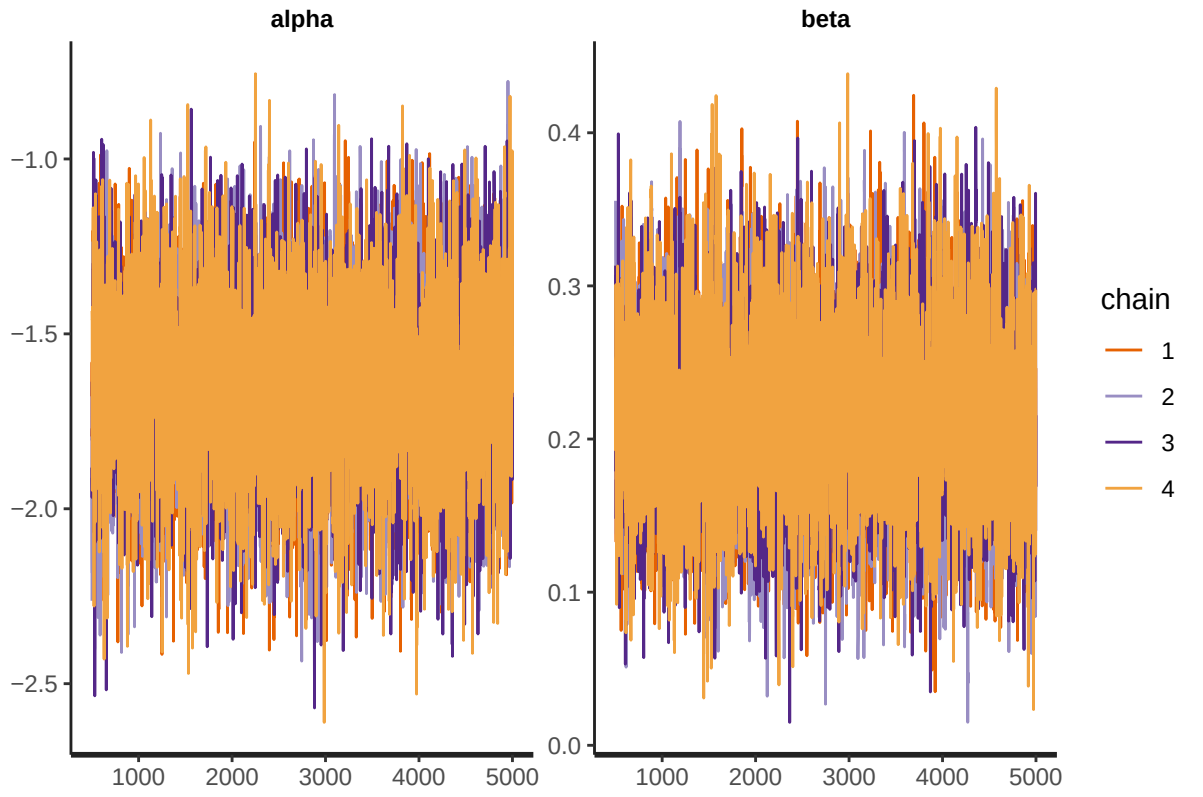
```
input_bernoulli <- list(N = nrow(data_melanoma),
                        x = data_melanoma$thickness,
                        y = data_melanoma$status,
                        xpred = 16)
fit_bernoulli <- stan(file='bernoulli.stan', data=input_bernoulli, seed=SEED, iter=5000, warmup=500)
print(fit_bernoulli)
```

```
## Inference for Stan model: bernoulli.
## 4 chains, each with iter=5000; warmup=500; thin=1;
## post-warmup draws per chain=4500, total post-warmup draws=18000.
##
##          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff
## alpha   -1.64    0.00 0.24   -2.12   -1.80   -1.63   -1.47   -1.17  5694
## beta     0.22    0.00 0.06    0.11    0.18    0.21    0.25    0.33  5695
## ypred    0.84    0.00 0.37    0.00    1.00    1.00    1.00    1.00 16133
## lp__  -114.08    0.01 1.00 -116.80 -114.47 -113.77 -113.36 -113.10  5781
##       Rhat
## alpha    1
## beta     1
## ypred    1
## lp__     1
##
## Samples were drawn using NUTS(diag_e) at Sun Dec 08 23:46:16 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Chain converging

Check converge of chains:

```
rstan::traceplot(fit_bernoulli, pars=c('alpha','beta'))
```



Chains seems to be pretty well converged.

HMC diagnostics

HMC diagnostics:

```
check_hmc_diagnostics(fit_bernoulli)
```

```
##
## Divergences:

## 0 of 18000 iterations ended with a divergence.

##
## Tree depth:

## 0 of 18000 iterations saturated the maximum tree depth of 10.
```

```
##
## Energy:

## E-BFMI indicated no pathological behavior.
```

None of the values gives any warnings, so everything went well in the fitting.

Rhat

Rhat diagnostics:

```
rhat(fit_bernoulli)
```

```
##    alpha     beta    ypred     lp__
## 1.000377 1.000387 1.000093 1.001194
```

Rhats are very close to one, so the chains are well converged.

ESS -values

From monitor -function we can see the effective sample sizes (ESS -values).

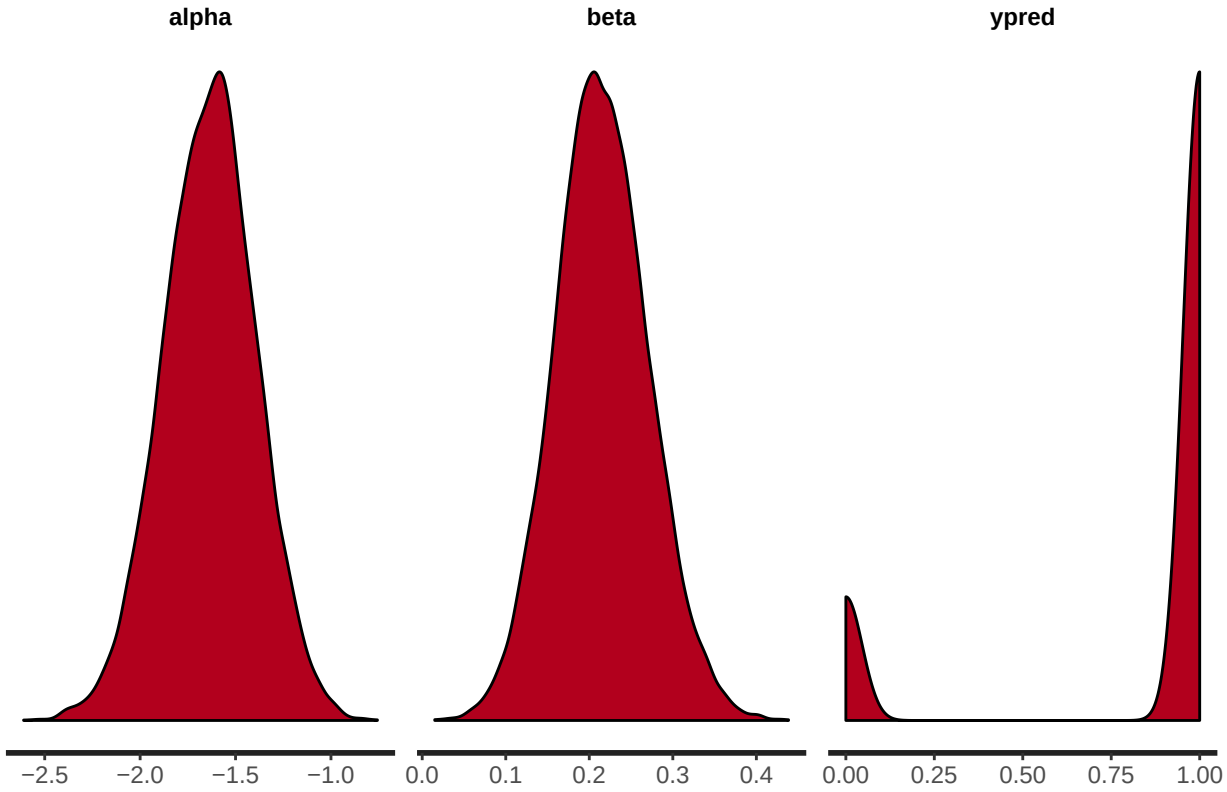
```
monitor(fit_bernoulli, warmup = 500)
```

```
## Inference for the input samples (4 chains: each with iter = 5000; warmup = 0):
##
##           Q5    Q50    Q95   Mean  SD  Rhat Bulk_ESS Tail_ESS
## alpha   -2.0   -1.6   -1.2   -1.6 0.2     1     5696     7882
## beta     0.1    0.2    0.3    0.2 0.1     1     5704     7410
## ypred    0.0    1.0    1.0    0.8 0.4     1    16171    16171
## lp__  -116.1 -113.8 -113.1 -114.1 1.0     1     6504     8084
##
## For each parameter, Bulk_ESS and Tail_ESS are crude measures of
## effective sample size for bulk and tail quantities respectively (an ESS > 100
## per chain is considered good), and Rhat is the potential scale reduction
## factor on rank normalized split chains (at convergence, Rhat <= 1.05).
```

Posterior densities

Plot posterior densities:

```
stan_dens(fit_bernoulli)
```



As can be seen from the beta, the expected probability of survival from melanoma is approximately 0.8 (chance of death is 0.2). It is dependant from the thickness of the tumour.

Posterior predictive checks

In the figure above is plotted density of ypred (status, 0=alive, 1=dead), which was predicted with the xpred = 16 (tumour thickness). 16 in the top end of the tumour thickness. When we look at the ypred, it clearly shows that there is high chance of dying when the tumour thickness is that high, and low chance of survival. The x label represents the probability. From the prediction we can see that the model works as expected.

Comparison of models

We used two different statistical models to analyze the data. Another was linear regression and another was bernoulli model. In these two models we approached the dataset in a bit different way. In linear regression we used variables: survival time and tumour thickness to predict the survival time when the tumour thickness is known. In bernoulli model we used status and thickness to predict the probability of death when the tumour thickness is known. They both measure the same thing in the end, which is survivability, but they address the problem in a different way and using partly different variables.

Conclusion

According to the data it can be concluded that the size of a tumor even if it is removed is in linear correlation with the patient having a shorter life span after the surgery. It can also be concluded that the size of the melanoma tumor has a correlation with the patient dying of cancer related reasons later on in life.

Problems and potential improvements

The data set used was rather old having being made in the 60s and 70s. Also the data set was not that large making the results a bit less reliable. To improve this analysis we recommend taking more data points that correspond to the situation with current medical technology. Also the data should be gathered from a larger area than just one university hospital.

To further improve the accuracy, other variables such as ulcer and age could be taken into account in the models.

Also normal distribution is not the best choice for this kind of data. The data does not follow normal distribution very well as can be seen from the scatter plot in the beginning of this report.

In the linear regression posterior predictive checking the prediction is kind of "wrong way", because there is predicted the size of the tumour with the time. It should be other way around so that the survival time is predicted with the size of tumour.