Chapter 10

Missing data

10.1 Introduction

Strategies for dealing with missing data take up a lot of space in the literature. It is easy to see why: experiments fail, surveys are not returned, files get corrupted etc..

But does it matter if some data are missing? When we update our priors given the likelihood, we use the data that we have seen. The data that we are yet to see can be dealt with by utilising the posterior predictive distribution. However, what if the data we have not seen are not exchangeable with the data that we have seen?

There are three common classifications for missing data:

(1) Missing Completely at Random

The data that are missing has no describable pattern. From one data observation to the next, there is no way that we will know how likely a missing value is. This is a strong assumption and a personal judgement about how the missing data should be treated.

(2) Missing at Random

In this case, we can describe the process for the occurrence of missing through a stochastic mechanism. For instance, we may know that there is a 5% chance that a data point will not be recorded regardless of the value. We can model this, but, when we have independent data, it is not worth it.

The missing at random situation is related to what you saw in the first half of the course with hidden Markov models and what you saw in the last chapter on errors-in-variables. In both of those cases, we had unknown latent variables with a clear statistical model to explain their relationship to what we can see.

(3) Missing Not at Random

Here, there is some rule or decision that has been taken to exclude certain data. Two key examples are

1. we have no data for a subpopulation,

2. we have not been able to measure something because of threshold effect.

The former can be handled using hierarchical models and the latter gives us censored data.

Example 10.1.1

Imagine a situation where we are expecting ten observations of count data. We receive the following:

12, missing, 9, 12, 8, 10, missing, missing, 7, 12.

If we have the following model,

$$\begin{aligned} X_i|\lambda &\sim \text{Po}(\lambda), \quad i = 1, \ldots, 10, \\ \lambda &\sim \text{Gamma}(\alpha, \beta), \end{aligned}$$

then we can calculate the posterior whilst treating the missing values as unknown parameters:

$$\begin{aligned} \pi(\lambda, \underline{x}_{\text{miss}}|\underline{x}_{\text{obs}}) &\propto \pi(\lambda, \underline{x}_{\text{miss}}, \underline{x}_{\text{obs}}) \\ &\propto \pi(\underline{x}_{\text{obs}}|\lambda)\pi(\underline{x}_{\text{miss}}|\lambda)\pi(\lambda). \end{aligned}$$

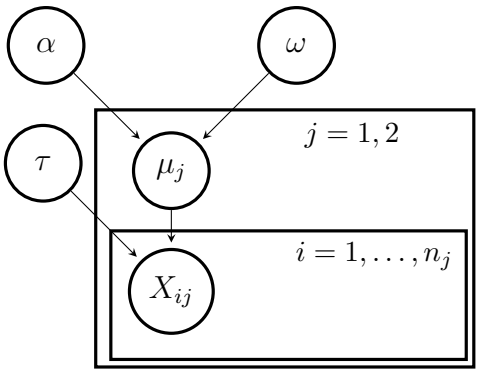
Now, we remove the nuisance parameters by summing over possible values of the missing variables:

$$\begin{aligned} \pi(\lambda|\underline{x}_{\text{obs}}) &\propto \pi(\underline{x}_{\text{obs}}|\lambda)\pi(\lambda) \sum_{x_2=0}^{\infty} \sum_{x_7=0}^{\infty} \sum_{x_8=0}^{\infty} \pi(\underline{x}_{\text{miss}}|\lambda) \\ &= \pi(\underline{x}_{\text{obs}}|\lambda)\pi(\lambda) \sum_{x_2=0}^{\infty} \pi(x_2|\lambda) \sum_{x_7=0}^{\infty} \pi(x_7|\lambda) \sum_{x_8=0}^{\infty} \pi(x_8|\lambda) \\ &= \pi(\underline{x}_{\text{obs}}|\lambda)\pi(\lambda) \end{aligned}$$

So considering the missing data at all was a waste of time.

Example 10.1.2

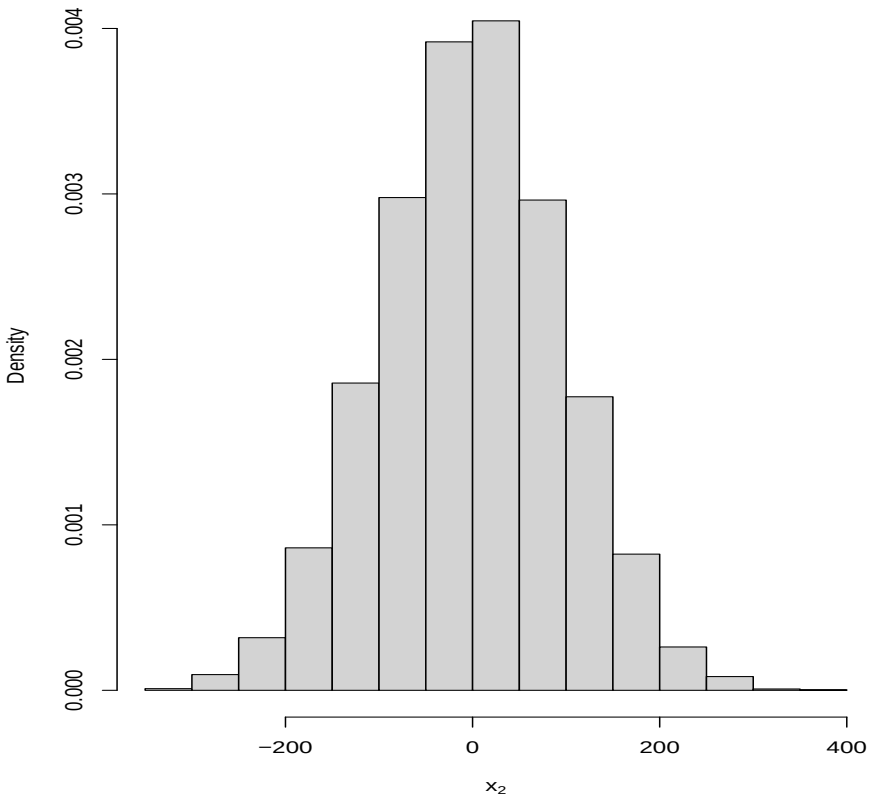
It might be unethical to experiment on a subpopulation. However, we have clear results for another subpopulation, and we strongly suspect that the results will carry over. This can be dealt with directly using hierarchical models.



We can imagine that we do not observe any X_{i2}, but we can use our model to look at the predictive distribution for this unseen population.

```
data {
  int<lower=0> N;
  real x1[N];
}
parameters {
  real mu[2];
  real alpha;
  real<lower=0> tau;
  real<lower=0> omega;
}
model {
  // Prior
  tau ~ gamma(10,1);
  alpha ~ normal(0,100);
  omega ~ gamma(10,1);
  mu ~ normal(alpha,sqrt(1/omega));

  // Likelihood
  for (i in 1:N){
    x1[i] ~ normal(mu[1],sqrt(1/tau));
  }
}
generated quantities {
  real x2;
  x2 = normal_rng(mu[2],sqrt(1/tau));
}
```

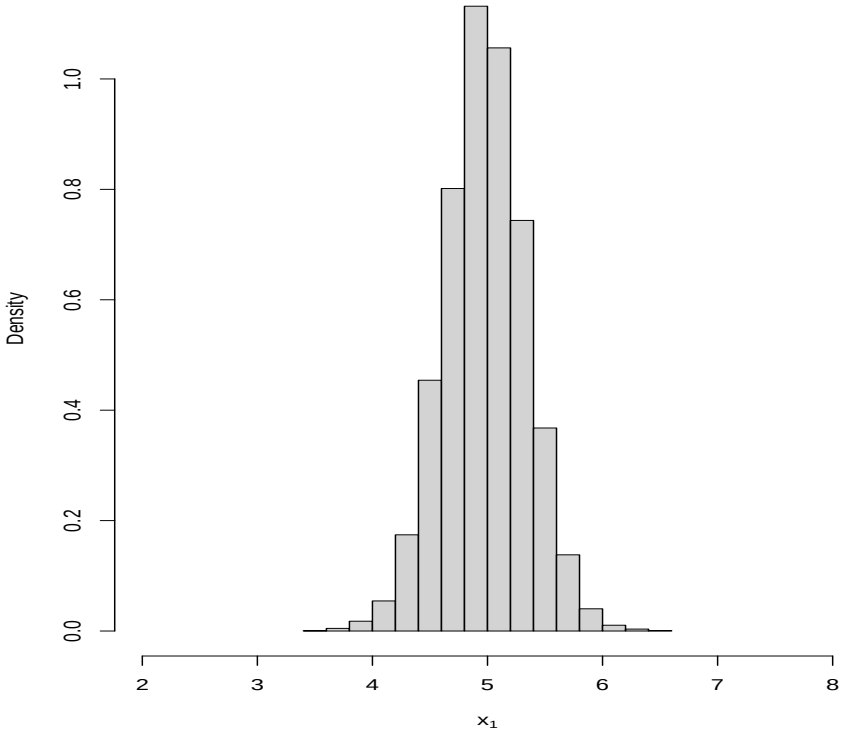
If we comment out the likelihood in the Stan model, we get a sample of X_{i2} from the preposterior distribution.

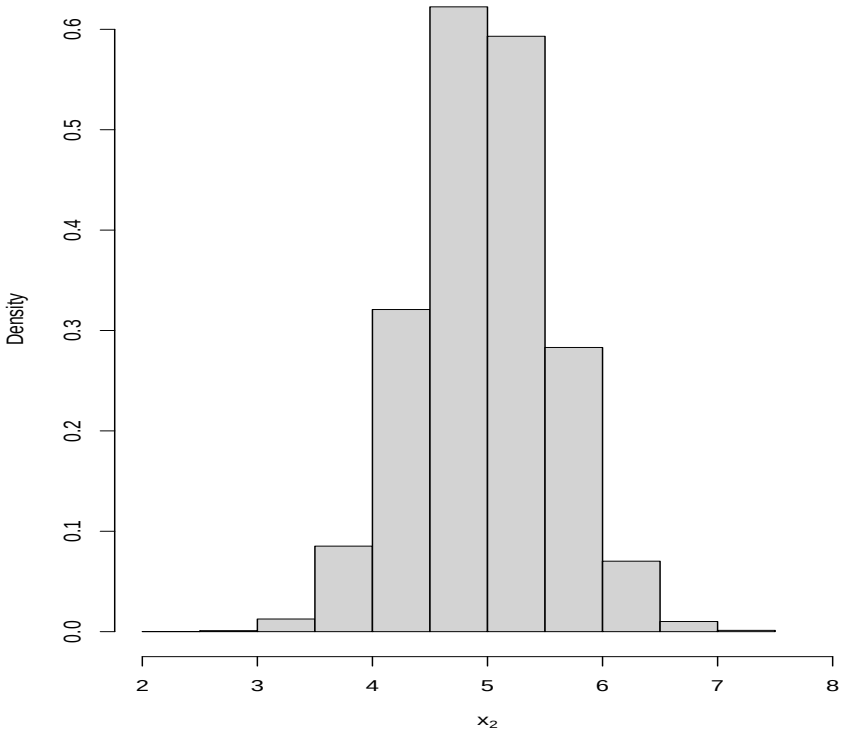


We use the following code to sample from the predictive distributions using the model:

```
fit <- sampling(model,
                data = list(N = 12,
                            x1 = c(5.3,5.1,4.8,4.5,
                                   5.5,5.2,5.0,5.0,
                                   5.1,4.6,4.3,5.3)),
                iter = 10000)
```

So we expect our predictive distribution to be centred in the vicinity of 5 because we have not explicitly modelled any expected bias between the two subpopulations.


Density
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
x1


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2
3
4
5
6
7
8
x2

10.2 Missing explanatory variables

A more interesting situation arises when we are modelling the relationship between response and explanatory variables. In particular, what if some of the explanatory measurements are missing?

Example 10.2.1

Let's consider simple linear regression:

$$\begin{aligned} y_i|\alpha,\beta,\sigma^2,x_i &\sim \mathsf{N}(\alpha+\beta x_i,\sigma^2),\\ \alpha|\sigma_\alpha^2 &\sim \mathsf{N}(0,10000),\\ \beta|\sigma_\beta^2 &\sim \mathsf{N}(0,10000),\\ \sigma^2 &\sim \mathsf{InvGamma}(0.01,0.01), \end{aligned}$$

We have the following information:

x	y
1.02	2.67
1.52	3.45
1.89	4.49
1.91	4.50
2.51	5.62
2.62	5.70
?	2.21
?	3.45

If we are going to treat the missing values as unknown, we need to put a prior distribution on them:

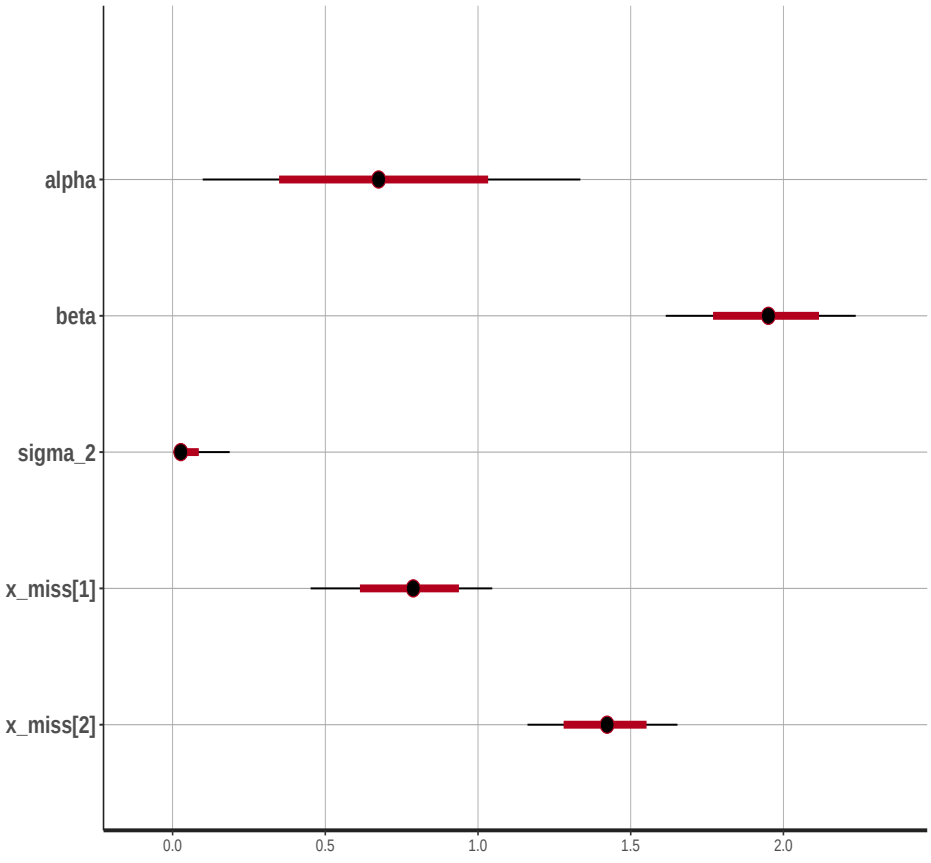
$$x_i \sim \mathsf{Uniform}(0,5), \quad i=7,8.$$

The first blocks in the Stan model become:

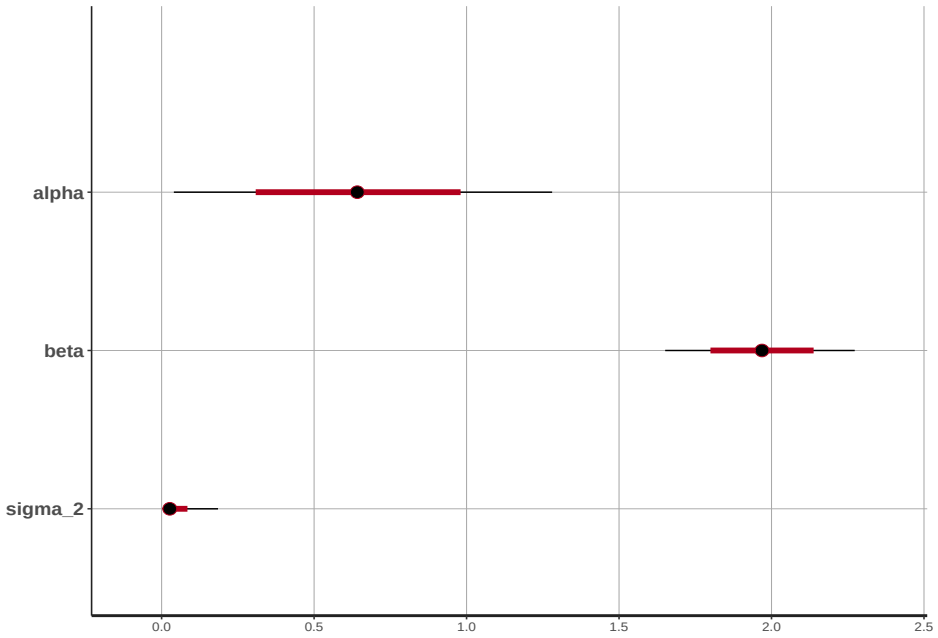
```
data {
  int<lower=0> N;   // num observations
  int<lower=0> M;   // num missing
  real x[N-M];      // observed explanatory variable
  real y[N];        // observed response variable
}
parameters {
  real alpha;                       // intercept
  real beta;                        // gradient
  real<lower=0> sigma_2;            // error variance
  real<lower=0,upper=5> x_miss[2];  // missing explanatory variables
}
```

And our model becomes:

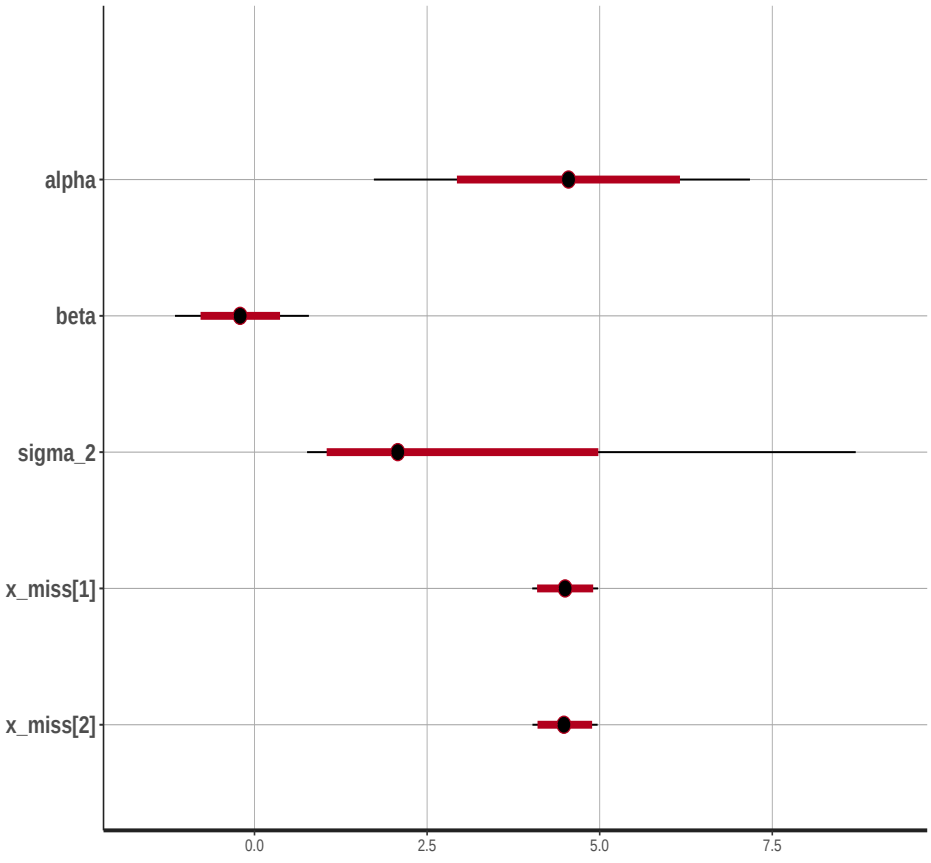
```
model {
  // Prior
  alpha ~ normal(0,10000);
  beta ~ normal(0,10000);
  sigma_2 ~ inv_gamma(0.01,0.01);
  x_miss ~ uniform(0,5);

  // Likelihood
  for (n in 1:(N-M))
      y[n] ~ normal(alpha + beta*x[n], sqrt(sigma_2));
  for (n in 1:M)
      y[N-M+n] ~ normal(alpha + beta*x_miss[n], sqrt(sigma_2));
}
```



Was there any point in including the missing data? Here are the marginal posteriors ignoring the data rows with missing x.



But what if we know the missing x should have been in [4,5]?



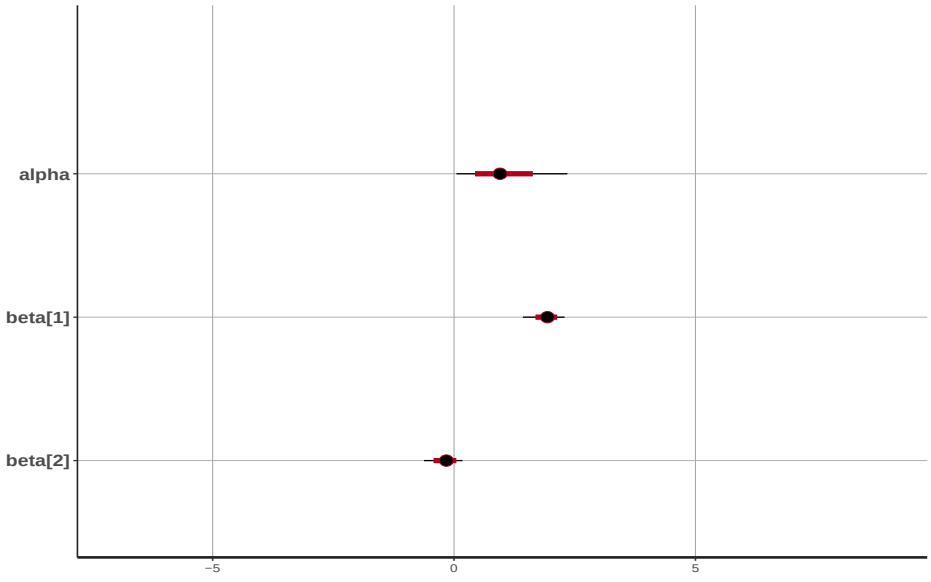
Example 10.2.2

If we start ignoring missing data with the following data, we will lose half of the observations.

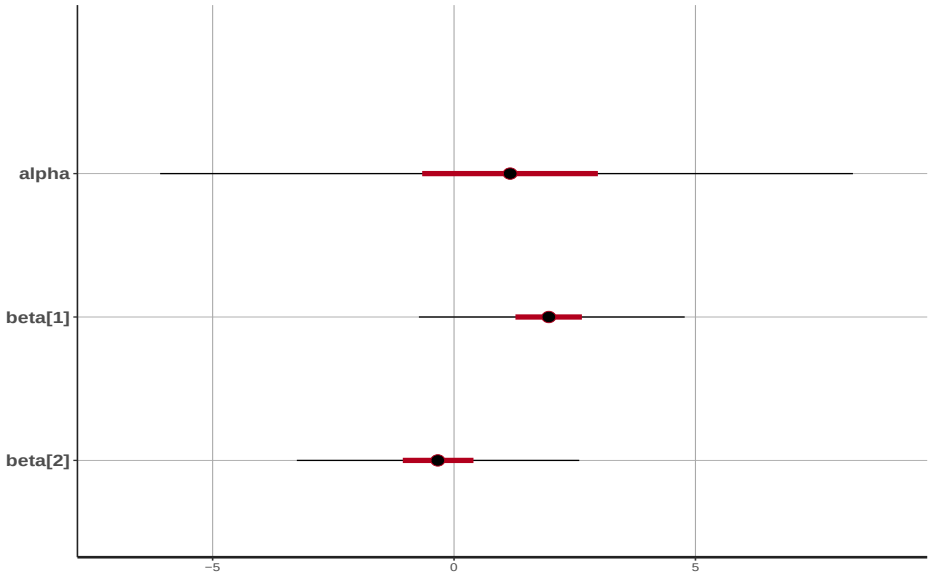
x_1	x_2	y
1.02	?	2.67
1.52	2.01	3.45
1.89	1.02	4.49
1.91	1.35	4.50
2.51	?	5.62
2.62	1.75	5.70
?	4.25	2.21
?	4.00	3.45

```
data {
  int<lower=0> N;           // num observations
  int<lower=0, upper=2> m;  // missingness indicator
  real x[N,2];  // observed explanatory variable (-9999 for missing)
  real y[N];    // observed response variable
}
parameters {
  real alpha;                          // intercept
  real beta[2];                        // gradient
  real<lower=0> sigma_2;               // error variance
  real<lower=0,upper=5> x_full[N,2];  // complete data
}
transformed parameters {
  real mu[N];
  for (n in 1:N) {
    if (m[n] = 1) {
        mu[n] = alpha + beta[1]*x_full[n,1]+beta[2]*x[n,2];
    }
    else if (m[n] = 2) {
        mu[n] = alpha + beta[1]*x[n,1]+beta[2]*x_full[n,2];
    }
    else {
        mu[n] = alpha + beta[1]*x[n,1]+beta[2]*x[n,2];
    }
}
model {
  // Prior
  alpha ~ normal(0,10000);
  beta ~ normal(0,10000);
  sigma_2 ~ inv_gamma(0.01,0.01);
  for (i in 1:N) {
```

```
      for (j in 1:2) {
        x_full[i,j] ~ uniform(0,5);
      }
  }

  // Likelihood
  for (n in 1:N)
      y[n] ~ normal(mu[n], sqrt(sigma_2));
}
```



And, if we ignore the rows with the missing data...



10.3 Censored data

There are many situations when studying time to failures, recoveries and transitions where the items being studied do nothing during the observation period. One example is waiting for components to fail during stress testing. This is an example of *right censored data*.

Left censored data occur when the data point falls before our observation period. For example, the component has failed before we have started our observations.

Note that censored data are not the same as data coming from a truncated distribution. The latter gives zero probability to any data beyond the range of the truncation interval.

Imagine that we have times of failures from the following model,

$$X_i|\lambda \sim \text{Exp}(\lambda), \quad i = 1, \ldots, n,$$

for the n observations that occurred during our observation period $(0, T)$, and we have m experimental units yet to return as failures when we get to time T. When we have censoring, the likelihood is a mix of standard data-generating probability density functions and cumulative density functions:

$$\begin{aligned} \pi(\text{Data}|\lambda) &\propto \prod_{i=1}^{n} \pi(x_i|\lambda) \prod_{i=1}^{m} \left[\int_T^{\infty} \pi(x|\lambda) dx \right] \\ &\propto \lambda^n \exp\left(-\lambda \sum x_i\right) \exp\left(-\lambda T\right)^m. \end{aligned}$$

In Stan, we can handle this type of likelihood directly by using `target +=`:

```
model {
  // Likelihood
  target += n*log(lambda) - lambda*sum(x) - m*lambda*T;
}
```

Or we can use the original model as usual and treat the censored data as parameters:

```
parameters {
  real<lower=0> lambda;
  real<lower=T> x_censored[m];
}
model {
  // Likelihood
  x ~ exp(lambda);              // in vectorised form
  x_censored ~ exp(lambda);     // in vectorised form
}
```

Example 10.3.1

We are modelling time to failure of a certain component in hours using the model given in this section. We will stop observing after 100 hours.

A priori, we believe that the mean failure time will be about 80 hours (probably in the range 60-100). This means that we believe that the failure rate is about $1/80 = 0.0125$ (with a range of 1/100-1/60). Of course, the rate needs to be positive so we believe that a Gamma distribution with mode

$$\frac{\alpha - 1}{\beta}$$

and variance

$$\frac{\alpha}{\beta^2}$$

would be adequate. Setting the prior mode to be $1/80$ and a four standard deviation range to be 4/600 (a standard deviation of 0.00167), we can rearrange the formulae to get $\alpha = 58.3$ and $\beta = 4580$ to three significant figures.

In the 100 hours, we see seven failures at times 15.4, 18.5, 60.7, 80.54, 84.12, 91.11 and 91.13. By conjugacy, we have

$$\lambda|\underline{x} \sim \text{Gamma}(65.3, 5020),$$

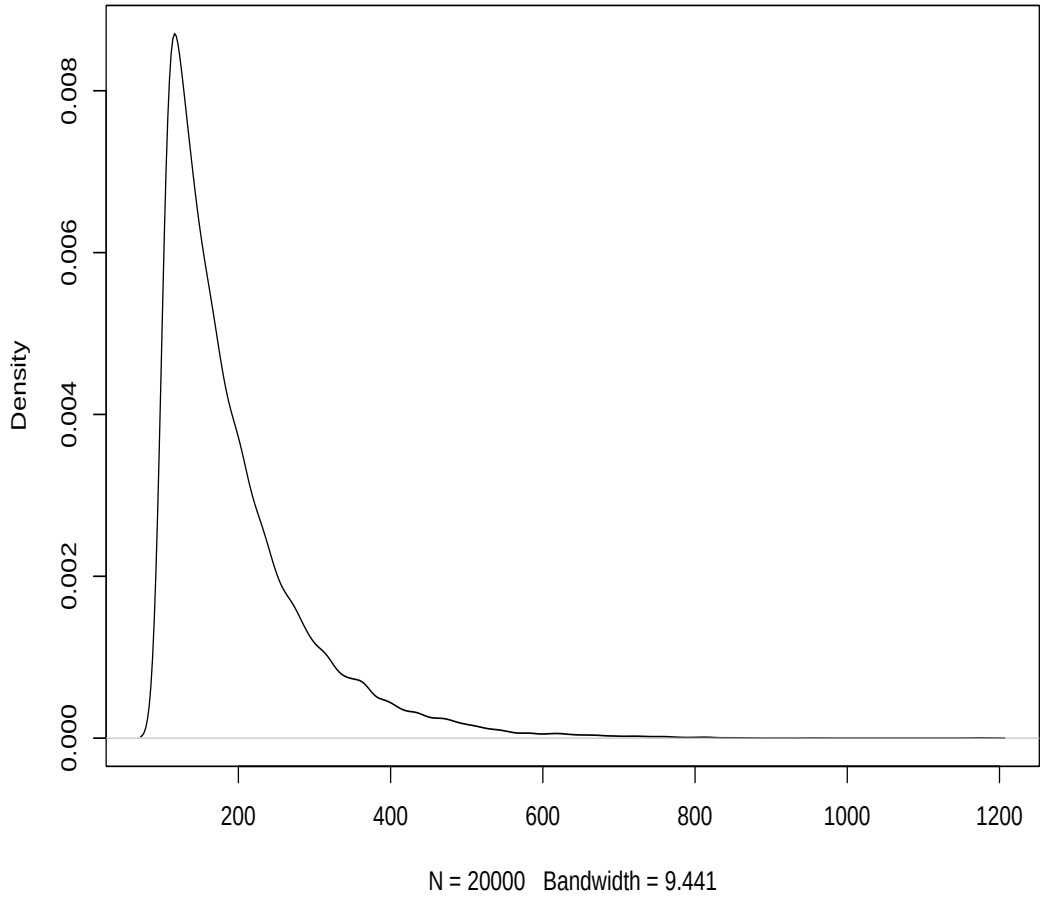
which gives us a posterior mode of 0.0130 and standard deviation of 0.00161.

If we know that there were ten additional units that did not fail during the 100 hours, then we can compute use Stan to get at the posterior for λ.

```
# Extract the posterior samples
post_sample <- extract(fit)

# Approximate the posterior mode for lambda
post_sample$lambda[which.max(post_sample$lp__)]

# Calculate the posterior standard deviation
sd(post_sample$lambda)
```

We get an approximate posterior mode of 0.0110 and standard deviation of 0.00135. We can also get an impression of when the remaining units might have failed using the posterior sample for the censored data.

```
# Kernel density estimation for x[1]
plot(density(post_sample$x_[,1]), main = NULL)
```


Density
0.000
0.002
0.004
0.006
0.008
200
400
600
800
1000
1200
N = 20000 Bandwidth = 9.441