Advanced Topics in Statistics

∼ Bayesian regression models ∼

4

UNIVERSITY OF EXETER

Bayesian regression models.

Standard (and non standard) regression models can be easily formulated within a Bayesian framework. We just have to

- Specify probability distribution (likelihood) for the data.
- Specify form of relationship between response and explanatory variables.
- Specify prior distributions for regression coefficients and any other unknown (nuisance) parameters.

Some advantages of a Bayesian formulation in regression modelling include:

- Easy to include parameter restrictions and other relevant prior knowledge.
- Easily extended to non-linear regression.
- Easily 'robustified' (see later).
- Easy to make inference about functions of regression parameters and/or predictions.
- Easily extended to handle missing data and covariate measurement error.

LINEAR REGRESSION.

Consider a simple linear regression with univariate Normal outcome y_i and a vector of covariates $x_{1i}, \ldots x_{pi}$, $i = 1, \ldots, n$

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \varepsilon_i,$$

$$\varepsilon_i \sim \text{Normal}(0, \sigma^2).$$

An equivalent Bayesian formulation would typically specify:

1. Likelihood: $y_i \sim N(\mu_i, \sigma^2)$,

   where $\mu_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}$.

2. Priors for all unknown parameters: $(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2) \sim$ Prior distributions

(Recall that in the BUGS language functions cannot be used as arguments in distributions, therefore μ_i has to be defined separately).

A typical choice of 'vague' prior distribution that will give numerical results similar to ordinary least squares or MLE (what we are used to frequentist statistics) is

$$\beta_k \sim N(0, 10^5), \quad k = 0, \ldots, p$$

$$1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$$

New York crime data.

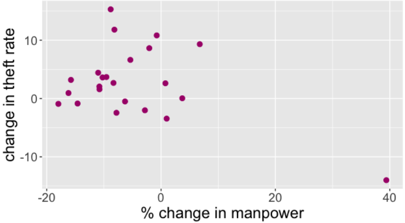
To give an example of Bayesian linear regression, we consider the following dataset that contains changes in police manpower and (seasonally adjusted) changes in thefts between a 27-week base period and a 58-week experimental period in 23 police precincts in New York.

MAN	THEFTS	DIST	MAN	THEFTS	DIST
-15.76	3.19	1	-6.3	-0.5	2
0.98	-3.45	1	39.4	-11	2
3.71	0.61	1	-10.79	2.05	2
-5.37	6.62	1	-8.16	11.8	2
-10.23	3.61	1	-2.82	-2.02	2
-8.32	2.67	1	-16.19	0.94	3
-7.8	-2.45	1	-11	4.42	3
6.77	9.31	1	-14.6	-0.86	3
-8.81	15.29	1	-17.96	-0.92	3
-9.56	3.68	1	0.76	2.61	3
-2.06	8.63	2	-10.77	1.58	3
-0.76	10.82	2			

In the dataset `MAN` is the changes in police manpower, `THEFT` is the changes in theft, and `DIST` is the district indicator (1 Downtown, 2 Mid-town, 3 Up-town).

New York crime data - Linear regression.

First we plot the relationship between the change in theft rate and the change in police manpower, and also look at the distribution of the change in theft rate for each district indicator.

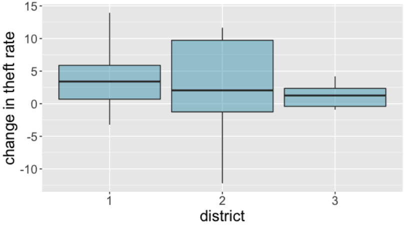


Then we fit a linear regression to the data with `THEFT` as the response, and `MAN` and `DIST` as explanatory variables.

That is, we fit the following model

$$\begin{aligned} \text{THEFT}_i &\sim \text{Normal}(\mu_i, \sigma^2), \quad i = 1, \ldots, 23 \\ \mu_i &= \alpha + \beta \times \text{MAN} + \langle \text{effect of DIST} \rangle \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \\ \alpha &\sim N(0, 100000) \\ \beta &\sim N(0, 100000) \end{aligned}$$

Prior on coefficients of DIST effect



The question is how we code up the effect of the district.

Specifying categorical covariates in BUGS.

The variable `DIST` is a 3-level categorical explanatory variable.

In BUGS there are two alternative ways of specifying a model with categorical variables.

1. Create the usual 'design matrix' in the data file:

```
  MAN    THEFT    DIST2   DIST3
-15.76    3.19        0       0  # district 1
  0.98   -3.45        0       0
  3.71    0.04        0       0
.......
 -9.56    3.68        0       0
 -2.06    8.63        1       0  # district 2
 -0.76   10.82        1       0
 -6.30   -0.50        1       0
.......
 -2.82   -2.02        1       0
-16.19    0.94        0       1  # district 3
-11.00    4.42        0       1
......
-10.77    1.58        0       1
```

That is, we add a column that indicates `DIST`=2, and another one that indicates `DIST`=3. Then the effect of `DIST`=1 is captured by the intercept, while `DIST2` and `DIST3` provides an adjustment to the intercept to the district 2 and 3 effects.

SPECIFYING CATEGORICAL COVARIATES IN BUGS.

Using the design matrix, the model in JAGS can be defined as

```
jags.mod <- function(){
    # likelihood
    for (i in 1:N) {
      THEFT[i] ~ dnorm(mu[i], tau)
      mu[i] <- alpha + beta*MAN[i] + delta2*DIST2[i] + delta3*DIST3[i]
    }
    # priors
    alpha ~ dnorm(0, 0.00001)
    beta ~ dnorm(0, 0.00001)
    delta2 ~ dnorm(0, 0.00001)
    delta3 ~ dnorm(0, 0.00001)
    tau ~ dgamma(0.001, 0.001)
    sigma2 <- 1/tau
}
```

Recall that BUGS parameterises the normal distribution in terms of the mean and precision τ, which is 1/variance.

But we are usually interested in the variance too, and adding the logical node `sigma2 <- 1/tau` allows us to get the posterior distribution of σ^2.

For this model the initial values would be something like

```
list(alpha = 1, beta = -2, delta2 = -2, delta3 = 4, tau = 2)
```

Specifying categorical covariates in BUGS.

2. Another option when specifying categorical covariates is to use the double-indexing feature of the BUGS language.
   In this case we can use the `DIST` column directly in the model definition as a way of indexing a parameter vector of length three. Similarly to the other method, `DIST`=1 is captured by the intercept `alpha`, thus we set `delta[1]` to zero.

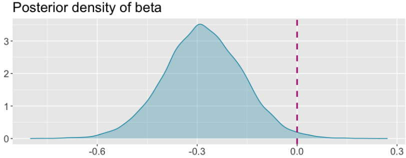
```
jags.mod <- function(){
  # likelihood
  for(i in 1:N){
    THEFT[i] ~ dnorm(mu[i], tau)
    mu[i] <- alpha + beta*MAN[i]+delta[DIST[i]]
  }
  # priors
  alpha ~ dnorm(0,1e-5)
  beta ~ dnorm(0,1e-5)
  delta[1] <- 0 # set coefficient for reference category to zero
  delta[2] ~ dnorm(0,1e-5)
  delta[3] ~ dnorm(0,1e-5)
  tau ~ dgamma(0.001,0.001)
  sigma2 <- 1/tau
}
```

In initial values file, we need to specify initial values for `delta[2]` and `delta[3]` but not `delta[1]`. Use following syntax:

```
list(alpha = 1, beta = -2, delta = c(NA, -2, 4), tau = 2)
```

New York crime data - Linear regression outcome.

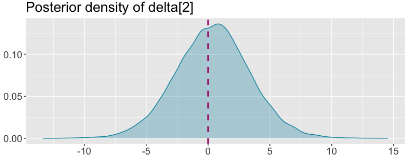
Change in theft rate per 1% increase in police manpower



Posterior mean: -0.282

95% credible interval: $(-0.518, -0.052)$

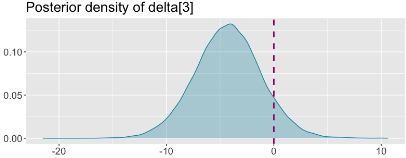
Change in theft rate in Midtown relative to Downtown



Posterior mean: 0.401

95% credible interval: $(-5.591, 6.457)$

Change in the theft rate in Uptown relative to Downtown



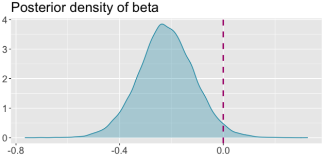
Posterior mean: -4.271

95% credible interval: $(-10.445, 1.919)$

New York crime data - Linear regression 2.

We can see that the 95% credible intervals for both `delta[2]` and `delta[3]` contain zero.

This implies that the `DIST` effect is not significant, therefore we can try removing it from the model, which gives the following output for `beta`.

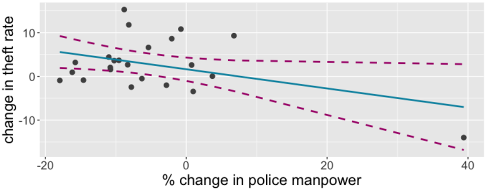


The remaining variable `beta` is significant.

- Posterior mean: -0.219
- 95% credible interval: $(-0.433, -0.004)$

We can also plot the fitted values (the mean of the posterior density of `mu[i]`) along with the 95% credible intervals.

Notice the outlying variable in the lower right corner.



How to deal with outliers?

Considering the outlying variable, do we trust the significance of `beta`?

The influential point corresponds to the 20th police precinct (which is observation number 14 in the dataset).

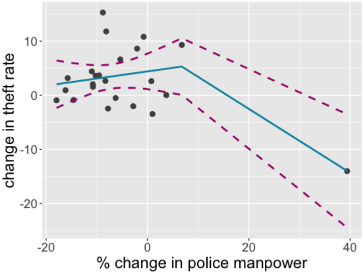
During the 2nd period, manpower assigned to this precinct was experimentally increased by about 40%.

There were no experimental increases in any other police precinct.

To understand how the model is influenced by the 20th precinct we can add an additional covariate corresponding to a binary indicator for precinct 20, which is equivalent to fitting a separate (saturated) model to this observation:

```
PREC20 <- rep(0,length(DIST))
PREC20[14] <- 1

jags.mod4 <- function(){
  for(i in 1:N){
   THEFT[i] ~ dnorm(mu[i], tau)
   mu[i] <- alpha + beta*MAN[i] + delta*PREC20[i]
  }
  alpha ~ dnorm(0,1e-5)
  beta ~ dnorm(0,1e-5)
  delta ~ dnorm(0,1e-5)
  tau ~ dgamma(0.001,0.001)
  sigma2 <- 1/tau
}
```



How to deal with outliers?

The coefficient `beta` in the previous enhanced model becomes insignificant. Therefore the significance of the change in manpower variable is caused by the outlier in our dataset. The question is, *what is the best way to deal with this outlier?*

In general, there are several approaches we can take when there are outliers in our dataset. These include

- Delete the outlier. This can be a reasonable approach e.g. if we are certain that the outlier is a result of incorrectly entered/measured data.
- Try a transformation. E.g. log transformation can pull in high values.
- Choose a model that is more robust to outliers.

Robust regression methods provide an alternative to least squares regression by imposing less restrictive assumptions. These methods attempt to dampen the influence of outlying observations in order to provide a better fit to the majority of the data.

E.g. Assume we have some observations far away from the mean. Linear regression assumes normally distributed errors. The normal distribution however has thin tails, thus it assigns relatively small probability to these observations, and can result in a biased model fit.

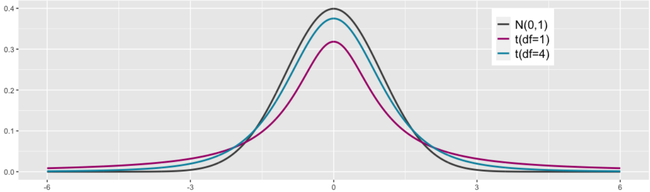
One approach to dealing with this problem could be choosing an error distribution with fatter tails, which can result in a model that is less affected by the outlying values.

Robust regression.

Recall that the t-distribution has fatter tails, that is, it assigns more weight to observations further away from the mean than the normal distribution.

The t-distribution with 1 degree of freedom has the heaviest tails, while as the degree of freedom increases, the t-distribution gets more and more similar to the normal distribution.

A t-distribution with 4-6 degrees of freedom is considered a good choice when fitting a robust regression model.



In the frequentist world, fitting a model with t-distributed errors (from scratch) can be quite tricky.

In JAGS however robust regression is relatively easy, since MCMC methods do not make a specific distributional assumption on the error term of the regression model.

New York crime data - Robust regression.

Since the observation corresponding to precinct 20 is not mismeasured data, it should not be thrown away.

Instead we could try to fit a robust regression where the errors are assumed to be t-distributed with 4 degrees of freedom.

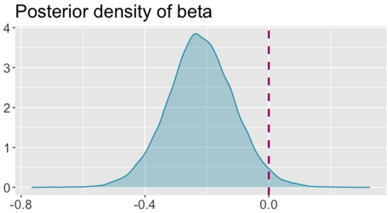
In JAGS the t-distribution has three parameters: the mean, the precision and the degrees of freedom.

Thus we have the following model definition for the robust regression:

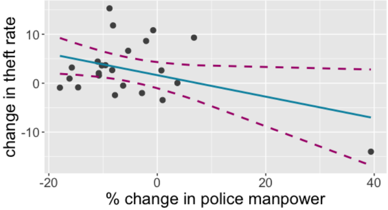
```
jags.mod <- function(){
  # likelihood
  for(i in 1:N){
    THEFT[i] ~ dt(mu[i], tau, 4)
    mu[i] <- alpha + beta*MAN[i]
  }
  # priors
  alpha ~ dnorm(0,1e-5)
  beta ~ dnorm(0,1e-5)
  tau ~ dgamma(0.001,0.001)
  sigma2 <- 1/tau
}
```

New York crime data - Robust regression outcome.

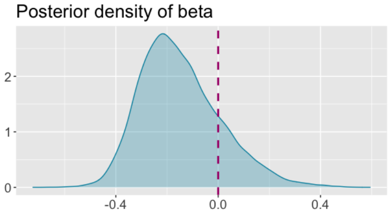
Linear regression outcome



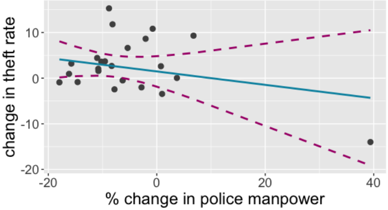
Posterior mean: -0.219
95% credible interval: (-0.433,-0.004)
(significant)



Robust regression outcome



Posterior mean: -0.147
95% credible interval: (-0.403,0.210)
(NOT significant)



TRADE UNION DENSITY.

Next we will look at an example where informative priors can help with estimating the effect of variables more precisely.

Recall that an informative prior expresses specific, definite information about a variable. We can use e.g. the posterior from another related problem or expert opinion when choosing informative priors.

Trade Union Density.

- We want to understand what explains cross-national variation in union density.

  Note that union density is defined as the percentage of the work force who belongs to a trade union.

- There are two competing theories:

  Wallerstein: union density depends on the size of the civilian labour force (LabF).

  Stephens: Union density depends on industrial concentration (IndC).

  Note that these two predictors correlate at -0.92.

The trade union density analysis is an example of regression analysis in comparative research.

Trade union density data.

Data: $n = 20$ countries with a continuous history of democracy since World War II.

Variables:

- Union density (Uden),
- (log) labour force size (LabF),
- industrial concentration (IndC),
- left wing government (LeftG), measured in late 1970s.

Uden	LabF	IndC	LeftG	Uden	LabF	IndC	LeftG
82.4	8.28	1.55	111.84	51.4	8.60	1.37	33.74
80.0	6.90	1.71	73.17	50.6	9.67	0.86	0.00
74.2	4.39	2.06	17.25	48.0	10.16	1.13	43.67
73.3	7.62	1.56	59.33	39.6	10.04	0.92	35.33
71.9	8.12	1.52	43.25	37.7	8.41	1.25	31.50
69.8	7.71	1.52	90.24	35.4	7.81	1.68	11.87
68.1	6.79	1.75	0.00	31.2	9.26	1.35	0.00
65.6	7.81	1.53	48.67	31.0	10.59	1.11	1.92
59.4	6.96	1.64	60.00	28.2	9.84	0.95	8.67
58.9	7.41	1.58	83.08	24.5	11.44	1.00	0.00

Trade union density - vague priors.

To compare the two theories we fit a linear regression model with centred covariates:

$$\text{Uden}_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = b_0 + b_1(\text{LeftG}_i - \overline{\text{LeftG}}) + b_2(\text{LabF}_i - \overline{\text{LabF}}) + b_3(\text{IndC}_i - \overline{\text{IndC}})$$

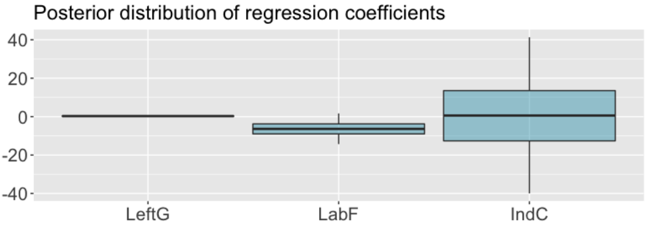
(Note that centring the covariates helps with convergence).

First we consider vague priors:

$$1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$$

$$b_0 \sim N(0{,}10^5) \qquad b_1 \sim N(0, 10^5)$$

$$b_2 \sim N(0{,}10^5) \qquad b_3 \sim N(0, 10^5)$$



Trade union density - Motivation for informative priors.

Data tend to favour Wallerstein's theory (union density depends on labour force size), but we can't draw strong conclusions from the previous analysis (notice that the 95% credible intervals for LabF and IndC both contain zero).

The problem is the small sample size and the multicollinear variables, which prevents us to adjudicate between the two theories.

However, other historical data are available that could provide further relevant information.

Incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients.

- **Both Wallerstein and Stephens** believe that left-wing governments assist union growth.
- Assuming 1 year of left-wing government increases union density by about 1% translates to effect size of 0.3.
- Confidence in direction of effect is represented by a prior SD that gives a 95% interval that excludes 0. That is

$$b_1 \sim N(0.3, 0.15^2)$$

- Vague prior is assumed for the intercept: $b_0 \sim N(0, 10^5)$.

Wallerstein and Stephens informative priors.

Wallerstein and Stephens agree in the LeftG effect (and the intercept), but the two theories differ in their assumptions about the LabF effect and the IndC effect.

Wallerstein informative prior

Believes in negative labour force.

- Comparison of Sweden and Norway in 1950: doubling of labour force corresponds to 3.5-4% drop in union density.

  On the log scale this corresponds to a labour force effect size of $\approx -3.5/\log(2) \approx -5$.
- Confidence in direction of effect is represented by a prior SD that gives a 95% interval that excludes 0:

$$b_2 \sim N(-5, 2.5^2)$$

- Vague prior assumed for IndC effect: $b_3 \sim N(0, 10^5)$.

Stephens informative prior

Believes in positive industrial concentration effect.

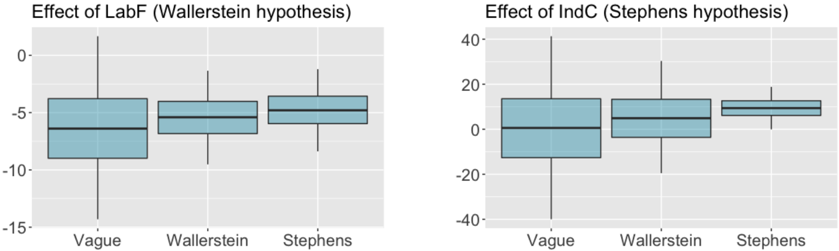
- Decline in industrial concentration in the UK in the 1980s: drop of 0.3 in industrial concentration corresponds to about 3% drop in union denisty.

  Thus the industrial concentration effect size is $\approx 3/0.3 = 10$.
- Confidence in direction of effect is represented by a prior SD that gives 95% interval that excludes 0:

$$b_3 \sim N(10, 5^2)$$

- Vague prior assumed for LabF effect: $b_2 \sim N(0, 10^5)$.

Trade Union density - Informative priors outcome.



Comments.

- Both sets of prior beliefs support inference that labour-force size decreases union density. (The whiskers of the above boxplots show the 95% CrI of the posteriors).
- Only Stephens prior supports conclusion that industrial concentration increases union density. (Wallerstein's 95% CrI for IndC contains zero).
- But we should note that the choice of prior is subjective. In case there is no consensus, can we be satisfied that data have been interpreted fairly? In this case we should consider the analysis of *sensitivity to priors* (e.g. repeat analysis using priors with increasing variance) and *sensitivity to data* (e.g. residuals, influence diagnostics).

Non linear regression models

In Bayesian statistics it is easy to define non-linear regression models.

Specification of these Bayesian non-linear models follows straightforwardly from previous discussion of linear models.

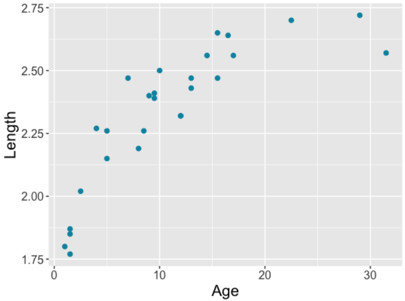
Even if there is no closed form solution available, it is still straightforward to obtain samples from posterior using MCMC.

Example: Dugongs

Carlin and Gelfand (1991) consider data on length (y_i) and age (x_i) measurements for 27 dugongs (sea cows) captured off the coast of Queensland.

age x_i	length y_i	age x_i	length y_i	age x_i	length y_i
1.0	1.80	8.0	2.19	13.0	2.47
1.5	1.85	8.5	2.26	14.5	2.56
1.5	1.87	9.0	2.40	15.5	2.65
1.5	1.77	9.5	2.39	15.5	2.47
2.5	2.02	9.5	2.41	16.5	2.64
4.0	2.27	10.0	2.50	17.0	2.56
5.0	2.15	12.0	2.32	22.5	2.70
5.0	2.26	12.0	2.32	29.0	2.72
7.0	2.47	13.0	2.43	31.5	2.57

Dugongs model.



To model the length of the dugongs we can use a nonlinear growth curve

$$y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \alpha - \beta\gamma^{x_i},$$

where $\alpha, \beta > 0$ and $\gamma \in (0, 1)$.

The following vague prior distributions satisfy the constraints we have on the parameters:

$$\alpha \sim \text{Unif}(0, 100)$$
$$\beta \sim \text{Unif}(0, 100)$$
$$\gamma \sim \text{Unif}(0, 1)$$

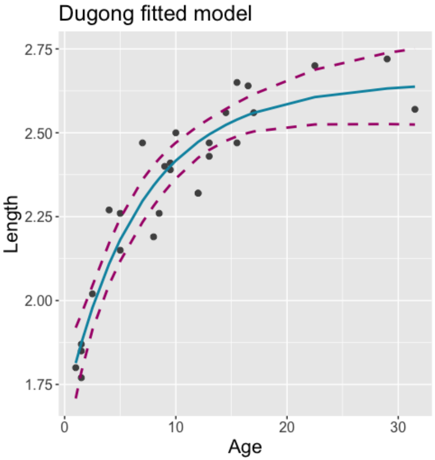
For the precision we can use $\tau = 1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$.

Dugongs in JAGS.

In JAGS we have the following non-linear model, which can be fitted the same way as the linear regression models.

```
jags.mod <- function(){
  for (i in 1 : N) {
    Y[i] ~ dnorm(mu[i],tau)
    mu[i] <- alpha-beta*gamma^x[i]
  }
  # priors
  alpha ~ dunif(0,100)
  beta ~ dunif(0,100)
  gamma ~ dunif(0,1.0)

  tau ~ dgamma(0.001,0.001)
  sigma <- 1/sqrt(tau)
}
```



Making predictions.

It's important to be able to predict unobserved quantities for

- 'Filling-in' missing or censored data.
- Model checking - are predictions 'similar' to observed data?
- Making predictions!

Making predictions is easy in JAGS. We just have to specify a stochastic node without a data-value, and this will be automatically predicted.

Note that this provides an automatic imputation of missing data.

As an example, consider again the dugongs dataset. Suppose we want to project beyond current observations, e.g. at ages 35 and 40.

There are two ways of doing this:

- We could either explicitly set up predictions, or
- We can choose to set the prediction up as missing data, and JAGS will automatically predict it.

Direct prediction.

When making direct prediction we can add direct statements our previous model definition to predict the dugong length at ages 35 and 40.

```
jags.mod <- function(){
  for( i in 1 : N ) {
    Y[i] ~ dnorm(mu[i], tau)
    mu[i] <- alpha - beta * gamma^x[i]
  }
  alpha ~ dunif(0, 100)
  beta ~ dunif(0, 100)
  gamma ~ dunif(0, 1.0)

  tau ~ dgamma(0.001, 0.001)
  sigma <- 1 / sqrt(tau)

  mu35 <- alpha-beta*gamma^35
  mu40 <- alpha-beta*gamma^40
  y35 ~ dnorm(mu35,tau)
  y40 ~ dnorm(mu40,tau)
}
```

Note that the interval around mu40 will reflect uncertainty concerning fitted parameters.

The interval around y40 will additionally reflect sampling error sigma and uncertainty about sigma.

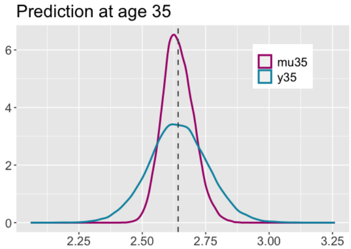
Direct prediction.

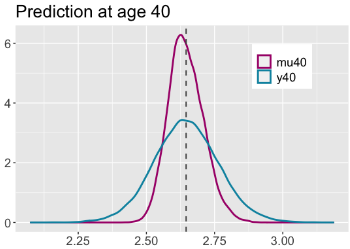
Since we have additional stochastic nodes in the model, these will have to be initialised. These are single nodes, thus we can use the form of the following list to initialise the chains

```
list(alpha = 1, beta = 1, tau = 1, gamma = 0.9, y35=2.4, y40=2.62)
```

Fitting the model gives the following outcome

```
      mu.vect sd.vect   2.5%    25%    50%    75%  97.5%  Rhat n.eff
mu35    2.642   0.061  2.528  2.601  2.639  2.679  2.772 1.001  8600
mu40    2.646   0.065  2.528  2.603  2.642  2.685  2.788 1.001  7400
y35     2.641   0.118  2.410  2.563  2.641  2.717  2.876 1.001  9300
y40     2.647   0.119  2.415  2.568  2.646  2.725  2.889 1.001  5700
```


Prediction at age 35


Prediction at age 40

PREDICTION AS MISSING DATA.

Prediction is asier to set up as missing data – JAGS will automatically predict it. In this case we use the original model definition!

```
x <- c( 1.0,  1.5,  1.5,  1.5, 2.5,   4.0,  5.0,  5.0,  7.0,
        8.0,  8.5,  9.0,  9.5, 9.5,  10.0, 12.0, 12.0, 13.0,
        13.0, 14.5, 15.5, 15.5, 16.5, 17.0, 22.5, 29.0, 31.5,35,40)

Y = c(1.80, 1.85, 1.87, 1.77, 2.02, 2.27, 2.15, 2.26, 2.47,
      2.19, 2.26, 2.40, 2.39, 2.41, 2.50, 2.32, 2.32, 2.43,
      2.47, 2.56, 2.65, 2.47, 2.64, 2.56, 2.70, 2.72, 2.57,NA,NA)

N = 29

jags.data <- list("x", "Y", "N")
```

The unknown quantities should still be initialised.

Since the ‘missing values’ are part of the data that we pass onto JAGS, we need to use this data structure when initialising the chains.

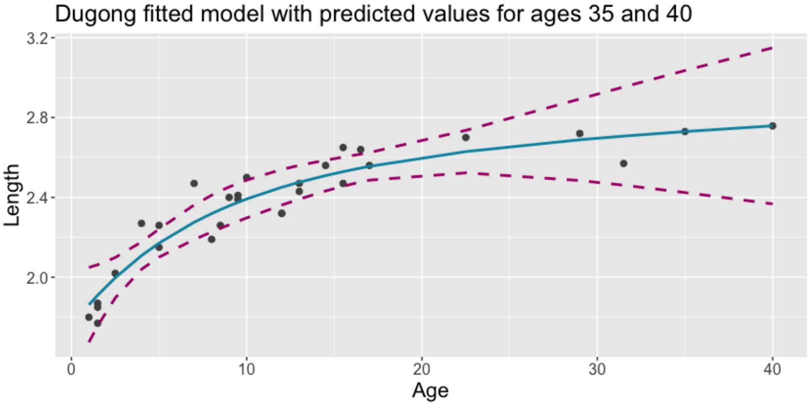
Use `NA` where the data is available (since these are not stochastic nodes), and initialise where data is missing:

```
list(alpha = 10, beta = 3, tau = 5, gamma = 0.1,Y = c(rep(NA,27),2.5,2.6))
```

Prediction as missing data - outcome.

Plotting the fitted model shows the imputed missing values.

Notice that we have greater uncertainty around the imputed nodes.


Dugong fitted model with predicted values for ages 35 and 40
Length
Age