# Model assessment

- **After** you have chosen the model for data $\pi(x|\theta)$ and the prior $\pi(\theta)$, and computed the posterior distribution $\pi(\theta|x)$…
  - Ok, we know it contains the complete information about the unknown parameter(s), and/or predicted variable(s).
  - **But how good is it?**

# Model assessment

- **BUGS-book:**

*”Model criticism and comparison inevitably involve a degree of judgement and cannot be reduced to a set of formal rules”*

# Model assessment

- If the posterior probabilities concern **unique events**, e.g. 'mass of jupiter', it is not possible to assess how ***often*** the model predicts this well.

- If possible to later find out the exact value for that, we will eventually see how close the posterior was…
  - This may not help us with modeling the next completely unique event!
  - Yet Bayesian inference is internally coherent approach

  – obeys the logic of probability calculus.

  – transparently shows the priors and conditional probabilities – open for anyone to see, criticise, and recompute with different choices. (Document your DAG well! → gives structure for your assessment)

# Model assessment

- A ’unique event’ may never become observable...
  - Then cannot check if the posterior probability pointed at the right answer or not.
    - E.g. population size of extinct animals in some geographical area, long time ago.
    - Maybe the best we can do, is to make sure all the evidence and uncertainties are used at least internally coherently in the analysis – if it is a probabilitistic analysis at all.
    - Transparency of the inference, the DAG.

# Model assessment

- If the posterior probabilities concern **repeatable events**, e.g. 'daily temperatures', we can repeatedly compare predictions with observable outcomes.
  - Frequentist properties of the predictions could be examined.
  - The predictive approach emphasizes observable variables instead of model parameters (which cannot be observed and therefore cannot be proven right or wrong).

# Model assessment

- *In all model assessments: both the prior and the conditional probability model of data (likelihood) are subject to model criticism.*
  - Sensitivity analysis: compute with different priors, different model choices. Could also check sensitivity to some data values (’outliers’?).
  - ”All models are wrong, but some of them are useful” – *George Box*.

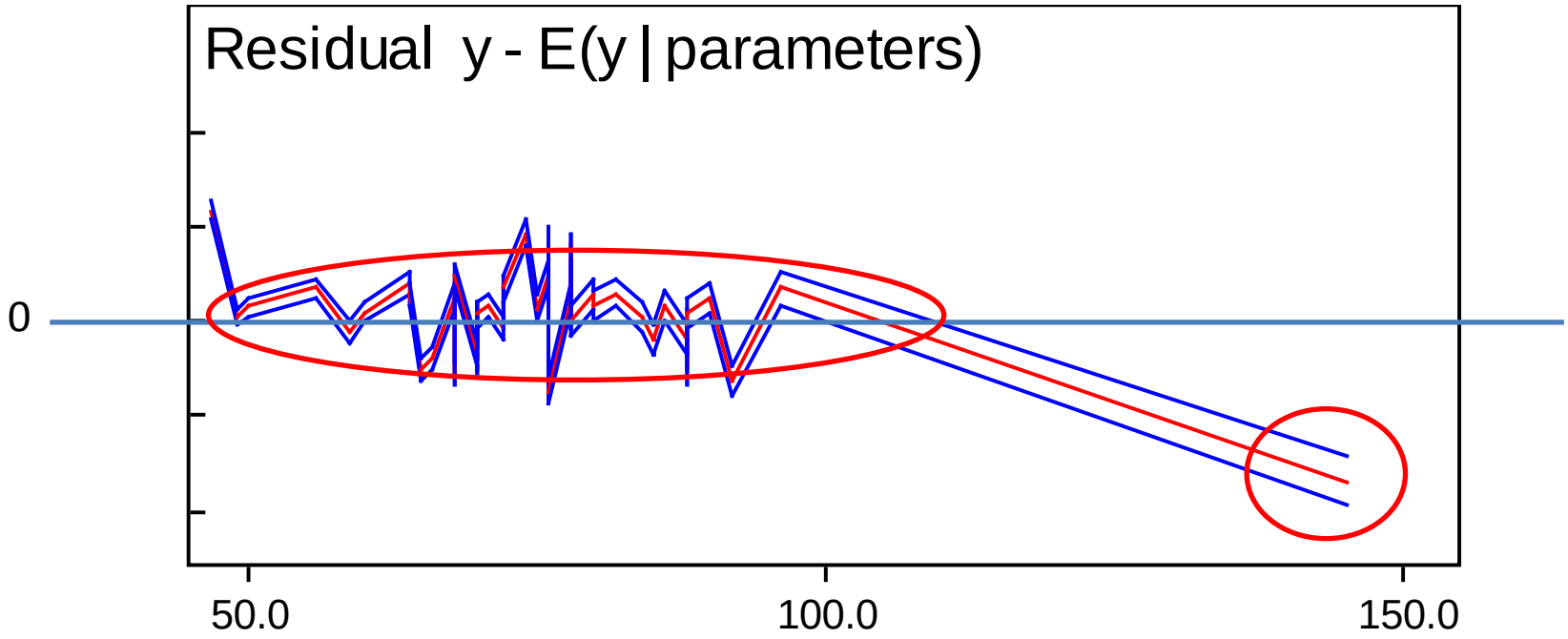
# Model assessment

- **Model fit** also usually depends on the number of parameters in it (model complexity).
- A parsimonious model that fits nearly as well as a more complex model, is thought to be better.
  - With too many parameters, overfitting can occur, and the model becomes less robust. (e.g. higher order regression terms)
  - Need to balance between model fit and model complexity.
  - The most accurate and complex description of the world is the world itself – but the purpose of a model is not to repeat all that – it should encapsulate the information, not inflate it.
  - Bayesian models combined with tools such as BUGS are so flexible that we can create and extend lots of models with slightly different structures.

# Model assessment

- **Residual plots**
  - In regression models where y = response, x = explanatory variables, classical residual plots are made of the residuals $y_i - E(y|x_i,\theta^*)$ calculated for the fitted parameter $\theta^*$. This ignores the uncertainty about $\theta$.
  - Bayesian residual plots could be made using $y_i - E(y|x_i,\theta)$ where the parameter(s) $\theta$ is sampled from the posterior.

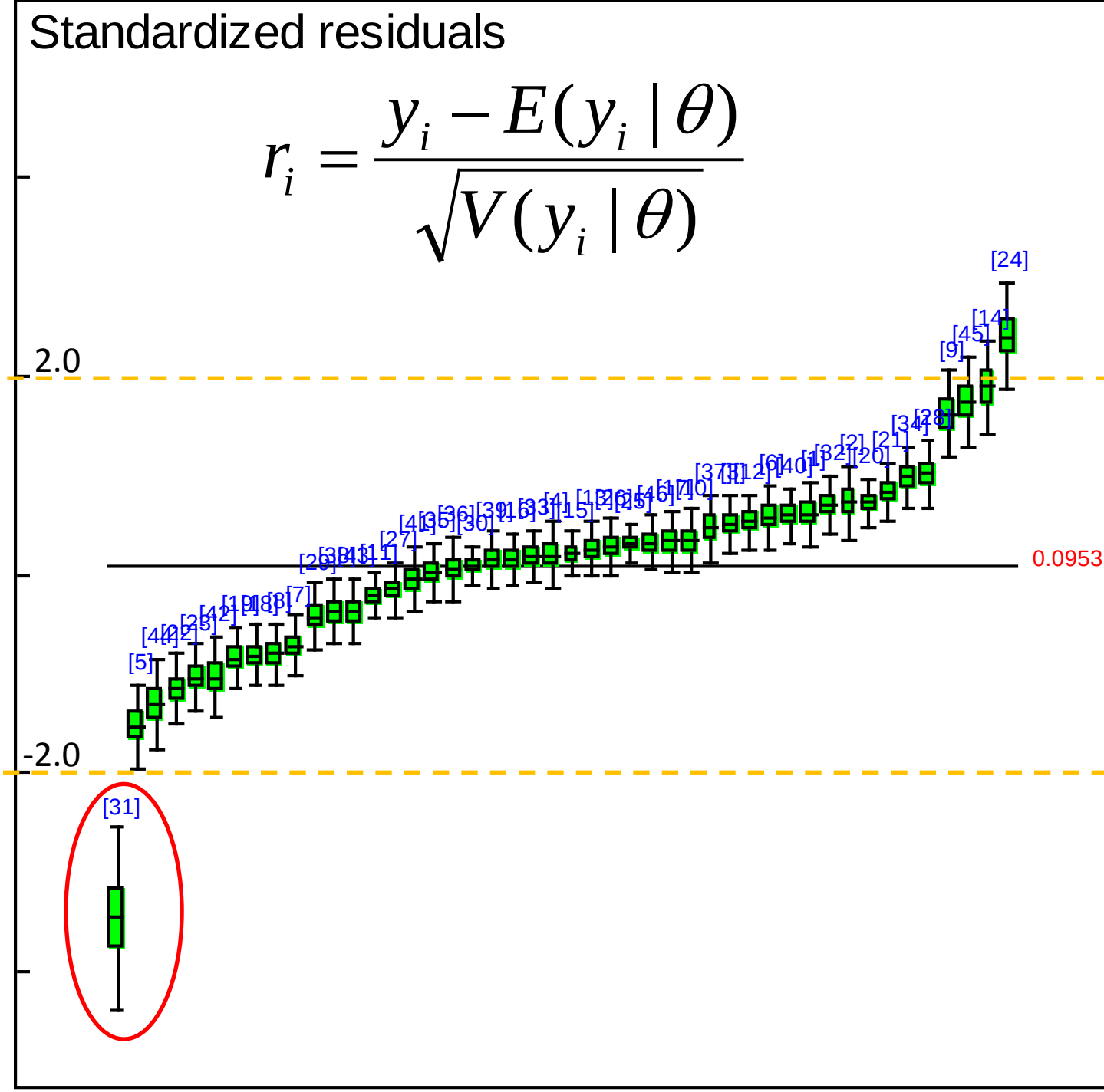
# Residual plot




- Should be evenly distributed around 0, large deviation indicates either outlier in data - or badly fitting model
  - Remove outliers?
  - Extension of model to better fit all data points?

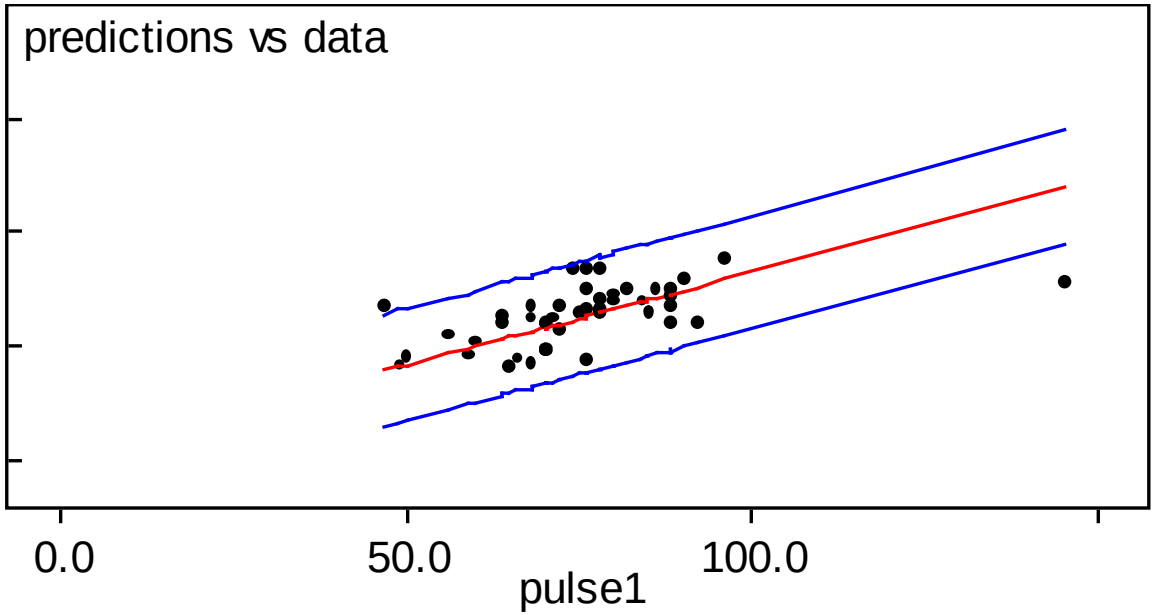
Box-plots plotted here as ranked.

Should be mostly between -2 and 2.



# Model assessment

- **Predictive model assessment**
  - First: check if the model can predict similar data points as those observed.
  - In regression model: produce predictive distributions for $y_i$ and compare if actual $y_i$ fall within 95% CI of posterior predictive distributions.



# Model assessment

- **Predictive model assessment**
  - Classical p-value $P(T(X^{pred}) > T(X^{obs}) \mid \theta)$
    - The value of $\theta$ is typically determined by a 'null hypothesis'.
  - Bayesian generalization of this:

$$P(T(X^{pred}, \theta) > T(X^{obs}, \theta) \mid X^{obs})$$

  - With fixed $\theta$, the classical p-value is a special case.
  - Values close to 0 or close to 1 indicate lack of fit.
  - Graphically, can compare $T(X^{pred},\theta)$ with $T(X^{obs},\theta)$ by plotting a scatter plot from MCMC sample. This should be symmetric about the 45° line.
  - Ideally, T should be chosen to reflect aspects that are relevant to the scientific purposes of the model.
  - Note: Bayesian predictive checks are not used to 'accept' or 'reject' a model, but rather to understand the limits of its applicability.

# Model assessment

- E.g. 'Omnibus' discrepancies:

$$T(X^{obs},\theta)=\sum\frac{(X_i-E(X\mid\theta))^2}{V(X\mid\theta)}$$

$$T(X^{pred},\theta)=\sum\frac{(X_i^{pred}-E(X\mid\theta))^2}{V(X\mid\theta)}$$

- E.g. Checking the lower tail:

$$T(X^{obs},\theta)=\mid X^{obs}_{smallest}-\theta\mid$$

$$T(X^{pred},\theta)=\mid X^{pred}_{smallest}-\theta\mid$$

# Example: linear model & max abs residual

- Checking the max absolute residual:

  ($Y_i$ = 2nd pulse, $X_i$ =1st pulse)

$$\text{regression mean}: E(Y_i \mid \theta_i) = \theta_i = \alpha X_i$$

$$T(Y^{obs}, \theta) = \max \mid Y_i^{obs} - \theta_i \mid$$

$$T(Y^{pred}, \theta) = \max \mid Y_i^{pred} - \theta_i \mid$$

$$P(T(Y^{pred}, \theta) > T(Y^{obs}, \theta) \mid \text{data}) \approx 0.065$$

  which is near 0, indicating some model fit problem in this respect.

- Easily computed from BUGS simulations

# Model assessment

- **DIC = Deviance Information Criteria**
  - Available in BUGS
  - Can be used for some model comparisons: better model has smaller DIC.
- Deviance: $D(x,\theta)=-2\log(\pi(x|\theta))$
  - If e.g. N-model with fixed variance, and mean $\theta$ as a parameter, then D is the same as the statistics

$$T(x,\theta)=\frac{1}{n}\sum_{i=1}^{n}(x_i-E(x_i\,|\,\theta))^2$$

  - This is mean squared error. Deviance can be thought as a generalization of this. Smaller deviance means a better fit.

# Model assessment

- **Posterior mean deviance** can be computed from MCMC sample of the parameter $\theta$:

$$\overline{D}(x) = E(D(x,\theta) \mid x) \approx \frac{1}{K}\sum_{k=1}^{K} D(x,\theta_k)$$

- Likewise, we may compute deviance by using an estimate of $\theta$, namely its posterior mean $E(\theta|x)$. Using this we get

$$\hat{D}(x) = D(x,\hat{\theta})$$

- Mean deviance describes better the errors of the *model*: it computes the average error over all possible values of $\theta$, not just with the fitted estimate of $\theta$.

# Model assessment

- The difference between these can be seen as the gain that can be achieved by fitting the model by $\hat{\theta}$. This gives the 'effective number of parameters':

$$p_D = \overline{D}(x) - \hat{D}(x)$$

- Also: 'the number of unconstrained parameters in the model'
  - A parameter does not count if it is completely constrained, or if all information about it comes from the prior.

# Model assessment

- As an example: consider multinomial model

$$x_1,\ldots,x_k \sim \text{Multinom}(p_1,\ldots,p_k,N)$$

- This has k parameters, but only k-1 are free, because of the constraint that the sum needs to be one.
- What is the effective number of parameters in a Bayesian model? This depends on prior.
  - Try computing $p_D$ with Dirichlet(1,…,1)-prior versus a more informative Dirichlet(100,…,100)-prior. In the latter case, the 'freedom' of parameters is smaller because prior is more concentrated (has smaller variance).

# Model assessment

- In comparing two models, better one has smaller DIC value:

$$E(D(x^{rep}, \hat{\theta}(x))) \approx 2\overline{D}(x) - \hat{D}(x) = DIC$$

- This can be automatically obtained from BUGS for each model (when applicable…).
- Small differences <5 are not meaningful.
- DIC looks for a good fitting parsimonious model.
- The model that would best predict a replicate dataset of the same structure as that currently observed.
- Works if posterior mean is a good estimate of model parameters, not if very skewed distributions.

# Model assessment

- **All the previous use the same data for computing the posterior and for the assessment of fit.**
  - Could also divide the data in 3 parts: one for constructing a prior, another for computing posterior, and the third for model assessment.
  - Cross-validation techniques: leave one observation out, predict it using the rest. Repeat this for every observation.
  - Real prediction of new observations, conditionally on the past data.

# Model assessment

- **Compute probabilities for different models?**
  - Could define prior probabilities for each model, then compute posterior probabilities and choose the one with highest probability.
  - Does not necessarily make sense, unless each model corresponds to some statement of real world and one of them must be true.

- **Bayesian model averaging (BMA)**, a mixture of several models can perform better than a single model.
  - E.g. mixture densities with k components, k=1,…,K, define K different possible models.
  - Could also let K to be unknown parameter.

- **Nonparametric Bayesian modeling (NPB)**
  - Priors are given for probability densities, in the space of probability density functions.
  - Could e.g. define a Poisson intensity with unknown number of change points in a piecewise constant function. → average of this is a nonparametric function. (or infinite parametric!)

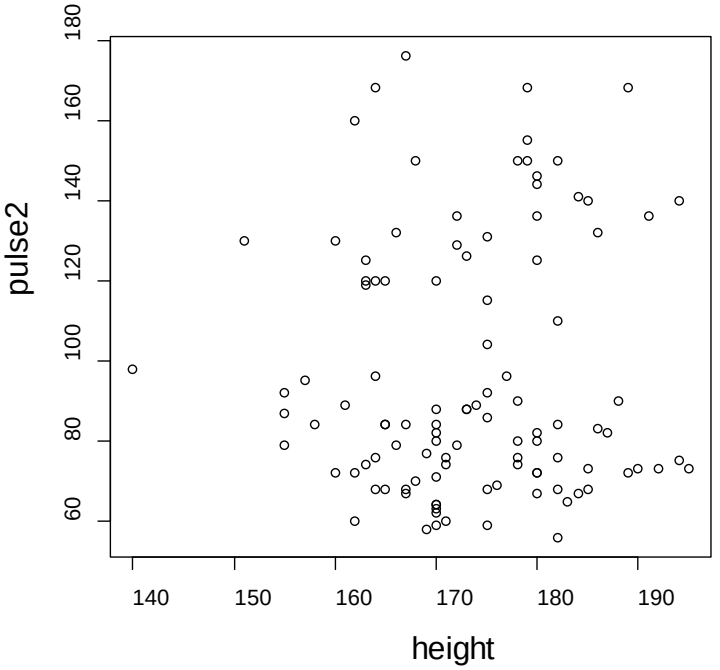
# Example: variable selection

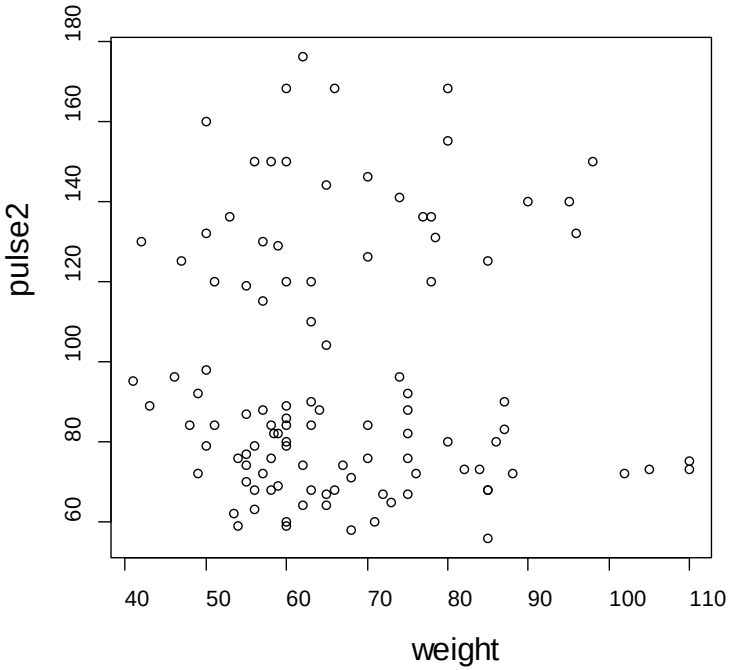
- Linear regression model for $\text{pulse}_2$
- Explanatory variables: height, weight, age, $\text{pulse}_1$
- Model choice problem: which variables to include? There are $2^4$ possible models.

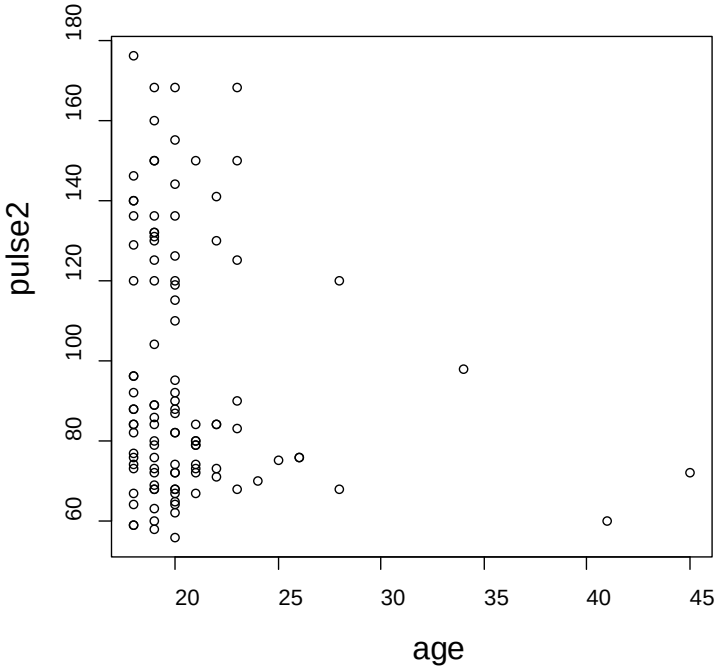
- Simple approach: include all & investigate posterior distribution of regression coefficients.
  - If estimated to be ≈0, if 95% CI includes 0, then the corresponding variable does not seem to have a significant effect.

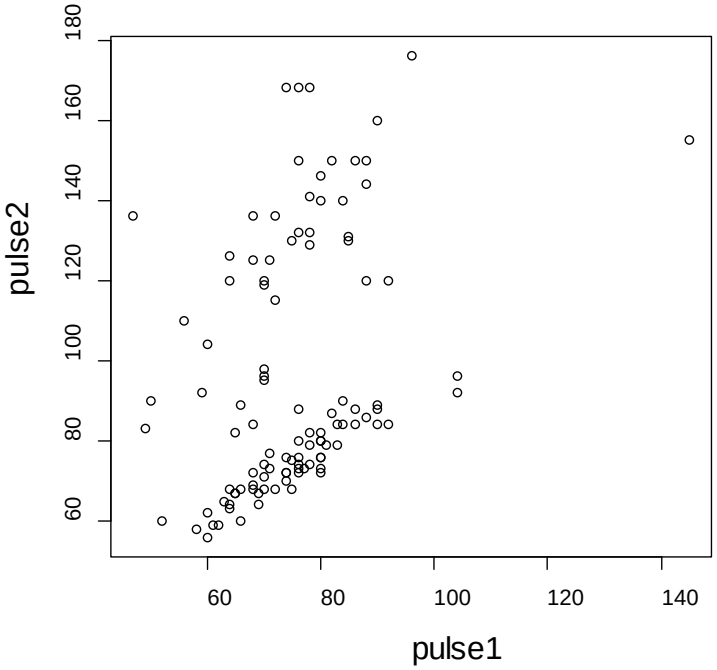
# Example: variable selection

- From scatter plots of data, it can be seen that $pulse_1$ is clearly correlated with $pulse_2$. Other variables not much.
  - Although there appears to be 2 groups: in one group pulse remains nearly the same, in another it becomes higher (due to running exercise between measurements)
  - Also: in original data 2 heights were <100cm. Errors? These were removed from data below


pulse2
height
60
80
100
120
140
160
180
140
150
160
170
180
190


pulse2
weight
60
80
100
120
140
160
180
40
50
60
70
80
90
100
110


pulse2
age
60
80
100
120
140
160
180
20
25
30
35
40
45


pulse2
pulse1
60
80
100
120
140
160
180
60
80
100
120
140

# Example: variable selection

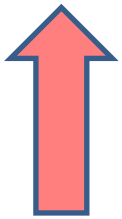
- Linear regression model for $pulse_2$

$E(pulse_2) = \mu_0 + \alpha_1 X_1 \gamma_1 + \alpha_2 X_2 \gamma_2 + \alpha_3 X_3 \gamma_3 + \alpha_4 X_4 \gamma_4$

- Each variable is included or excluded according to indicator variable $\gamma_i$.
- Prior: $P(\gamma_i=1)=0.5$
- $pulse_2 \sim$ Normal($E(pulse_2)$, $\tau$)
- Posterior distribution is now computed for all model parameters and indicators $\gamma$ jointly.
- Investigate which variables have high inclusion probability.

# Example: variable selection

| | mean | sd | val2.5pc | median | val97.5pc | start | sample |
|---|---|---|---|---|---|---|---|
| gamma[1] | 0.01854 | 0.1349 | 0.0 | 0.0 | 0.0 | 1000 | 79001 |
| gamma[2] | 0.007607 | 0.08689 | 0.0 | 0.0 | 0.0 | 1000 | 79001 |
| gamma[3] | 0.03049 | 0.1719 | 0.0 | 0.0 | 1.0 | 1000 | 79001 |
| gamma[4] | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1000 | 79001 |



**Posterior inclusion probability for each variable**

*Without going into detailed theory, some flavour of the technique is given in next slides.*

# Example: variable selection

- In MCMC, the simulation of $\gamma_i$ is the same as jumping between models over simulations.
  - Each model has a model specific set of parameters.
  - What happens to model parameters during simulations when that model is not chosen?
  - Compare with the simple model choice:

```
model{
x ~ dbin(p[gamma],n);  x <- 9; n <- 10
p[1] <- 0.5          # M0: model for unbiased coin
p[2] ~ dunif(0,1)  # M1: model for biased coin
gamma <- gamma0+1;  gamma0 ~ dbern(0.5)
}
```

  - Here: either the coin is 'unbiased' so that p=0.5, or 'biased' so that p is some value in [0,1]
  - gamma=1 is indicator for unbiased, gamma=2 for biased.
  - x=9 observed heads in n=10 coin tosses.

# Example: variable selection

- During iterations when gamma=1, parameter p[2] is simulated from its prior only (it's not connected to likelihood then).
- To get posterior of p, under model M2, we need to collect those iterations when M2 was actually chosen.
- The values of p when M2 was *not chosen* have no meaning.
- When M2 is not chosen p[2] can take values that are badly compatible with data → when jumping between models, MCMC is comparing the model with p=0.5 to a model with maybe p=0.01, and chooses with high probability the one with better p → MCMC can get jammed into one model.
- Trick: pseudo priors. A prior that depends on which model is currently chosen. When the model is not chosen, 'prior' is artificially defined around 'good estimates', but if model becomes chosen, the prior for its parameters is again the original prior.

# Example: variable selection

```
model{
  x ~ dbin(p[gamma],n);  x <- 9; n <- 10
  p[1] <- 0.5
  p[2] <- v*gamma0  +  u*(1-gamma0)
  u ~ dunif(0,1)
  v ~ distribution centered at x/n,   or  maybe  even constant  v <- x/n
  gamma <- gamma0+1;  gamma0 ~ dbern(0.5)
  }
```

- Here U(0,1) is the real prior for p[2], under model M2, but it is effective only when that model is chosen.
- Jumping between different models, each with different parameters is a complicated issue!
- With categorical explanatory variables, similar but multivariate priors needed.
- Below just one BUGS example for the Normal regression with variable choice…

# Example: variable selection

```
model{
for(i in 1:N){
pulse2[i] ~ dnorm(mu[i],tau)   # Normal model for observed 2nd pulse of ith person
mu[i] <- mu0+sum(z[i,])          # linear mean for regression model
for(p in 1:P){
z[i,p] <- alpha[p]*x[i,p]*gamma[p]
}
#  possible explanatory variables: height, weight, age, pulse1:
x[i,1] <- height[i]    # height of the ith person
x[i,2] <- weight[i]   # weight of the ith person
x[i,3] <- age[i]        # age of the ith person
x[i,4] <- pulse1[i]   # 1st pulse of the ith person
}
for(p in 1:P){
gamma[p] ~ dbern(0.5)            # inclusion indicator
alpha[p] ~ dnorm(m[p],t[p])      # pseudo priors  and actual N(0,0.001)-priors
m[p] <- gamma[p]*0 + (1-gamma[p])*pseudom[p]
t[p] <- gamma[p]*0.001 + (1-gamma[p])*pow(pseudos[p],-2)
}
mu0 ~ dnorm(0,0.001);  tau ~ dgamma(0.01,0.01)
}
```