Modelling missing and censored data

Overview

1. Missing data in Bayesian regression models
    - Data missing at random: full Bayesian modelling and Bayesian multiple imputation (30 min)

      Practical session (1) (30 min)
    - Data missing not at random (15 min)

      Practical session (2) (30 min, inc break)
2. Censored data and Bayesian survival modelling (30 min)

   Practical session (3) (45 min)

Missing data in Bayesian regression models

Missing data in regression models (“item non-response”)

Example from Lecture 3.

- Predict cholesterol level `chol_4yr` at next visit, based on current value and other cardiovascular risk factors
- Many of these are missing at different observations

```
##   age gender chol_curr sbp  dbp diabetes    hba1c chol_4yr [etc.]
## 1  67   Male   NA 119.0 71.0        0       NA        5.5
## 2  47 Female  6.7 131.5 75.5        0       NA        8.6
## 3  53 Female  8.6 141.0 95.5        0 30.05475        8.4
## 4  57 Female  8.4    NA   NA        0 34.42635        6.2
## 5  62 Female   NA 132.5 74.0        0       NA        7.5
## 6  65   Male  7.0 174.5 86.0        0       NA        3.8
[etc.]
```

1. Outcomes missing at random
2. Covariates missing at random
3. Outcomes or covariates missing not at random

Missing data in regression models ("item non-response")

Example from Lecture 3.

- Predict cholesterol level `chol_4yr` at next visit, based on current value and other cardiovascular risk factors
- Many of these are missing at different observations

```
##   age gender chol_curr sbp  dbp diabetes    hba1c chol_4yr [etc.]
## 1  67   Male   NA 119.0 71.0        0        NA        5.5
## 2  47 Female  6.7 131.5 75.5        0        NA        8.6
## 3  53 Female  8.6 141.0 95.5        0 30.05475        8.4
## 4  57 Female  8.4    NA   NA        0 34.42635        6.2
## 5  62 Female   NA 132.5 74.0        0        NA        7.5
## 6  65   Male  7.0 174.5 86.0        0        NA        3.8
[etc.]
```

1. Outcomes missing at random
2. Covariates missing at random
3. Outcomes or covariates missing not at random

Mechanisms for missing data: informal definitions

- Missing completely at random (MCAR): whether an observation is made doesn't depend on any other variables
- Missing at random (MAR): whether an observation is made depends (at most) on variables that are observed.
- Missing not at random (MNAR): whether an observation is made depends on the value of the missing observation itself
    - Miss medical appointments when disease less severe.
    - Sensitive questions not answered in surveys.

Why not just drop all records with missing data in?

- Increase precision by making use of all information (e.g. some covariates observed but not others)
- Alleviate bias that can occur if complete data are not a representative subset

Mechanisms for missing data: informal definitions

- Missing completely at random (MCAR): whether an observation is made doesn't depend on any other variables
- Missing at random (MAR): whether an observation is made depends (at most) on variables that are observed.
- Missing not at random (MNAR): whether an observation is made depends on the value of the missing observation itself
  - Miss medical appointments when disease less severe.
  - Sensitive questions not answered in surveys.

Why not just drop all records with missing data in?

- Increase precision by making use of all information (e.g. some covariates observed but not others)
- Alleviate bias that can occur if complete data are not a representative subset

Outcomes missing, but not covariates?

If outcomes are missing at random, OK to exclude the observation for regression, since:

- records with missing outcomes Y provide no information about the relation between outcome Y and covariates X.
- The same is true for records where all covariates are missing.

In a model $p(Y|X, \theta)$, we could impute missing Y_i by sampling from the posterior predictive distribution in a model fitted to the complete data.

- Simple to do in JAGS and similar software (see Session 3)
- But doing this will not change the estimate of θ

However, if outcomes are missing not at random, the fact that the observation is missing provides information (see later . . .)

Outcomes missing, but not covariates?

If outcomes are missing at random, OK to exclude the observation for regression, since:

- records with missing outcomes Y provide no information about the relation between outcome Y and covariates X.
- The same is true for records where all covariates are missing.

In a model $p(Y|X,\theta)$, we could impute missing Y_i by sampling from the posterior predictive distribution in a model fitted to the complete data.

- Simple to do in JAGS and similar software (see Session 3)
- But doing this will not change the estimate of θ

However, if outcomes are missing not at random, the fact that the observation is missing provides information (see later . . .)

Bayesian regression with missing covariates

General regression model for an outcome Y with multiple covariates X, where some covariates are missing for some observations

Define a model for the joint distribution of $p(X, Y)$ given parameters ϕ, θ, usually decomposed as

$$p(X, Y|\phi, \theta) = p(X|\phi)p(Y|X, \theta)$$

- $p(X|\phi)$: model with the covariates considered as the outcomes
- $p(Y|X, \theta)$: model of interest for outcomes given covariates

(assuming θ and ϕ are distinct)

Two approaches:

1. Full Bayesian modelling
2. Multiple imputation

Bayesian regression with missing covariates

General regression model for an outcome Y with multiple covariates X, where some covariates are missing for some observations

Define a model for the joint distribution of $p(X, Y)$ given parameters ϕ, θ, usually decomposed as

$$p(X, Y|\phi, \theta) = p(X|\phi)p(Y|X, \theta)$$

- $p(X|\phi)$: model with the covariates considered as the outcomes
- $p(Y|X, \theta)$: model of interest for outcomes given covariates

(assuming θ and ϕ are distinct)

Two approaches:

1. Full Bayesian modelling
2. Multiple imputation

Full Bayesian modelling with missing data

Covariate model $p(X|\phi)$, model of interest $p(Y|X,\theta)$

- Define a full joint probability model for Y and X, decomposed as $p(X|\phi)$ and $p(Y|X,\theta)$
- Partition $X = (X_{mis}, X_{obs})$ into missing and observed values

Missing observations X_{mis} considered as unknown parameters

Estimated along with other parameters of the Bayesian model

- The model $p(X|\phi)$ acts as
  - the "prior" for the X_{mis} and
  - the "likelihood" for the observed X_{obs}
- Estimate the joint posterior $(\theta, \phi, X_{mis}|Y, X_{obs})$ in a single fit

May be challenging in practice to define / fit / check a joint model – posterior distribution may be complicated and slow to sample from

Full Bayesian modelling with missing data

Covariate model $p(X|\phi)$, model of interest $p(Y|X,\theta)$

- Define a full joint probability model for Y and X, decomposed as $p(X|\phi)$ and $p(Y|X,\theta)$
- Partition $X = (X_{mis}, X_{obs})$ into missing and observed values

Missing observations X_{mis} considered as unknown parameters

Estimated along with other parameters of the Bayesian model

- The model $p(X|\phi)$ acts as
  - the "prior" for the X_{mis} and
  - the "likelihood" for the observed X_{obs}
- Estimate the joint posterior $(\theta, \phi, X_{mis}|Y, X_{obs})$ in a single fit

May be challenging in practice to define / fit / check a joint model – posterior distribution may be complicated and slow to sample from

Full Bayesian modelling with missing data

Covariate model $p(X|\phi)$, model of interest $p(Y|X,\theta)$

- Define a full joint probability model for Y and X, decomposed as $p(X|\phi)$ and $p(Y|X,\theta)$
- Partition $X = (X_{mis}, X_{obs})$ into missing and observed values

Missing observations X_{mis} considered as unknown parameters

Estimated along with other parameters of the Bayesian model

- The model $p(X|\phi)$ acts as
  - the "prior" for the X_{mis} and
  - the "likelihood" for the observed X_{obs}
- Estimate the joint posterior $(\theta, \phi, X_{mis}|Y, X_{obs})$ in a single fit

May be challenging in practice to define / fit / check a joint model – posterior distribution may be complicated and slow to sample from

Example: predicting cholesterol levels

```
##   age male chol_curr  sbp  dbp diabetes    hba1c chol_4yr
## 1  67    1   NA 119.0 71.0        0       NA       5.5
## 2  47    0  6.7 131.5 75.5        0       NA       8.6
## 3  53    1  8.6 141.0 95.5        0 30.05475       8.4
## 4  57    1  8.4    NA   NA        0 34.42635       6.2
## 5  62    1   NA 132.5 74.0        0       NA       7.5
```

- Model of interest: predict $Y =$ `chol_4yr` from a linear regression model based on $X =$ `chol_curr`, age, `male`
- Covariate model to predict missing values of `chol_curr`
  - e.g. regress `chol_curr` on, age, `male`, using data where all of these are observed

A full Bayesian missing data model in JAGS

Regression models defined simultaneously for `chol_curr` X and `chol\_4yr` Y.

`NA` values are supplied in the data in R for the elements of `chol_curr` that are missing X_{mis}.

MCMC sample drawn from joint posterior of missing values and all other regression parameters ϕ : `ac`, `bcmale`, `bcage`, `prec_c`, θ : `acn`, `bcnmale`, `bcnage`, `bcnchol`, `prec_cn`

```
model {
  for (i in 1:n){
    ## Covariate model: predict missing
    ## current cholesterol based on age and gender
    chol_curr[i] ~ dnorm(mu_c[i], prec_c)
    mu_c[i] <- ac + bcmale*male[i] + bcage*age[i]

    ## Model of interest: predict next cholesterol
    ## based on current cholesterol, age and gender
    chol_4yr[i] ~ dnorm(mu_cn[i], prec_cn)
    mu_cn[i]     <- acn + bcnmale*male[i] + bcnage*age[i] +
                    bcnchol*chol_curr[i]
```

What predictors to include in the covariate model?

No need to put the outcome of interest Y as a predictor in the model for X

This is because the regression model specified for $(Y|X, \theta)$ already allows (informally)

- the relationship between X and Y to be estimated given cases where both variables are observed (Y and X_{obs})
- missing values of X (X_{mis}) to be inferred from the corresponding values of Y.

More formally: the "likelihood" for $(Y|X_{obs}, X_{mis}, \theta)$, "likelihood" for $X_{obs}|\phi$, and the "prior" for $(X_{mis}|\phi)$ together give the joint posterior for $\theta, \phi, X_{mis}|Y, X_{obs}$.

Data-generating assumptions would be unclear if we simultaneously had a regression model of X on Y, as well as Y on X

What predictors to include in the covariate model?

No need to put the outcome of interest Y as a predictor in the model for X

This is because the regression model specified for $(Y|X,\theta)$ already allows (informally)

- the relationship between X and Y to be estimated given cases where both variables are observed (Y and X_{obs})
- missing values of X (X_{mis}) to be inferred from the corresponding values of Y.

More formally: the "likelihood" for $(Y|X_{obs}, X_{mis}, \theta)$, "likelihood" for $X_{obs}|\phi$, and the "prior" for $(X_{mis}|\phi)$ together give the joint posterior for $\theta, \phi, X_{mis}|Y, X_{obs}$.

Data-generating assumptions would be unclear if we simultaneously had a regression model of X on Y, as well as Y on X

What predictors to include in the covariate model?

No need to put the outcome of interest Y as a predictor in the model for X

This is because the regression model specified for $(Y|X, \theta)$ already allows (informally)

- the relationship between X and Y to be estimated given cases where both variables are observed (Y and X_{obs})
- missing values of X (X_{mis}) to be inferred from the corresponding values of Y.

More formally: the "likelihood" for $(Y|X_{obs}, X_{mis}, \theta)$, "likelihood" for $X_{obs}|\phi$, and the "prior" for $(X_{mis}|\phi)$ together give the joint posterior for $\theta, \phi, X_{mis}|Y, X_{obs}$.

Data-generating assumptions would be unclear if we simultaneously had a regression model of X on Y, as well as Y on X

Another approach to missing data: Bayesian multiple imputation

1. Fit the covariate model $p(X|\phi)$
2. Draw missing covariate values X_{mis} from the posterior predictive distribution of $p(X|\phi)$.
3. Do multiple draws, giving multiple imputed, completely-observed datasets $X^{(1)}, X^{(2)}, \ldots$
4. Fit the model of interest $p(Y|X^{(r)}, \theta)$ to each imputed dataset r in turn (giving a posterior sample $\theta^{(r,1)}, \theta^{(r,2)}, \theta^{(r,3)}, \ldots$)
5. Pool together the posterior samples of θ from all analyses r done in step 4.

Note: this is an approximation to a full Bayesian posterior $(\theta, \phi, X_{mis}|Y, X_{obs})$ — since we ignore the information from Y provided by the model $p(Y|X, \theta)$ when imputing X_{mis}.

Another approach to missing data: Bayesian multiple imputation

1. Fit the covariate model $p(X|\phi)$
2. Draw missing covariate values X_{mis} from the posterior predictive distribution of $p(X|\phi)$.
3. Do multiple draws, giving multiple imputed, completely-observed datasets $X^{(1)}, X^{(2)}, \ldots$
4. Fit the model of interest $p(Y|X^{(r)}, \theta)$ to each imputed dataset r in turn (giving a posterior sample $\theta^{(r,1)}, \theta^{(r,2)}, \theta^{(r,3)}, \ldots$)
5. Pool together the posterior samples of θ from all analyses r done in step 4.

Note: this is an approximation to a full Bayesian posterior $(\theta, \phi, X_{mis}|Y, X_{obs})$ — since we ignore the information from Y provided by the model $p(Y|X, \theta)$ when imputing X_{mis}.

Another approach to missing data: Bayesian multiple imputation

1. Fit the covariate model $p(X|\phi)$
2. Draw missing covariate values X_{mis} from the posterior predictive distribution of $p(X|\phi)$.
3. Do multiple draws, giving multiple imputed, completely-observed datasets $X^{(1)}, X^{(2)}, \ldots$
4. Fit the model of interest $p(Y|X^{(r)}, \theta)$ to each imputed dataset r in turn (giving a posterior sample $\theta^{(r,1)}, \theta^{(r,2)}, \theta^{(r,3)}, \ldots$)
5. Pool together the posterior samples of θ from all analyses r done in step 4.

Note: this is an approximation to a full Bayesian posterior $(\theta, \phi, X_{mis}|Y, X_{obs})$ — since we ignore the information from Y provided by the model $p(Y|X, \theta)$ when imputing X_{mis}.

Advantages of multiple imputation over full Bayesian missing data model?

If not much missingness, full joint model may not be worthwhile.

Multiple imputation splits up the estimation into two simpler stages

- may be computationally faster than MCMC for joint model
- easier than developing/checking/revising a full joint model

How to define covariate imputation model $p(X|\phi)$?
Typically works well to use all variables to predict all others

- Needn't be Bayesian. MICE ("multivariate imputation using chained equations") is a popular non-Bayesian method
- 10-100 imputations is usually sufficient to characterise posterior of θ {(Zhou and Reiter (https://www.tandfonline.com/doi/abs/10.1198/tast.2010.09109))}

Model of interest $p(y|X, \theta)$? Concentrate most effort to get this correct (choice of covariates, model form, priors. . .)

Advantages of multiple imputation over full Bayesian missing data model?

If not much missingness, full joint model may not be worthwhile.

Multiple imputation splits up the estimation into two simpler stages

- may be computationally faster than MCMC for joint model
- easier than developing/checking/revising a full joint model

How to define covariate imputation model $p(X|\phi)$?
Typically works well to use all variables to predict all others

- Needn't be Bayesian. MICE ("multivariate imputation using chained equations") is a popular non-Bayesian method
- 10-100 imputations is usually sufficient to characterise posterior of θ {(Zhou and Reiter (https://www.tandfonline.com/doi/abs/10.1198/tast.2010.09109))}

Model of interest $p(y|X, \theta)$? Concentrate most effort to get this correct (choice of covariates, model form, priors. . .)

Advantages of multiple imputation over full Bayesian missing data model?

If not much missingness, full joint model may not be worthwhile.

Multiple imputation splits up the estimation into two simpler stages

- may be computationally faster than MCMC for joint model
- easier than developing/checking/revising a full joint model

How to define covariate imputation model $p(X|\phi)$?
Typically works well to use all variables to predict all others

- Needn't be Bayesian. MICE ("multivariate imputation using chained equations") is a popular non-Bayesian method
- 10-100 imputations is usually sufficient to characterise posterior of θ {(Zhou and Reiter (https://www.tandfonline.com/doi/abs/10.1198/tast.2010.09109))}

Model of interest $p(y|X,\theta)$? Concentrate most effort to get this correct (choice of covariates, model form, priors. . .)

Practical question (1): effect of modelling missing data

Model of interest is now a logistic regression to predict diabetes given current cholesterol, but some cholesterol values are missing.

Compare the posterior odds ratios of interest, between models with

(a) missing current cholesterol values estimated from a full Bayesian model

(b) records with missing current cholesterol excluded

Modelling missing data may or may not affect these much.
Balance between potential:

- reduction in bias + increase in precision from including additional data
- decrease in precision due to accounting for uncertainty about the missing values, and having more parameters in the model

Implementing multiple imputation in JAGS

No special new techniques needed. Define a loop (e.g. in R) to

(a) impute a dataset from the covariate model, and

(b) fit the model of interest to the imputed data in JAGS, saving the posterior sample for θ.

Finally combine all samples for θ, giving the final approximated posterior distribution.

Demonstration code provided in the practical session material.

- uses a non-Bayesian linear regression model for imputation,
- multiple ways in R to process the posterior samples cleanly.

Non-random missingness. Some theory

New idea: consider the fact that an observation is missing as a modelled outcome in itself

Denote by M the set of indicators that each data point is missing or observed. And suppose Y is all data (outcomes and covariates)

Assume we can define a joint model as

$p(Y, M|\psi, \theta) = p(Y|\theta)p(M|Y, \psi)$, with $Y = (Y_{obs}, Y_{mis})$

If data missing at random (only depends on data that are observed), then $p(M|Y_{obs}, Y_{mis}, \psi) = p(M|Y_{obs}, \psi)$. Hence:

$p(Y_{obs}, M|\psi, \theta) = p(Y_{obs}|\theta)p(M|Y_{obs}, \psi)$ (by integrating over Y_{mis})

Then we can estimate our parameters of interest θ from Y_{obs} alone

- and ignore Y_{mis} and M: "missingness is ignorable".

Otherwise, have to model the missingness mechanism $p(M|Y, \ldots)$

Non-random missingness. Some theory

New idea: consider the fact that an observation is missing as a modelled outcome in itself

Denote by M the set of indicators that each data point is missing or observed. And suppose Y is all data (outcomes and covariates)

Assume we can define a joint model as

$p(Y, M|\psi, \theta) = p(Y|\theta)p(M|Y, \psi)$, with $Y = (Y_{obs}, Y_{mis})$

If data missing at random (only depends on data that are observed), then $p(M|Y_{obs}, Y_{mis}, \psi) = p(M|Y_{obs}, \psi)$. Hence:

$p(Y_{obs}, M|\psi, \theta) = p(Y_{obs}|\theta)p(M|Y_{obs}, \psi)$ (by integrating over Y_{mis})

Then we can estimate our parameters of interest θ from Y_{obs} alone

- and ignore Y_{mis} and M: "missingness is ignorable".

Otherwise, have to model the missingness mechanism $p(M|Y, \ldots)$

Non-random missingness. Some theory

New idea: consider the fact that an observation is missing as a modelled outcome in itself

Denote by M the set of indicators that each data point is missing or observed. And suppose Y is all data (outcomes and covariates)

Assume we can define a joint model as

$p(Y, M|\psi, \theta) = p(Y|\theta)p(M|Y, \psi)$, with $Y = (Y_{obs}, Y_{mis})$

If data missing at random (only depends on data that are observed), then $p(M|Y_{obs}, Y_{mis}, \psi) = p(M|Y_{obs}, \psi)$. Hence:

$p(Y_{obs}, M|\psi, \theta) = p(Y_{obs}|\theta)p(M|Y_{obs}, \psi)$ (by integrating over Y_{mis})

Then we can estimate our parameters of interest θ from Y_{obs} alone

- and ignore Y_{mis} and M: "missingness is ignorable".

Otherwise, have to model the missingness mechanism $p(M|Y, \ldots)$

Non-random missingness. Some theory

New idea: consider the fact that an observation is missing as a modelled outcome in itself

Denote by M the set of indicators that each data point is missing or observed. And suppose Y is all data (outcomes and covariates)

Assume we can define a joint model as

$p(Y, M|\psi, \theta) = p(Y|\theta)p(M|Y, \psi)$, with $Y = (Y_{obs}, Y_{mis})$

If data missing at random (only depends on data that are observed), then $p(M|Y_{obs}, Y_{mis}, \psi) = p(M|Y_{obs}, \psi)$. Hence:

$p(Y_{obs}, M|\psi, \theta) = p(Y_{obs}|\theta)p(M|Y_{obs}, \psi)$ (by integrating over Y_{mis})

Then we can estimate our parameters of interest θ from Y_{obs} alone

- and ignore Y_{mis} and M: "missingness is ignorable".

Otherwise, have to model the missingness mechanism $p(M|Y, \dots)$

Non-random missingness. Some theory

New idea: consider the fact that an observation is missing as a modelled outcome in itself

Denote by M the set of indicators that each data point is missing or observed. And suppose Y is all data (outcomes and covariates)

Assume we can define a joint model as

$p(Y, M|\psi, \theta) = p(Y|\theta)p(M|Y, \psi)$, with $Y = (Y_{obs}, Y_{mis})$

If data missing at random (only depends on data that are observed), then $p(M|Y_{obs}, Y_{mis}, \psi) = p(M|Y_{obs}, \psi)$. Hence:

$p(Y_{obs}, M|\psi, \theta) = p(Y_{obs}|\theta)p(M|Y_{obs}, \psi)$ (by integrating over Y_{mis})

Then we can estimate our parameters of interest θ from Y_{obs} alone

- and ignore Y_{mis} and M: "missingness is ignorable".

Otherwise, have to model the missingness mechanism $p(M|Y, \ldots)$

Non-random missingness. Some theory

New idea: consider the fact that an observation is missing as a modelled outcome in itself

Denote by M the set of indicators that each data point is missing or observed. And suppose Y is all data (outcomes and covariates)

Assume we can define a joint model as

$p(Y, M|\psi, \theta) = p(Y|\theta)p(M|Y, \psi)$, with $Y = (Y_{obs}, Y_{mis})$

If data missing at random (only depends on data that are observed), then $p(M|Y_{obs}, Y_{mis}, \psi) = p(M|Y_{obs}, \psi)$. Hence:

$p(Y_{obs}, M|\psi, \theta) = p(Y_{obs}|\theta)p(M|Y_{obs}, \psi)$ (by integrating over Y_{mis})

Then we can estimate our parameters of interest θ from Y_{obs} alone

- and ignore Y_{mis} and M: "missingness is ignorable".

Otherwise, have to model the missingness mechanism $p(M|Y, \dots)$

Outcomes that are missing not at random

Example: People more likely to miss their clinic appointment (where cholesterol etc. are measured) at times when they are at lower risk (e.g. lower cholesterol)?

In this case, the record with missing outcome provides information.

Impossible to estimate the effect of the value of Y on the chance that Y is not observed, since we don't know the value of Y when it is missing!

But we can

- make an assumption about this effect
- model the consequences of different assumptions for the results of interest

“Selection model” for non-random missing outcomes

Missingness model: logistic regression for the 0/1 indicator m_i that outcome i is missing.

Assume, e.g. the odds of missingness reduces by 10% for each unit increase in cholesterol:

$$\begin{aligned} m_i &\sim Bern(p_i) \\ \log(p_i/(1-p_i)) &= a + by_i \\ & \quad b = \log(0.9) \text{ assumed known} \end{aligned}$$

Couple this with the model of interest:

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- Full Bayesian model, estimate the joint posterior of a, α, β, σ and the values of the missing y_i.

Intuitively: assumption about b allows us to infer what the y_i might have been for i when they were not observed

“Selection model” for non-random missing outcomes

Missingness model: logistic regression for the 0/1 indicator m_i that outcome i is missing.

Assume, e.g. the odds of missingness reduces by 10% for each unit increase in cholesterol:

$$\begin{aligned} m_i &\sim Bern(p_i) \\ \log(p_i/(1-p_i)) &= a + by_i \\ & \quad b = \log(0.9) \text{ assumed known} \end{aligned}$$

Couple this with the model of interest:

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- Full Bayesian model, estimate the joint posterior of a, α, β, σ and the values of the missing y_i.

Intuitively: assumption about b allows us to infer what the y_i might have been for i when they were not observed

“Selection model” for non-random missing outcomes

Missingness model: logistic regression for the 0/1 indicator m_i that outcome i is missing.

Assume, e.g. the odds of missingness reduces by 10% for each unit increase in cholesterol:

$$\begin{aligned} m_i &\sim Bern(p_i) \\ \log(p_i/(1-p_i)) &= a + by_i \\ & \quad b = \log(0.9) \text{ assumed known} \end{aligned}$$

Couple this with the model of interest:

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- Full Bayesian model, estimate the joint posterior of a, α, β, σ and the values of the missing y_i.

Intuitively: assumption about b allows us to infer what the y_i might have been for i when they were not observed

JAGS implementation of non-random missing outcomes

Missingness indicators miss[i] provided as 0/1 data.
Provide NA for any missing values of chol_4yr[i] in the data.

```
model {
  for (i in 1:n){
    ## Missingness model: missing value of chol_4yr influences
    ## chance p of missingness, by known amount b
    miss[i]      ~ dbern(p[i])
    logit(p[i]) <- a + b*chol_4yr[i]

    ## Model of interest: predict cholesterol at next visit in 4 years
    ## based on current cholesterol, age and gender
    chol_4yr[i] ~ dnorm(mu_cn[i], prec_cn)
    mu_cn[i]     <- acn + bcnmale*male[i] + bcnage*age[i]
        + bcnchol*chol_curr[i]
```

Instead of a fixed value for the effect *b* of missing value on missingness, could use an informative prior distribution based on expert belief, but sensitivity analysis advised if this is uncertain.

Covariates missing not at random

Specify three different models:

- Missingness model for the chance that covariate is missing
- Covariate model to describe how the missing value of a covariate X_i depends on the values of other covariates, as before
- Model of interest for how outcome depends on covariates

Estimate joint posterior for all unknown parameters and missing values, given observed data and model assumptions, by MCMC.

Practical session (2): outcomes missing not at random

See practical sheet, Section 2.

JAGS code provided for a linear regression with outcome missing not at random. Exercises:

- Practice specifying priors for a linear regression based on substantive knowledge
- Extend JAGS code to account for a non-random missingness mechanism
- Compare results with missing data ignored

Censored data and Bayesian survival modelling

Censoring

A data point x is censored when we

don’t know its exact value. . .

but we know it is within some interval: $x \in (a, b)$

Example: a weighing scale only represents weights from 0kg to 9kg. Objects heavier than 9kg are shown as 9kg

Censoring

A data point x is censored when we

don't know its exact value...
but we know it is within some interval: $x \in (a, b)$

Example: a weighing scale only represents weights from 0kg to 9kg. Objects heavier than 9kg are shown as 9kg

Censoring

A data point x is (interval-)censored when we

don't know its exact value...
but we know it is within some interval: $x \in (a, b)$

Example: a weighing scale only represents weights from 0kg to 9kg. Objects heavier than 9kg are shown as 9kg

Censoring

A data point x is (interval-)censored when we

don't know its exact value...
but we know it is within some interval: $x \in (a, b)$

- Right-censoring: we just know $x > a$ (thus $b = \infty$, assuming x comes from unbounded distribution)
- Left-censoring: we just know $x < b$ (thus $a = -\infty$, or the lowest possible value)

Censoring

A data point x is (interval-)censored when we

don't know its exact value...
but we know it is within some interval: $x \in (a, b)$

- Right-censoring: we just know $x > a$ (thus $b = \infty$, assuming x comes from unbounded distribution)
- Left-censoring: we just know $x < b$ (thus $a = -\infty$, or the lowest possible value)

Note that censoring is different from truncation:

- where an observation is only included in the data if it is in (a, b), i.e. observations outside (a, b) are impossible

Likelihood in models for censored data

Censored data are partially known: the exact value is missing, but they contribute information that should be included in the likelihood.

Let $\{y_i : i = 1, \ldots\}$ be the true underlying data.

- Assume generated as $y_i \sim p_i(y|\theta)$
- We know exact y_i for some i (observed data), e.g. $i <= 6$
- but only know $y_i > c_i$ for $i = 7, 8, 9$ (right-censored data)

Likelihood contribution: $\prod_{i=1}^{6} p_i(y_i|\theta) \prod_{i=7}^{9}(1 - F_i(c_i|\theta))$

- where $p()$ is the probability density function and $F()$ is the cumulative distribution function

Thus we learn about θ through both observed and censored data

- we could also estimate the posterior predictive distribution of the censored values $y_i : i = 7, 8, 9$

Likelihood in models for censored data

Censored data are partially known: the exact value is missing, but they contribute information that should be included in the likelihood.

Let $\{y_i : i = 1, \ldots\}$ be the true underlying data.

- Assume generated as $y_i \sim p_i(y|\theta)$
- We know exact y_i for some i (observed data), e.g. $i <= 6$
- but only know $y_i > c_i$ for $i = 7, 8, 9$ (right-censored data)

Likelihood contribution: $\prod_{i=1}^{6} p_i(y_i|\theta) \prod_{i=7}^{9} (1 - F_i(c_i|\theta))$

- where $p()$ is the probability density function and $F()$ is the cumulative distribution function

Thus we learn about θ through both observed and censored data

- we could also estimate the posterior predictive distribution of the censored values $y_i : i = 7, 8, 9$

Reminder of “survival” data, and an example

Or “time-to-event” data, as the “event” need not be death.

Example: length of stay y_i in hospital for a set of people i admitted to hospital for some condition.

- Some people i were discharged from hospital at known times y_i.
- The remaining people i are still in hospital now, so $y_i > c_i$: right-censored times from admission to discharge

Estimate distribution (between individuals) of length of stay →

- $E(Y)$, expected length of stay, and
- $P(Y > 20)$, probability that a person will spend more than 20 days in hospital

to plan future health care resources needed (e.g. in a pandemic)

Parametric models for survival

Bayesian inference requires a parametric distribution to model the time-to-event Y.

Time-to-event distributions can be understood in terms of their hazard $p(t)/S(t)$:

- rate of death at t for people who have survived up to time t, $S(t) = P(T > t)$ is the survivor function, and $p(t)$ is the PDF

Examples:

- exponential distribution : constant hazard λ.
- piecewise exponential: different constant values in different (pre-specified) time periods.
- Weibull: hazard $\lambda\alpha t^{\alpha-1}$ is either increasing with time ($\alpha > 1$), decreasing ($\alpha < 1$) or constant ($\alpha = 1$)

Many others, including gamma, log-normal, log-logistic, Gompertz, 3-parameter generalised Gamma…

Parametric models for survival

Bayesian inference requires a parametric distribution to model the time-to-event Y.

Time-to-event distributions can be understood in terms of their hazard $p(t)/S(t)$:

- rate of death at t for people who have survived up to time t, $S(t) = P(T > t)$ is the survivor function, and $p(t)$ is the PDF

Examples:

- exponential distribution : constant hazard λ.
- piecewise exponential: different constant values in different (pre-specified) time periods.
- Weibull: hazard $\lambda \alpha t^{\alpha-1}$ is either increasing with time ($\alpha > 1$), decreasing ($\alpha < 1$) or constant ($\alpha = 1$)

Many others, including gamma, log-normal, log-logistic, Gompertz, 3-parameter generalised Gamma...

Covariates in parametric survival models

Covariates included by expressing some parameter as a linear (or log-linear) function of covariates. This might define, e.g.

a proportional hazards model:

- hazard ratio between two groups, defined by different values of $\boldsymbol{x}$, is constant with time t.
- e.g. set $\lambda(\boldsymbol{x}) = exp(\boldsymbol{\beta}^T\boldsymbol{x})$ in the exponential or Weibull models

an accelerated failure time model:

- means that changing $\boldsymbol{x}$ speeds or slows the time unit t.
- reparameterise Weibull hazard as $\mu\alpha(\mu t)^{\alpha-1}$, survivor function is $S(t) = \exp(-(\mu t)^{\alpha})$ and set $\mu(\boldsymbol{x}) = exp(\boldsymbol{\beta}^T\boldsymbol{x})$

More advanced parametric survival models with covariates

Any parameter of a parametric survival distribution can be modelled as a function of covariates.

- e.g. in the Weibull with rate λ and shape α, $\lambda(\boldsymbol{x}) = exp(\boldsymbol{\beta}^T \boldsymbol{x})$ only gives proportional hazards for constant α
- But can also model α as a function of $\boldsymbol{x}$
  - e.g. different α for different subgroups
  - gives non-proportional hazards between these groups.

Might also include random effects in the linear predictor, giving a hierarchical model

- e.g, $\lambda_{ij} = exp(U_j + \boldsymbol{\beta}^T \boldsymbol{x}_{ij})$, $U_j \sim N(\mu, \sigma^2)$ for person i in group j. a "shared frailty" model.
  Constant hazard ratio between two people in different groups with same $\boldsymbol{x}$, e.g. $\exp(U_2 - U_1)$

Bayesian considerations for survival modelling

Parameters of parametric survival models don't always have real-world interpretations.

Prior beliefs can be influential. If data are censored / incomplete, you may have less information in your data than you think.

We'll give some examples of

- defining → checking by simulation → revising prior beliefs
- for the parameters of a simple parametric survival model (Weibull)

Choosing priors in a Weibull survival model

Hazard $h(t) = \lambda \alpha t^{\alpha - 1}$. Survivor function $S(t) = \exp(-\lambda t^{\alpha})$
α and λ hard to interpret in isolation, but functions of them are meaningful.

e.g. mean time to event is $(1/\lambda)^{\alpha} Gamma(1 + 1/\alpha)$

- which is $1/\lambda$ if $\alpha = 1$ (exponential distribution).

So we could choose priors on λ by specifying beliefs about the mean time to event with $\alpha = 1$

We might hope that the beliefs described by the resulting priors don't vary too much with different α.

- We will check this by simulation

Weibull distribution: priors on rate parameter. Example.

Expected length of stay in hospital?

Guess 10 days, with 97.5% prior belief that it is less than 30 days.

Could then define a log-normal prior:

$\log(1/\lambda) \sim N(\mu, \sigma^2)$ with
$\mu = \log(10) \quad \sigma = (\log(30) - \log(10))/1.96$

(upper 97.5% quantile is 1.96 standard deviations from the mean)

...note this is a judgement about the mean time $E(Y)$

- not a statement about the distribution between individuals of the time Y, like "we expect 97.5% of people to stay less than 30 days".

Weibull distribution: priors on shape parameter

Shape α describes the change through time in the hazard.

e.g. hazard ratio for a doubling of time (since $t = 0$) is

$HR = ((2t)^{\alpha-1}/t^{\alpha-1}) = 2^{\alpha-1}$, so $\alpha = \log_2(HR) + 1$.

So could translate beliefs on HR to beliefs on α

But note the Weibull is only defined for $\alpha > 0$, so only permits HR > 0.5

Or could use trial-and-error:

- specify a positive distribution on $\alpha \rightarrow$ check the implied HRs represent your prior belief $\rightarrow$ modify if needed . . .

Example: $\alpha \sim Gamma(1, 1)$. Implied 95% credible interval for HR:

```
> 2^(qgamma(c(0.025, 0.975), 1, 1) - 1)
[1] 0.5088519 6.4481238
```

Weibull distribution: priors on shape parameter

Shape α describes the change through time in the hazard.

e.g. hazard ratio for a doubling of time (since $t = 0$) is

$HR = ((2t)^{\alpha-1}/t^{\alpha-1}) = 2^{\alpha-1}$, so $\alpha = \log_2(HR) + 1$.

So could translate beliefs on HR to beliefs on α

But note the Weibull is only defined for $\alpha > 0$, so only permits HR > 0.5

Or could use trial-and-error:

- specify a positive distribution on α → check the implied HRs represent your prior belief → modify if needed . . .

Example: $\alpha \sim Gamma(1, 1)$. Implied 95% credible interval for HR:

```
> 2^(qgamma(c(0.025, 0.975), 1, 1) - 1)
[1] 0.5088519 6.4481238
```

Weibull distribution: priors on shape parameter

Shape α describes the change through time in the hazard.

e.g. hazard ratio for a doubling of time (since $t = 0$) is

$HR = ((2t)^{\alpha-1}/t^{\alpha-1}) = 2^{\alpha-1}$, so $\alpha = \log_2(HR) + 1$.

So could translate beliefs on HR to beliefs on α

But note the Weibull is only defined for $\alpha > 0$, so only permits HR > 0.5

Or could use trial-and-error:

- specify a positive distribution on α → check the implied HRs represent your prior belief → modify if needed ...

Example: $\alpha \sim Gamma(1, 1)$. Implied 95% credible interval for HR:

```
> 2^(qgamma(c(0.025, 0.975), 1, 1) - 1)
[1] 0.5088519 6.4481238
```

Simulating consequences of priors

Verify priors on λ, α are jointly plausible, by simulation

Could simulate implied prior for mean $E(Y)$ marginalised over α.

```
> alpha <- rgamma(10000, 1, 1)
> lambda <- 1 / exp(rnorm(10000, mean=log(10),
                          sd=(log(30) - log(10)) / 1.96))
> mean_time <- (1/lambda)^alpha * gamma(1 + 1/alpha)
> round(quantile(mean_time, c(0.25, 0.75, 0.975)))
25% 75%  97.5%
  8 344  2e+45
```

Prior on α implies surprisingly wide range of $E(Y)$

- Very long tail: implausibly large upper 97.5% quantile, instead of original guess of 30 days

Trial and error (or numerical search) could be used to determine prior SDs for α and λ that imply a more informative prior for $E(Y)$,
e.g. $\alpha \sim Gamma(k, k)$ has mean 1, but smaller variance for bigger k

Simulating from prior predictive distribution of data

Prior predictive distribution of individual lengths of hospital stay?

Simulate, e.g. using the `rweibull` function in R.

- Inconveniently, the parameterisation is different from the one used in JAGS.
- `dweibull` has the survivor function $\exp(-(t/\mathit{scale})^{\mathit{shape}})$ instead of $\exp(-\lambda t^{\alpha})$, so that $\mathit{shape} = \alpha$, $\lambda = (1/\mathit{scale})^{\mathit{shape}}$, $\mathit{scale} = \lambda^{-1/\alpha}$.

```
alpha <- rgamma(10000, 3, 3)
prior_pred <- rweibull(10000, scale=lambda^(-1/alpha), shape=alpha)
round(quantile(prior_pred, c(0.025, 0.5, 0.9, 0.975)))
```

```
##  2.5%   50%   90% 97.5%
##     0     7   409 72275
```

10% chance of hospital stays longer than around 409 days

Again, may want to decrease the prior SD on α and/or λ to give a smaller chance of such extreme values.

Bayesian survival models, and further resources

Survival models are just another form of Bayesian model: same issues of model building, priors, checking, comparing apply here.

- Censoring makes some implementation details trickier, e.g. JAGS code for censored data, computing deviance / DIC, posterior predictive model checking.

Further resources: Wide and useful range of Bayesian survival models implemented in various R packages

- e.g. `rstanarm`, `brms`, `survHE`, `INLA`
- many parametric distributions, including piecewise-exponential analogue of the Cox model, flexible spline-based models
- left-censoring, interval censoring, left-truncation, hierarchical (multilevel) models
- Alvares et al.: examples of survival models in JAGS `https://onlinelibrary.wiley.com/doi/epdf/10.1002/sim.8933`

Practical: Implementing survival models in JAGS

1. Demonstration of how to model censored survival data in JAGS. The JAGS syntax for censoring is unusual!
2. Basic exercise: implementing a Weibull survival model, and calculating posterior distributions of quantities of interest.
3. Advanced exercise: how to supply prior information about long survival times in parametric survival models.