Bayesian regression models

Outline

Linear regression models from a Bayesian perspective (45 min)

- Building models, principles for defining priors
- Implementing regression models in JAGS

Practical exercises, questions 1 and 2: Multiple linear regression - demonstration and exercises (45 min, inc. break)

More general forms of Bayesian regression (30 min)

- Binary outcomes: logistic regression
- Non-linear and more general regression. Model building advice.

Continuing practical exercises. Question 3: Logistic regression (60 min, inc. break)

Outline

Linear regression models from a Bayesian perspective (45 min)

- Building models, principles for defining priors
- Implementing regression models in JAGS

Practical exercises, questions 1 and 2: Multiple linear regression - demonstration and exercises (45 min, inc. break)

More general forms of Bayesian regression (30 min)

- Binary outcomes: logistic regression
- Non-linear and more general regression. Model building advice.

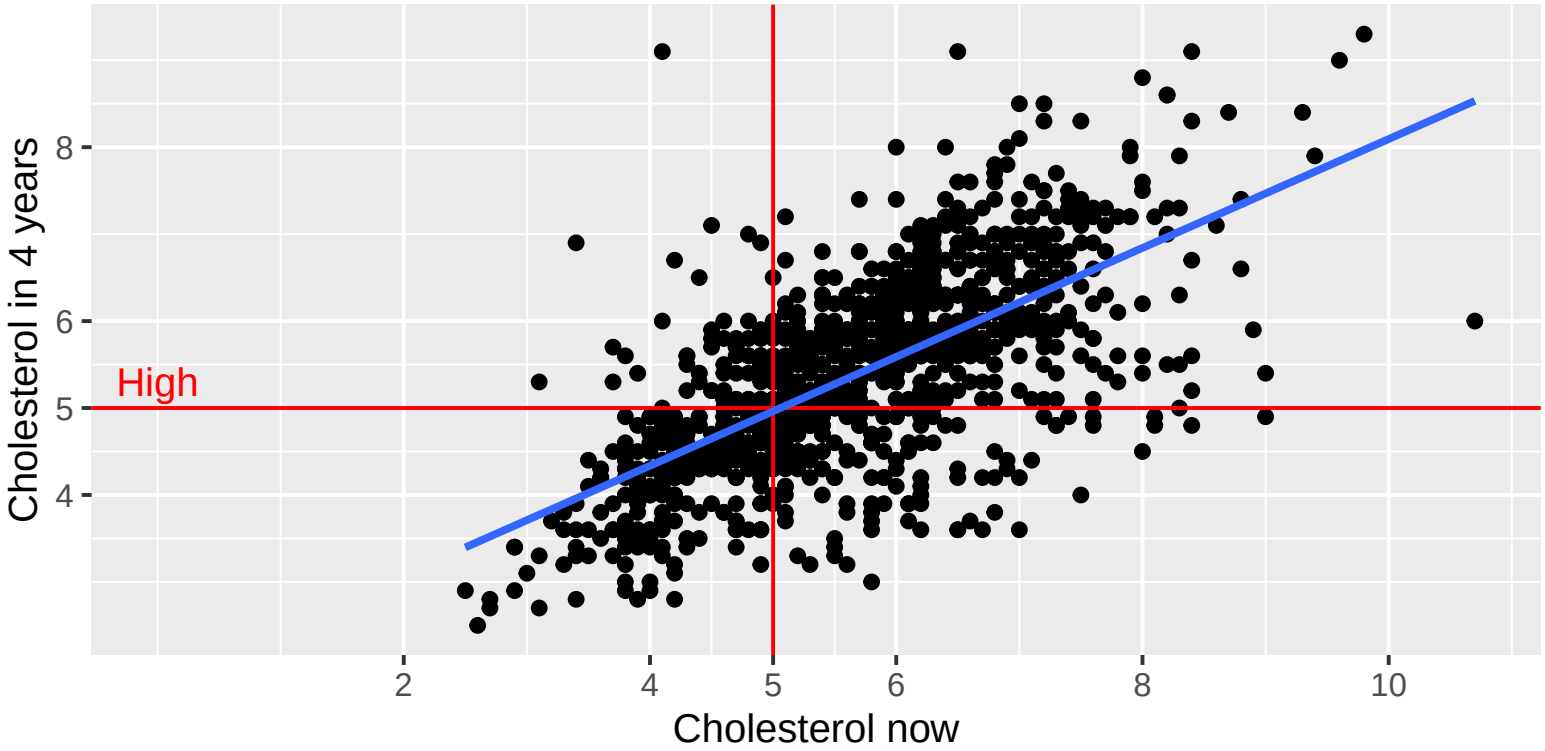
Continuing practical exercises. Question 3: Logistic regression (60 min, inc. break)

Bayesian linear regression

Linear regression

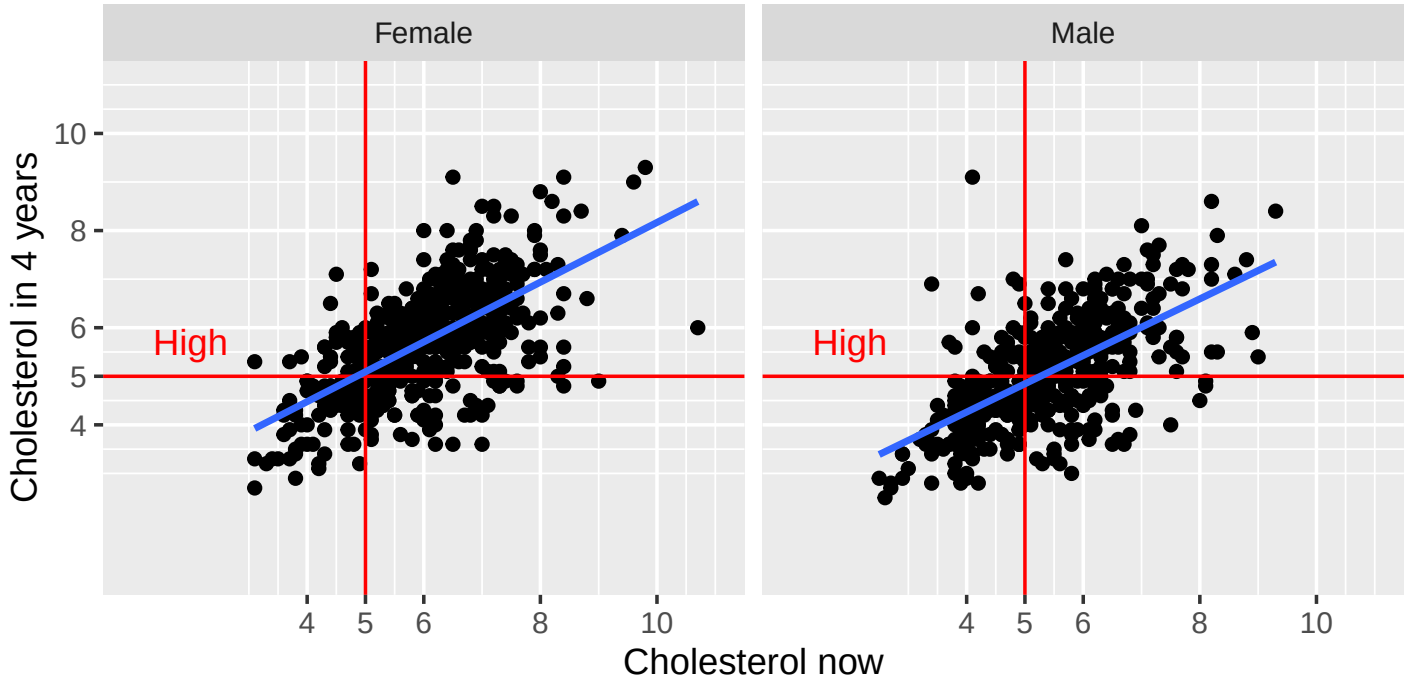
What will my cholesterol level be in 4 years' time (y)?

Predict this based on my cholesterol value now (x).
Prediction model built from data:



$$y_i = \alpha + \beta x_i + \epsilon_i, \quad E(\epsilon_i) = 0$$

Linear regression model with multiple variables



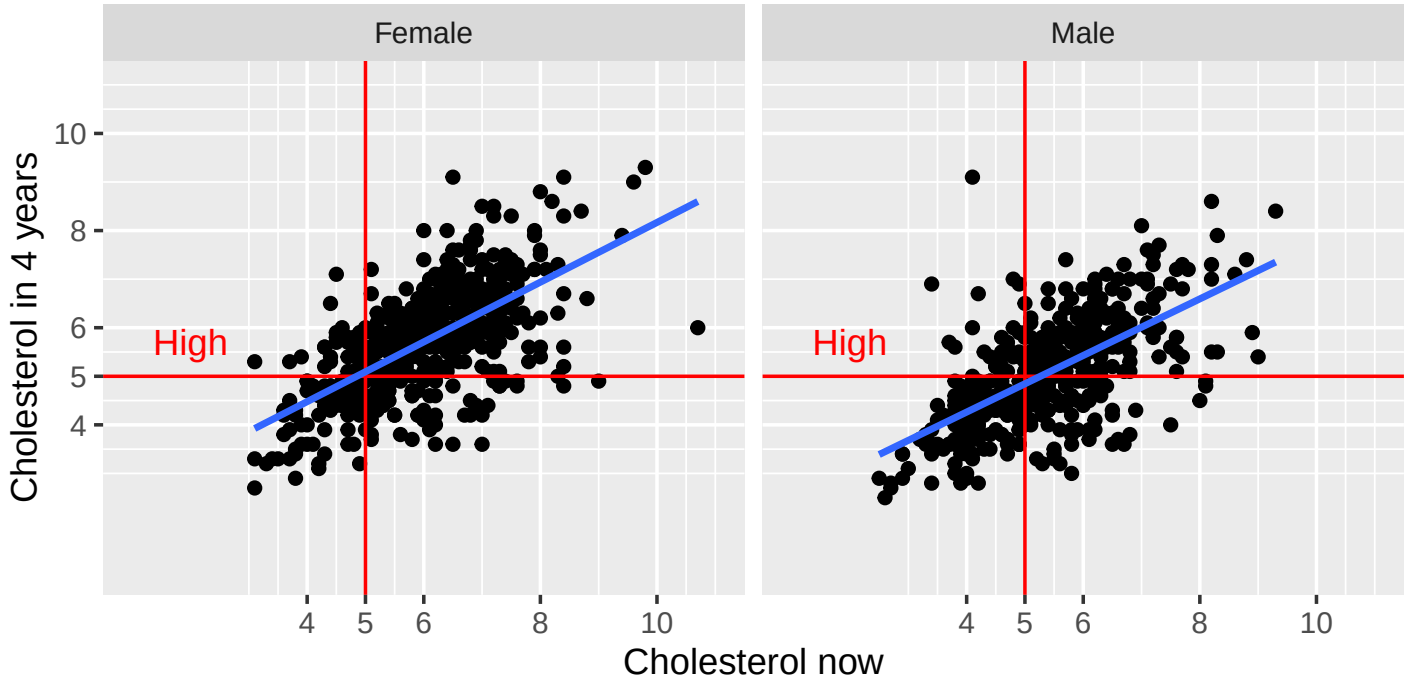
Sex *male* as binary covariate.

$$y_i = \alpha + \beta_1 x_i + \beta_2 male_i + \epsilon_i$$

Or with an interaction between *male* and x:

$$y_i = \alpha + \beta_1 x_i + \beta_2 male_i + \beta_3 x_i male_i + \epsilon_i$$

Linear regression model with multiple variables



Sex *male* as binary covariate.

$$y_i = \alpha + \beta_1 x_i + \beta_2 male_i + \epsilon_i$$

Or with an interaction between *male* and x:

$$y_i = \alpha + \beta_1 x_i + \beta_2 male_i + \beta_3 x_i male_i + \epsilon_i$$

Assumptions of regression models

Model is suitable to answer the scientific question!

- sample relevant to population, no excessive extrapolation outside data, any causal inferences made appropriately...

Structure of the model represents the data

- Relation of expected outcome to predictor: linear?
  - could have log-transformed the outcome or predictor
  - could have fitted a curve, e.g. quadratic
- With multiple covariates: are there interactions?
- Errors, are they independent and/or identically-distributed?

see, e.g. Gelman, Hill and Vehtari "Regression and Other Stories"

https://avehtari.github.io/ROS-Examples

Assumptions of regression models

Model is suitable to answer the scientific question!

- sample relevant to population, no excessive extrapolation outside data, any causal inferences made appropriately...

Structure of the model represents the data

- Relation of expected outcome to predictor: linear?
  - could have log-transformed the outcome or predictor
  - could have fitted a curve, e.g. quadratic
- With multiple covariates: are there interactions?
- Errors, are they independent and/or identically-distributed?

see, e.g. Gelman, Hill and Vehtari "Regression and Other Stories"

https://avehtari.github.io/ROS-Examples

Assumptions of regression models

Model is suitable to answer the scientific question!

- sample relevant to population, no excessive extrapolation outside data, any causal inferences made appropriately...

Structure of the model represents the data

- Relation of expected outcome to predictor: linear?
  - could have log-transformed the outcome or predictor
  - could have fitted a curve, e.g. quadratic
- With multiple covariates: are there interactions?
- Errors, are they independent and/or identically-distributed?

see, e.g. Gelman, Hill and Vehtari "Regression and Other Stories"

https://avehtari.github.io/ROS-Examples

Assumptions of regression models

Model is suitable to answer the scientific question!

- sample relevant to population, no excessive extrapolation outside data, any causal inferences made appropriately...

Structure of the model represents the data

- Relation of expected outcome to predictor: linear?
  - could have log-transformed the outcome or predictor
  - could have fitted a curve, e.g. quadratic
- With multiple covariates: are there interactions?
- Errors, are they independent and/or identically-distributed?

see, e.g. Gelman, Hill and Vehtari "Regression and Other Stories"

https://avehtari.github.io/ROS-Examples

Bayesian regression: additional assumptions

Full probability model required, e.g. a specific distribution on the error term.

$$y_i = \alpha + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

Doesn't have to be Normal (e.g. could use t-distribution, heavier-tailed to accomodate outliers)

In a non-Bayesian setting, normally-distributed errors required to interpret "least squares" estimates as "maximum likelihood" estimates

Prior information required about parameters

- beyond the information in the data itself

In the example of predicting cholesterol levels, we roughly know what is a plausible range of cholesterol levels for people.

We can use this information to bound the parameters within reasonable ranges

Bayesian regression: additional assumptions

Full probability model required, e.g. a specific distribution on the error term.

$$y_i = \alpha + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

Doesn't have to be Normal (e.g. could use t-distribution, heavier-tailed to accomodate outliers)

In a non-Bayesian setting, normally-distributed errors required to interpret "least squares" estimates as "maximum likelihood" estimates

Prior information required about parameters

- beyond the information in the data itself

In the example of predicting cholesterol levels, we roughly know what is a plausible range of cholesterol levels for people.

We can use this information to bound the parameters within reasonable ranges

Principles for choosing priors: (a) Vague priors

Uniform or near uniform, e.g. Normal(mean 0, SD 100000)

- Uniform for α does not imply a uniform prior on $\log(\alpha)$!
  - Priors best placed on easily interpretable quantities
- Make sure SD high enough to include all plausible values
- Priors like Uniform$(-\infty, \infty)$ are improper (don't integrate to 1) — only valid to use if the resulting posteriors are proper.

"Reference" priors

Defined mathematically in some way to be minimally informative.

- Many ways to do this from theoretical literature (e.g. "Jeffreys" priors) though hard to specify except for simplest models
- Example: in a linear regression — Jeffreys priors are uniform on α, β and $\log \sigma$

Principles for choosing priors: (a) Vague priors

Uniform or near uniform, e.g. Normal(mean 0, SD 100000)

- Uniform for α does not imply a uniform prior on $\log(\alpha)$!
  - Priors best placed on easily interpretable quantities
- Make sure SD high enough to include all plausible values
- Priors like Uniform$(-\infty, \infty)$ are improper (don't integrate to 1) — only valid to use if the resulting posteriors are proper.

"Reference" priors

Defined mathematically in some way to be minimally informative.

- Many ways to do this from theoretical literature (e.g. "Jeffreys" priors) though hard to specify except for simplest models
- Example: in a linear regression — Jeffreys priors are uniform on α, β and $\log \sigma$

Principles for choosing priors: (a) Vague priors

Uniform or near uniform, e.g. Normal(mean 0, SD 100000)

- Uniform for α does not imply a uniform prior on $\log(\alpha)$!
  - Priors best placed on easily interpretable quantities
- Make sure SD high enough to include all plausible values
- Priors like Uniform$(-\infty, \infty)$ are improper (don't integrate to 1) — only valid to use if the resulting posteriors are proper.

"Reference" priors

Defined mathematically in some way to be minimally informative.

- Many ways to do this from theoretical literature (e.g. "Jeffreys" priors) though hard to specify except for simplest models
- Example: in a linear regression — Jeffreys priors are uniform on α, β and $\log \sigma$

Principles for choosing priors: (b) Substantive

Informative prior

Based on knowledge that exists outside the data
e.g. obtained by expert elicitation

Weakly informative prior

"Regularises" parameter estimates, i.e.

- Rules out unreasonable parameter values
  e.g. different order of magnitude
- Doesn't rule out values that might make sense
- Facilitates computation, and robustifies inferences, especially when not much data

If different reasonable choices of prior might make a difference to the results, they should be made transparent and subjected to sensitivity analysis, and need for more data highlighted

Principles for choosing priors: (b) Substantive

Informative prior

Based on knowledge that exists outside the data
e.g. obtained by expert elicitation

Weakly informative prior

"Regularises" parameter estimates, i.e.

- Rules out unreasonable parameter values
  e.g. different order of magnitude
- Doesn't rule out values that might make sense
- Facilitates computation, and robustifies inferences, especially when not much data

If different reasonable choices of prior might make a difference to the results, they should be made transparent and subjected to sensitivity analysis, and need for more data highlighted

Principles for choosing priors: (b) Substantive

Informative prior

Based on knowledge that exists outside the data
e.g. obtained by expert elicitation

Weakly informative prior

"Regularises" parameter estimates, i.e.

- Rules out unreasonable parameter values
  e.g. different order of magnitude
- Doesn't rule out values that might make sense
- Facilitates computation, and robustifies inferences, especially when not much data

If different reasonable choices of prior might make a difference to the results, they should be made transparent and subjected to sensitivity analysis, and need for more data highlighted

Priors for linear regression: intercept α

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

α is the expected outcome for a person with $x_i = 0$, which may not happen in practice (e.g. if x is cholesterol level)

Reparameterise the regression by centering around some interesting "baseline" value, e.g. the mean $\bar{x}$.

$$y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

Then define a prior for the expected outcome α with $x = \bar{x}$.

Centering covariates in regression also helps with computation using MCMC (reduces posterior correlation between coefficients, see Session 2)

Weakly informative prior for an expected value

Assuming normal prior $\alpha \sim N(\alpha_0, \sigma_0^2)$.

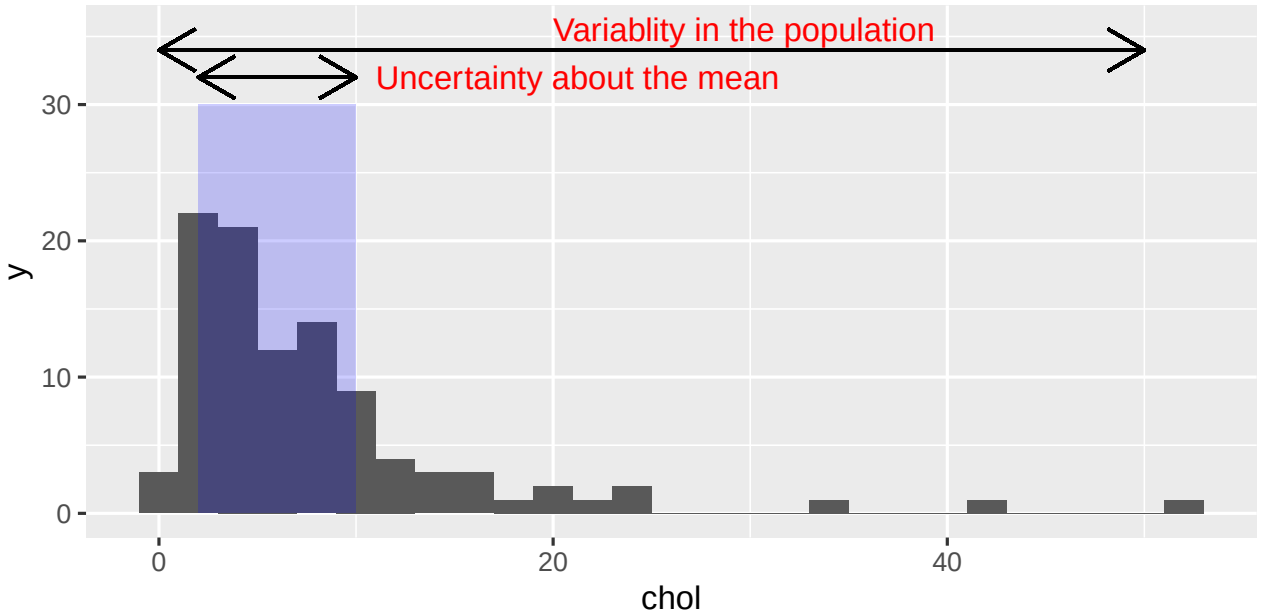
- Set prior mean or median based on knowledge of typical values of the outcome, e.g. $\alpha_0 = 5$
- Then suppose we think it is highly unlikely that the expected outcome α is greater than $u = 100$
  - Deduce the SD σ_0 to satisfy, e.g. $P(\alpha > u) = 0.001$
  - $\Phi((u - \alpha_0)/\sigma_0) = 0.999, \qquad \sigma_0 = (u - \alpha_0)/\Phi^{-1}(0.999)$
    In R, Φ is `pnorm`, and Φ^{-1} is `qnorm`

In the cholesterol example, we may want to log transform the outcome, giving $\alpha_0 = \log(5), u = \log(100)$, say.

Prior for mean is not the population distribution

Caution: Should the prior on α describe the distribution of cholesterol values we would expect to see among the population?

No - it's our belief about the average cholesterol value in our population! So be careful not to make the prior too vague.

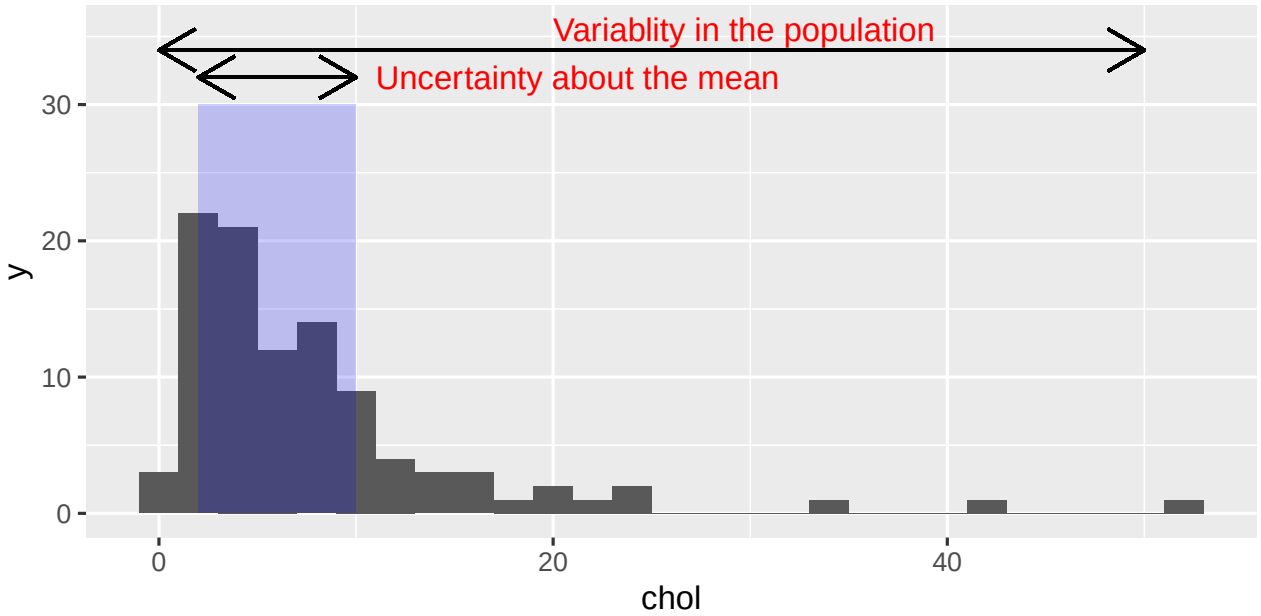


What population? General population, or participants in a study of hypercholesterolaemia? Range of ages? (cholesterol increases with age)?

Prior for mean is not the population distribution

Caution: Should the prior on α describe the distribution of cholesterol values we would expect to see among the population?

No - it's our belief about the average cholesterol value in our population! So be careful not to make the prior too vague.



What population? General population, or participants in a study of hypercholesterolaemia? Range of ages? (cholesterol increases with age)?

Priors for linear regression coefficients β

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Vague: Improper flat prior $p(\beta) = 1$, or near-uniform $\beta \sim N(0, [\text{big variance}])$?

- Posterior mode is the maximum likelihood estimate.
- OK if enough data.
  Though safer in general to rule out implausible values

Informative: Note that the coefficient β is $E(y|x = 1) - E(y|x = 0)$, expected change in y given a unit change in x

- So could define prior $p(\beta) \sim N(0, \sigma_\beta^2)$ to rule out extreme changes in y for a typical change in x.

Weakly informative prior on regression coefficients

Example heuristic

- Typical change in x might be defined as $SD(x)$
- Extreme changes in y defined in terms of $SD(y)$, e.g. 5 standard deviations
- e.g. $p(-5SD(y) < \beta SD(x) < 5SD(y)) = 0.95$
  "Strong belief that y won't change more than ± 5 SDs, given a typical change in x"
- Leads to $p(\beta) \sim N(0, \sigma_\beta^2)$ with $\sigma_\beta = 2.5SD(y)/SD(x)$
- $SD(y)$ and $SD(x)$ from our knowledge of typical scale of x and y.

Priors for linear regression: error variance

Exact choice of prior on error variance σ^2 doesn't typically affect inferences about the coefficients.
Data will usually help to estimate σ^2 well.

Vague priors

- $1/\sigma^2 \sim Gamma(\epsilon, \epsilon)$ for very small ϵ (Jeffreys' prior)
- Gelman et al. *Bayesian Data Analysis* book uses uniform prior on $\log(\sigma)$

Weakly informative prior, on an interpretable scale, centred around a prior guess $SD(y)$ for σ

- e.g. `rstanarm` package uses
  $\sigma \sim Exponential(\text{rate} = 1/SD(y))$
  mean $SD(y)$, 99% credible interval of (0.005, 5)*$SD(y)$

See later session on hierarchical models for examples where the prior on a variance matters more

Implementing Bayesian regression models

With conjugate priors (normal on α, β, Gamma on $\tau = 1/\sigma^2$) the joint posterior (given data Y) for a normal linear regression is closed form:

- $\tau|Y \sim \chi^2()$, $\beta|\tau, Y \sim \textit{Normal}()$. See, e.g. Gelman et al. ch. 14
- We could do inference / prediction by summarising / simulating from those distributions.

With more general outcome models and priors, the posterior doesn't have a nice form. MCMC or posterior approximations needed.

Many software options, e.g:

- R packages that can fit many kinds of "built-in" Bayesian regression models: e.g. `rstanarm`, `brms`, `INLA`
- Software that requires you to define the model algebraically in the program's language (e.g. BUGS, JAGS, Stan, NIMBLE)

Here we use MCMC, via JAGS and its R interface

Implementing Bayesian regression models

With conjugate priors (normal on α, β, Gamma on $\tau = 1/\sigma^2$) the joint posterior (given data Y) for a normal linear regression is closed form:

- $\tau|Y \sim \chi^2()$, $\beta|\tau, Y \sim \textit{Normal}()$. See, e.g. Gelman et al. ch. 14
- We could do inference / prediction by summarising / simulating from those distributions.

With more general outcome models and priors, the posterior doesn't have a nice form. MCMC or posterior approximations needed.

Many software options, e.g:

- R packages that can fit many kinds of "built-in" Bayesian regression models: e.g. `rstanarm`, `brms`, `INLA`
- Software that requires you to define the model algebraically in the program's language (e.g. BUGS, JAGS, Stan, NIMBLE)

Here we use MCMC, via JAGS and its R interface

Linear regression in JAGS

Observed outcome is the log cholesterol for person i, called `logchol[i]`. Normal distribution `dnorm` with mean `mu[i]`, a linear function of the previous log cholesterol `logprevchol[i]`).

Center covariate around its mean - required for convergence in this example!

Data will be supplied and MCMC sampling run through the `rjags` R package (see demonstration in the practical material)

```
model {
  for (i in 1:n) {
    logchol[i] ~ dnorm(mu[i], prec) # precision, not variance.
    mu[i] <- alpha + beta*(logprevchol[i] - mean(logprevchol[]))
    logchol_pred[i] ~ dnorm(mu[i], prec) # predicted value
    chol_mu[i] <- exp(mu[i])
    chol_pred[i] <- exp(logchol_pred[i]) # on non-log scale
  }
  alpha ~ dnorm(0, 1/100^2)
  beta ~ dnorm(0, 1/100^2)
  prec ~ dgamma(0.0001, 0.0001)
  sd <- 1/sqrt(prec)
}
```

Linear regression in JAGS (rjags interface)

Practical session 1 includes a demonstration of how to define, run and summarise a linear regression model in JAGS. Note in particular:

- Defining the model in JAGS code
- Providing data to JAGS - should be a list (not a data frame) and should include constants (e.g. the number of observations) needed in the JAGS code, as well as the data
- Providing initial values of parameters for MCMC simulation (recall Session 2)
- Running JAGS: "burn-in" and checking convergence (recall Session 2)
- Summarising the samples from the posterior for quantities of interest

Summarising linear regression models

Often interested in regression coefficients (for descriptive comparisons, or causal inference)

How to interpret the posterior distribution for coefficients β?

	`2.5%`	`50%`	`97.5%`
beta	0.629	0.675	0.722

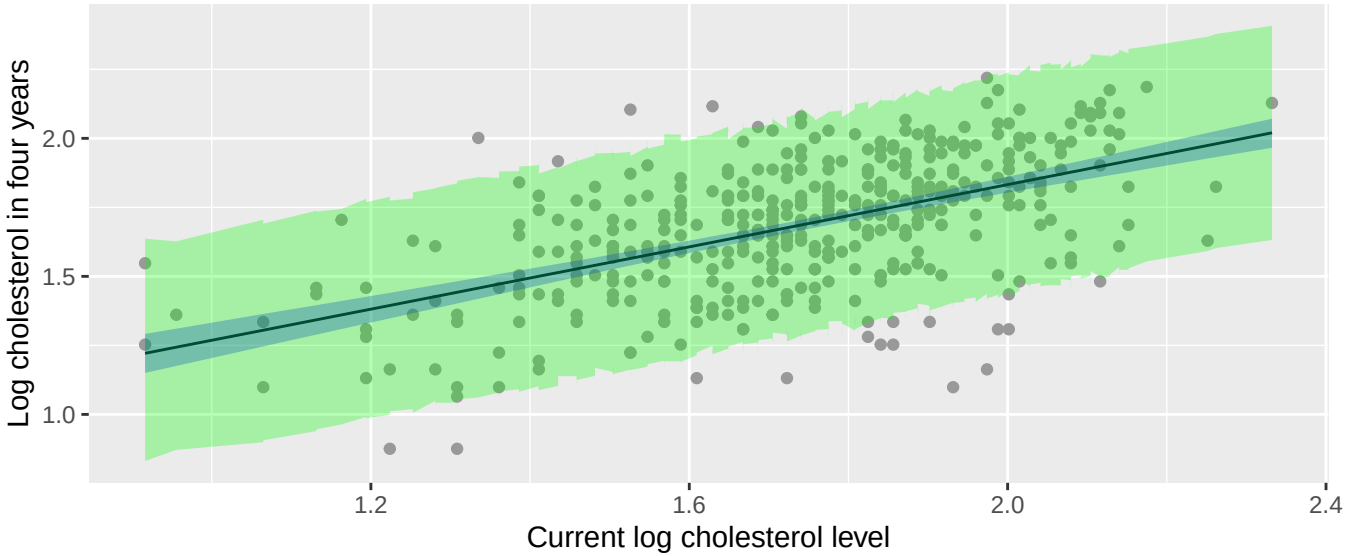
Difference in average log cholesterol in 4 years time, between two groups who differ by one unit in current log cholesterol

- is estimated to be 0.675 (posterior median)
- is between 0.63 and 0.72 with 95% probability (a credible interval)
- a statement of belief about β, unlike a confidence interval

Could also compute e.g. $P(\beta > \beta_0)$, probability of exceeding some practically important value say $\beta_0 = 0$

- a statement of belief about β, unlike a p-value

Summarising linear regression: expectation vs prediction

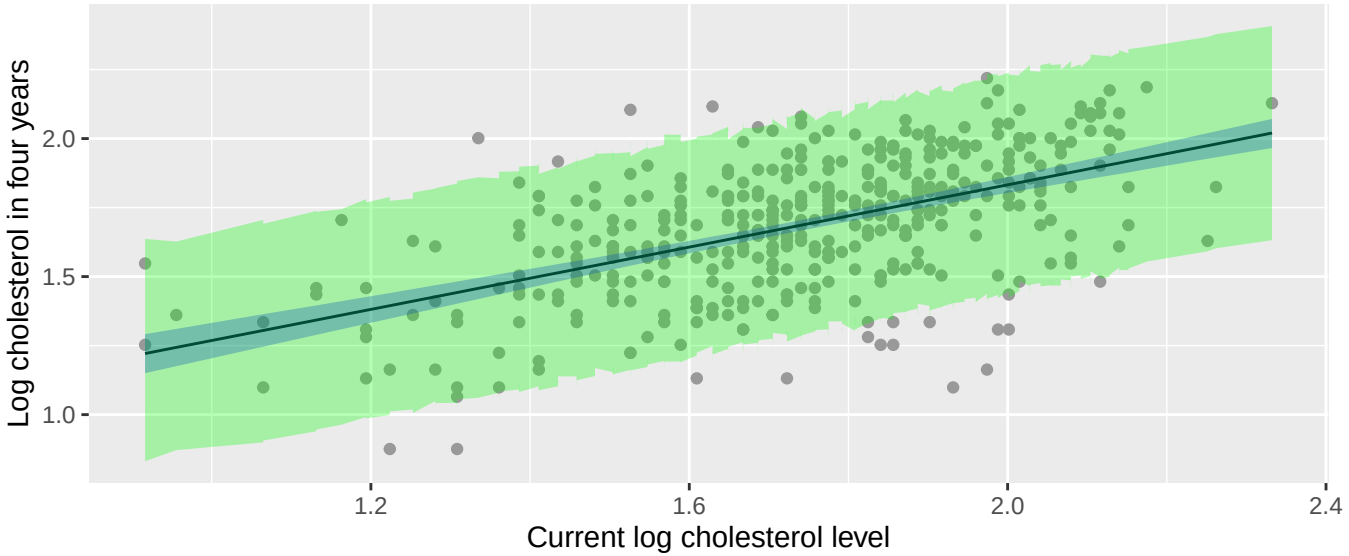


$$y_i = \mu_i + \epsilon_i, \qquad \mu_i = \alpha + \beta x_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

(a) Posterior credible interval for expected outcome μ_i shown as narrow blue band. Represents uncertainty about

- the average value of the outcome y (log cholesterol in 4 years)...
- ... in a population who all have predictor value x (current log cholesterol)

Summarising linear regression: expectation vs prediction



$$y_i = \mu_i + \epsilon_i, \qquad \mu_i = \alpha + \beta x_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

(b) Posterior credible interval for $\mu_i + \epsilon_i$ — prediction of the outcome for a new individual — shown as wide green band.

Represents uncertainty about the observed value of y for an individual with predictor x.

Wider than interval on μ_i — accounts for variability between people, as well as uncertainty about model parameters.

Practical: Multiple linear regression

Run through the linear regression demo (Question 1) and make sure it all makes sense.

Then tackle Question (2), which extends the regression to include a second covariate.

General topics:

- Defining priors in linear regression
- Editing JAGS code: from one predictor to multiple predictors
- Running JAGS and summarising posterior distributions of quantities of interest

Variable selection in regression models?

Any need to exclude variables that are not "significant" predictors?

- as in classical "stepwise" procedures, sequentially testing hypotheses that each coefficient is zero, or using AIC?

If model is for explanation / description?

- Full posterior describes beliefs given data, model and background information, may not need to exclude variables

Variable selection more useful if we need precise prediction, or in high-dimensional situations

- Specialised priors available, which can "shrink" effects of weak predictors towards zero (as in e.g. ridge or lasso regression)
- Model comparison e.g. based on deviance / cross-validation methods: see next lecture

Variable selection in regression models?

Any need to exclude variables that are not "significant" predictors?

- as in classical "stepwise" procedures, sequentially testing hypotheses that each coefficient is zero, or using AIC?

If model is for explanation / description?

- Full posterior describes beliefs given data, model and background information, may not need to exclude variables

Variable selection more useful if we need precise prediction, or in high-dimensional situations

- Specialised priors available, which can "shrink" effects of weak predictors towards zero (as in e.g. ridge or lasso regression)
- Model comparison e.g. based on deviance / cross-validation methods: see next lecture

Variable selection in regression models?

Any need to exclude variables that are not "significant" predictors?

- as in classical "stepwise" procedures, sequentially testing hypotheses that each coefficient is zero, or using AIC?

If model is for explanation / description?

- Full posterior describes beliefs given data, model and background information, may not need to exclude variables

Variable selection more useful if we need precise prediction, or in high-dimensional situations

- Specialised priors available, which can "shrink" effects of weak predictors towards zero (as in e.g. ridge or lasso regression)
- Model comparison e.g. based on deviance / cross-validation methods: see next lecture

Logistic and other generalised regression models

Logistic regression: example

Outcome: diabetes in next 4 years.
Predictors: current age, gender, cholesterol, blood pressure, smoking etc.

```
##  diabetes2 male age  tc  sbp   dbp [etc.]
## 1   0        1  65  4.2  172.5 95.0  ...
## 2   0        1  52  5.3  125.5 78.0
## 3   0        1  73  5.3  163.0 78.5
## 4   0        1  61  3.9  124.5 77.5
## 5   0        0  59  6.0  123.5 72.5
## 6   0        0  68  4.5  106.0 62.5
[etc.] ...
```

Logistic regression: the probability $p(\mathbf{x})$ of the outcome, given a vector of covariates $\mathbf{x}$ is

$$\log(p(\mathbf{x})/(1-p(\mathbf{x}))) = \alpha + \beta'\mathbf{x}$$

Simple to estimate by maximum likelihood, in standard software

Bayesian approach facilities using external information

- e.g. as prior, extending to hierarchical models, missing data, multiple sources of data ...

Logistic regression: prior distributions (intercept α)

Sensible again to recenter x around a value of interest, e.g. the mean $\bar{\mathbf{x}}$.

$$\log(p(\mathbf{x})/(1-p(\mathbf{x}))) = \alpha + \beta'(\mathbf{x} - \bar{\mathbf{x}})$$

Then α is the log odds of the outcome for $\mathbf{x} = \bar{\mathbf{x}}$.

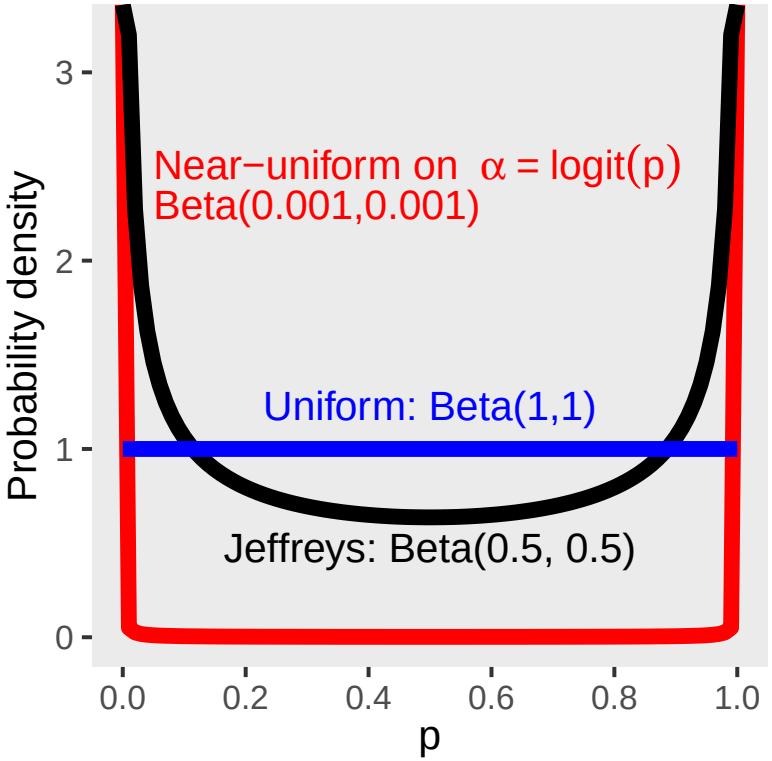
Informative priors: Put prior on interpretable quantity: $p(\bar{\mathbf{x}}) = exp(\alpha)/(1 + exp(\alpha))$ instead of α.

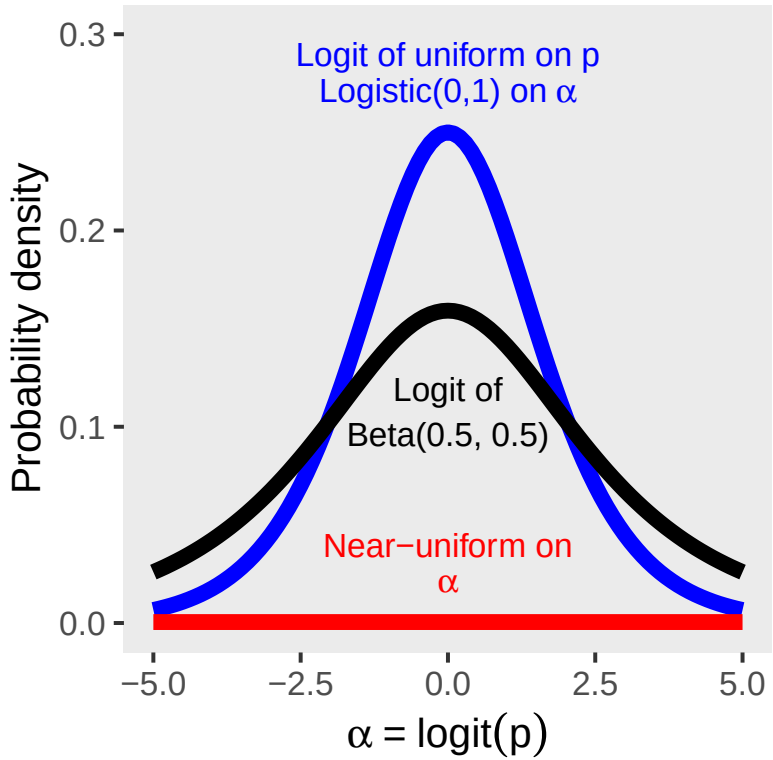
Recall lecture 1 for principles of defining Beta priors for probabilities, based on substantive beliefs (e.g. mean and credible intervals).

Uniform prior on p equivalent to standard logistic distribution for α.

Vague priors for probabilities

No single definition of a vague prior — depends on the scale it is defined on





- Uniform for $p = \exp(\alpha)/(1 + \exp(\alpha))$ is not uniform for log odds α, and vice-versa.
- Jeffreys' prior (based on information theory, see textbooks) is Beta(0.5, 0.5), a compromise.

Logistic regression: prior for coefficients β

e.g. prior on one β of the coefficients (log odds ratios) $\boldsymbol{\beta}$ in

$$\log(p(\mathbf{x})/(1-p(\mathbf{x}))) = \alpha + \boldsymbol{\beta}'(\mathbf{x} - \bar{\mathbf{x}})$$

As in linear regression, define a belief for change in p given unit change (or typical change) in x

Example: Is $\beta = 5$ or higher plausible?

- A change in 5 in $logit(p)$ gives an odds ratio of 148
- a change in the probability from $p = expit(0) = 0.01$ to $p = expit(5) = 0.5$, or from 0.5 to 0.99

Extreme in epidemiology! e.g. odds ratio for smoking and lung cancer is 40 {http://www.ncbi.nlm.nih.gov/pubmed/11700268}

Weakly informative prior e.g. $\beta \sim N(0, 2.5^2)$ gives low probability to this size of change given one unit change in corresponding x

Logistic regression: prior for coefficients β

e.g. prior on one β of the coefficients (log odds ratios) $\boldsymbol{\beta}$ in

$$\log(p(\mathbf{x})/(1-p(\mathbf{x}))) = \alpha + \boldsymbol{\beta}'(\mathbf{x} - \bar{\mathbf{x}})$$

As in linear regression, define a belief for change in p given unit change (or typical change) in x

Example: Is $\beta = 5$ or higher plausible?

- A change in 5 in $logit(p)$ gives an odds ratio of 148
- a change in the probability from $p = expit(0) = 0.01$ to $p = expit(5) = 0.5$, or from 0.5 to 0.99

Extreme in epidemiology! e.g. odds ratio for smoking and lung cancer is 40 {http://www.ncbi.nlm.nih.gov/pubmed/11700268}

Weakly informative prior e.g. $\beta \sim N(0, 2.5^2)$ gives low probability to this size of change given one unit change in corresponding x

JAGS implementation of logistic regression

Code the outcome y_i for each i as 0 or 1, and model y_i as a Bernoulli random variable with probability p_i, e.g.

```
for (i in 1:n){
  y[i] ~ dbern(p[i])
  logit(p[i]) <-  alpha + beta*x[i]
}
```

JAGS allows `logit` to be used on the left-hand side of a definition as a 'link function', mimicking how the model would be written mathematically

See Practical question (3) for a full example

Regression models for count data

Count data are common in epidemiology, e.g. y_i number of disease cases by area i, number of hospital admissions per month i...

Poisson distribution $y_i \sim Pois(\mu_i)$: basis for a wide range of models

- Simple Poisson regression $\log(\mu_i) = \alpha + \beta x_i$ on a covariate x_i
- Negative Binomial mean μ_i, dispersion parameter r
  Use when data are overdispersed: variance $>$ mean
  Tends to Poisson(μ_i) as $r \to \infty$.
- Zero-inflated and similar mixture models.
  - Uncertain proportion p of individuals have zero risk of events.
  - Remaining proportion $1 - p$ of individuals follow, e.g $Pois(\mu)$
  - Jointly estimate p and μ.
  - Extensible to regression models on both p and μ

Hierarchical models (see later): can represent more general sources of variation.

Regression models for count data

Count data are common in epidemiology, e.g. y_i number of disease cases by area i, number of hospital admissions per month i...

Poisson distribution $y_i \sim Pois(\mu_i)$: basis for a wide range of models

- Simple Poisson regression $\log(\mu_i) = \alpha + \beta x_i$ on a covariate x_i
- Negative Binomial mean μ_i, dispersion parameter r
  Use when data are overdispersed: variance $>$ mean
  Tends to Poisson(μ_i) as $r \to \infty$.
- Zero-inflated and similar mixture models.
  - Uncertain proportion p of individuals have zero risk of events.
  - Remaining proportion $1 - p$ of individuals follow, e.g $Pois(\mu)$
  - Jointly estimate p and μ.
  - Extensible to regression models on both p and μ

Hierarchical models (see later): can represent more general sources of variation.

Regression models for count data

Count data are common in epidemiology, e.g. y_i number of disease cases by area i, number of hospital admissions per month i...

Poisson distribution $y_i \sim Pois(\mu_i)$: basis for a wide range of models

- Simple Poisson regression $\log(\mu_i) = \alpha + \beta x_i$ on a covariate x_i
- Negative Binomial mean μ_i, dispersion parameter r
  Use when data are overdispersed: variance $>$ mean
  Tends to Poisson(μ_i) as $r \to \infty$.
- Zero-inflated and similar mixture models.
  - Uncertain proportion p of individuals have zero risk of events.
  - Remaining proportion $1 - p$ of individuals follow, e.g $Pois(\mu)$
  - Jointly estimate p and μ.
  - Extensible to regression models on both p and μ

Hierarchical models (see later): can represent more general sources of variation.

Regression models for count data

Count data are common in epidemiology, e.g. y_i number of disease cases by area i, number of hospital admissions per month i...

Poisson distribution $y_i \sim Pois(\mu_i)$: basis for a wide range of models

- Simple Poisson regression $\log(\mu_i) = \alpha + \beta x_i$ on a covariate x_i
- Negative Binomial mean μ_i, dispersion parameter r
  Use when data are overdispersed: variance $>$ mean
  Tends to Poisson(μ_i) as $r \to \infty$.
- Zero-inflated and similar mixture models.
  - Uncertain proportion p of individuals have zero risk of events.
  - Remaining proportion $1 - p$ of individuals follow, e.g $Pois(\mu)$
  - Jointly estimate p and μ.
  - Extensible to regression models on both p and μ

Hierarchical models (see later): can represent more general sources of variation.

JAGS implementation of Poisson regression

Again, suppose y_i is a count outcome (non-negative integer). The `log` link function can also be used in JAGS model definitions as follows:

```
for (i in 1:n) {
  y[i]          ~ dpois(mu[i])
  log(mu[i]) <- alpha + beta*x[i]
}
```

See the optional practical sheet for a full example and extension to negative binomial and zero-inflated models.

Probabilistic programming makes these (and further hierarchical) extensions of Poisson models straightforward to implement.

Non-linear regression

Used when scientific mechanism assumed for how one variable changes with another, e.g. from a differential equation

Example: "von Bertalanffy" growth curve: size at age a obeys

$$dL/da = k(L(\infty) - L(a))$$

Slower growth when closer to fully-grown size $L(\infty)$. Solved by

$$L(a) = \mu_\infty - \mu_{diff}\gamma^a$$

Here we have parameterised in a way that the parameters have interpretations (not always easy to do!).

- $\mu_\infty = L(\infty)$ is final size
- $\mu_{diff} = L(\infty) - L(0)$ is amount grown in lifetime
- γ is the proportion of final size still to grow at age $a = 1$

Makes it easier to define informative priors

Non-linear regression

Used when scientific mechanism assumed for how one variable changes with another, e.g. from a differential equation

Example: "von Bertalanffy" growth curve: size at age a obeys

$$dL/da = k(L(\infty) - L(a))$$

Slower growth when closer to fully-grown size $L(\infty)$. Solved by

$$L(a) = \mu_\infty - \mu_{diff}\gamma^a$$

Here we have parameterised in a way that the parameters have interpretations (not always easy to do!).

- $\mu_\infty = L(\infty)$ is final size
- $\mu_{diff} = L(\infty) - L(0)$ is amount grown in lifetime
- γ is the proportion of final size still to grow at age $a = 1$

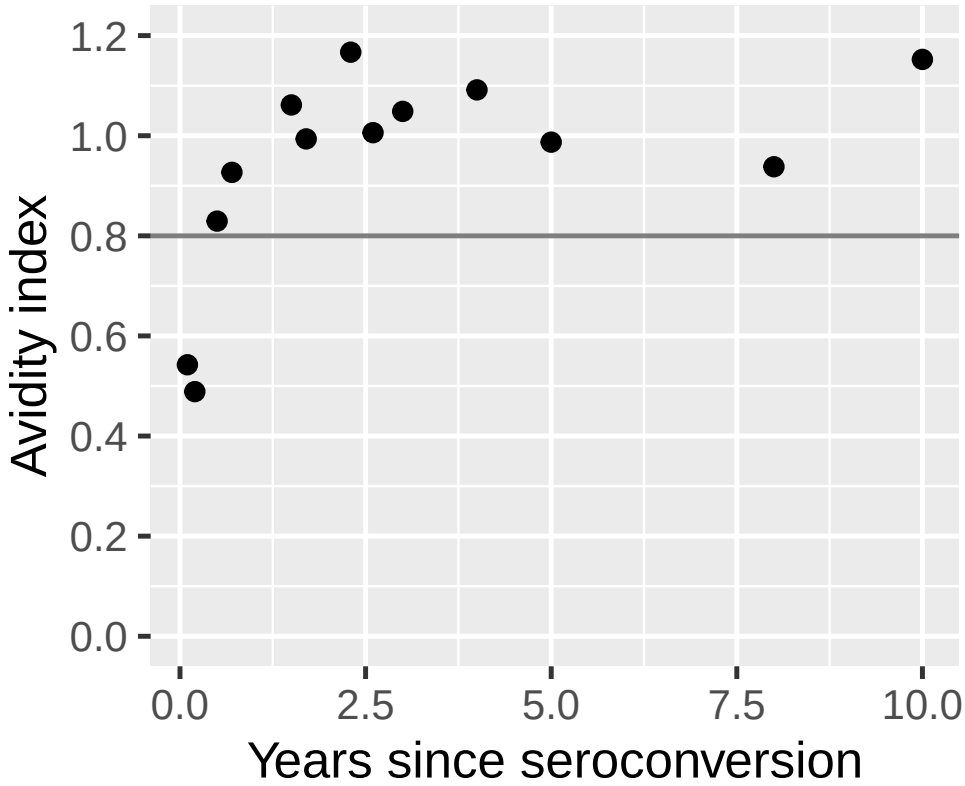
Makes it easier to define informative priors

Non-linear regression: detecting recent HIV infections

"Avidity index": immunological biomarker y_i, measured at a series of times. Classify as a "recent infection" if $y_i < 0.8$

$$y_i \sim N(\mu_i, \sigma^2) \text{ with } \mu_i = \mu_\infty - \mu_{diff}\gamma^{a_i}$$

Estimate growth curve parameters $\mu_\infty, \mu_{diff}, \gamma$ from data.
Error σ^2 represents variance between individuals i.

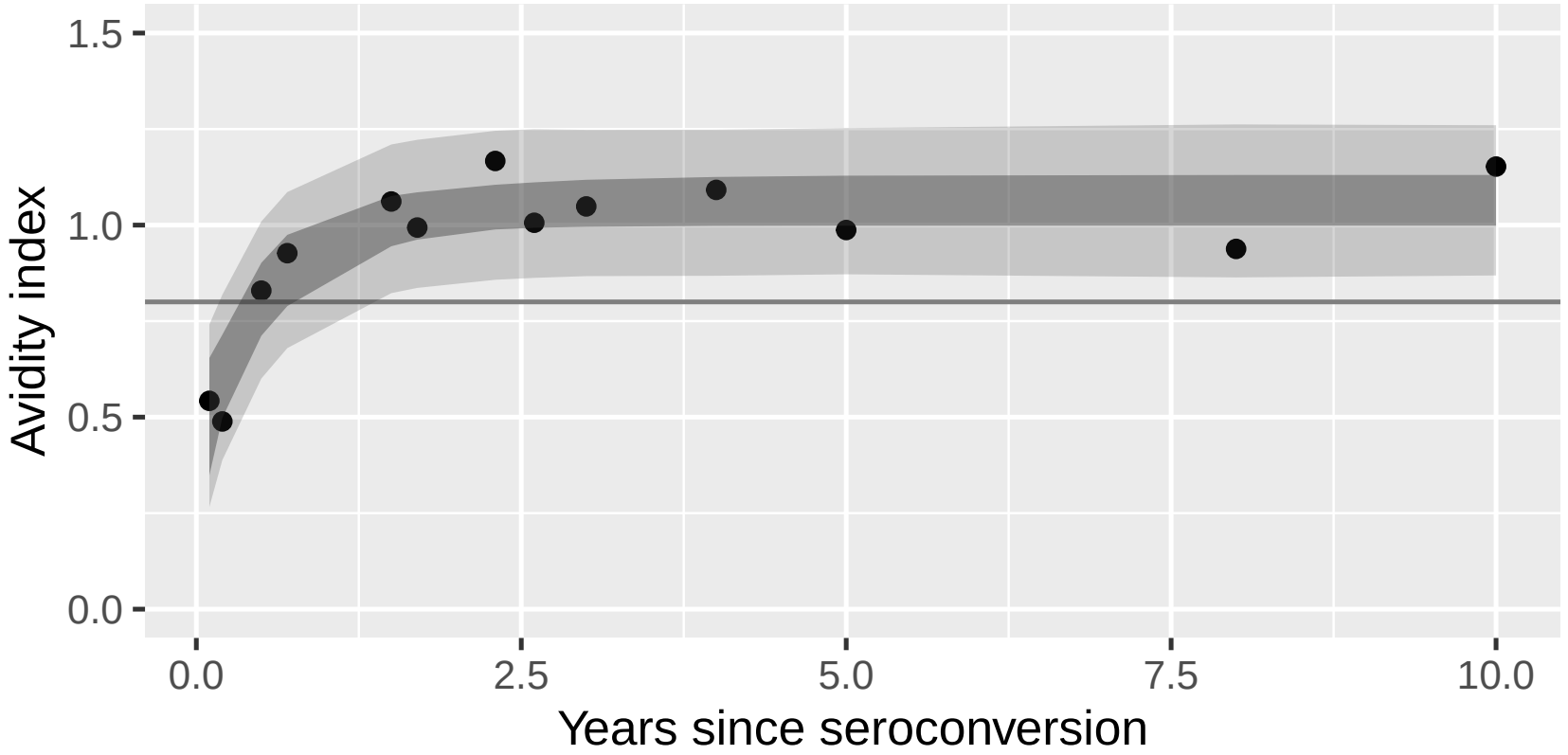


Priors for asymptote μ_∞, change μ_{diff}
Log-normal (since positive)
Center around "typical" value, e.g. 1
Determine variance such that prob of extreme value (> 5 say) is small

Prior for proportion to grow, γ
Uniform (or beta if had substantive information)

Non-linear regression: JAGS implementation

Underlying immune response at time i taken to equal the mean μ_i. Posterior 95% credible intervals for this:



Time zero ("seroconversion") defined here as midpoint beteen last negative and first positive HIV test. Bayesian approach can be extended to account for uncertainty about seroconversion time, and variability between individuals.

Any class of regression models can be made Bayesian

- Count data
  - Poisson and extensions: see optional practical sheet for examples
  - Multinomial models for categorical data (not covered here)
- Survival analysis (see Session 6)
- Nonlinear regression models (see in a bit)
- arbitrarily flexible regression relationships (e.g. nonlinear "spline" functions, interactions, non-constant errors) ...

Many implemented in R packages (e.g. `rstanarm`, `brms`, `INLA`).

Any Bayesian model can be implemented in JAGS (or similar) by writing out the model definition in code.

- more flexible, e.g. with multiple sources of data, complex hierarchical structures, missing data, ...

Considerations for Bayesian regression modelling

Very general models can be specified in modern software, but...

- bigger models may not be identifiable with less data
- models may not be computationally feasible with big datasets
- with more parameters, harder to specify substantive priors
  - spend more effort on those that are more influential / important

Starting small, and building up helps with implementation / computation

- See later for model comparison, selection and checking

General trick for choosing priors in any complex model

Simulate from the (joint) prior distribution, then

- transform parameters to quantities that can be interpreted
- simulate observable data from the prior predictive distribution

then check if the simulations match with beliefs, if not, revise and repeat.

Example in Session 6 (Weibull survival models)

Practical session

Question (3) : Logistic regression

- Posterior distributions for different model summaries of interest (odds ratio and risk difference)

Optional sheet for extra practice: Poisson regression

- Extending to negative binomial and zero-inflated models: implementation in JAGS and comparison

Looking forward to next sessions

Many other sessions will build on regression models and principles here, e.g.

- Session 4: checking and comparing models with different assumptions
- Session 5: extending to deal with hierarchical data (multiple measurements from individuals, areas...)
- Session 6: missing data and censored survival data

Probabilistic programming software like JAGS makes these all accessible in practice

Principles for thoughtful model building, choice of priors, checking, relevant throughout