Module 4: Bayesian Methods
Lecture 5: Linear regression

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Regression models

How does an outcome Y vary as a function of $\mathbf{x} = \{x_1, \dots, x_p\}$?

- Which x_j's have an effect?
- What are the effect sizes?
- Can we predict Y as a function of $\mathbf{x}$?

These questions can be assessed via a **regression model** $p(y|\mathbf{x})$.

Regression data

Parameters in a regression model can be estimated from data:

$$\begin{pmatrix} y_1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

These data are often expressed in matrix/vector form:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

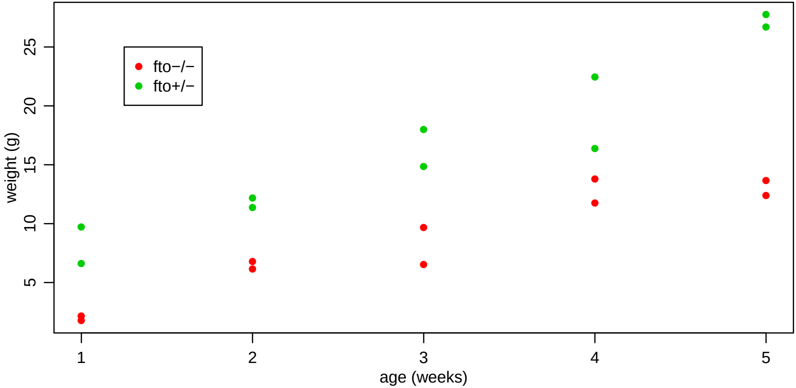
FTO experiment

FTO gene is hypothesized to be involved in growth and obesity.

Experimental design:

- 10 *fto* $+/-$ mice
- 10 *fto* $-/-$ mice
- Mice are sacrificed at the end of 1-5 weeks of age.
- Two mice in each group are sacrificed at each age.

FTO Data



Data analysis

- y = weight
- x_g = fto heterozygote $\in \{0, 1\}$ = number of "+" alleles
- x_a = age in weeks $\in \{1, 2, 3, 4, 5\}$

How can we estimate $p(y|x_g, x_a)$

Cell means model:

genotype	*age*				
$-/-$	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	$\theta_{0,4}$	$\theta_{0,5}$
$+/-$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$	$\theta_{1,4}$	$\theta_{1,5}$

Problem: Only two observations per cell.

Linear regression

Solution: Assume smoothness as a function of age. For each group,

$$y = \alpha_0 + \alpha_1 x_a + \epsilon$$

This is a *linear regression model*.
Linearity means "linear in the parameters".

We could also try the model

$$y = \alpha_0 + \alpha_1 x_a + \alpha_2 x_a^2 + \alpha_3 x_a^3 + \epsilon,$$

which is also a linear regression model.

Multiple linear regression

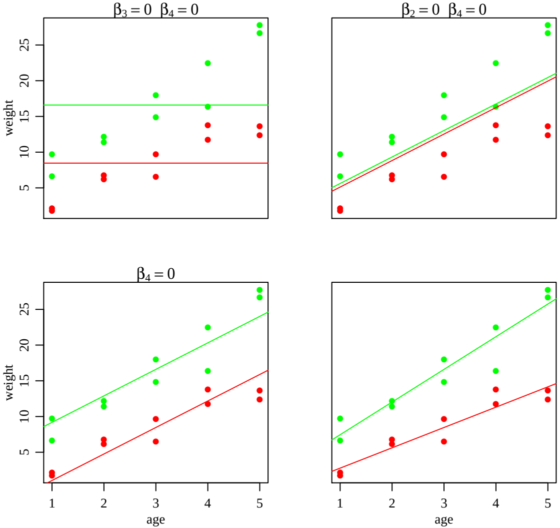
We can estimate the regressions for both groups simultaneously:

$$
\begin{aligned}
Y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i \ , \text{ where} \\
x_{i,1} &= 1 \text{ for each subject } i \\
x_{i,2} &= 0 \text{ if subject } i \text{ is homozygous, 1 if heterozygous} \\
x_{i,3} &= \text{age of subject } i \\
x_{i,4} &= x_{i,2} \times x_{i,3}
\end{aligned}
$$

Under this model,

$$
\begin{aligned}
\mathrm{E}[Y|\mathbf{x}] &= \beta_1 + \beta_3 \times \mathrm{age} \ \text{ if } x_2 = 0, \text{ and} \\
\mathrm{E}[Y|\mathbf{x}] &= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \mathrm{age} \ \text{ if } x_2 = 1.
\end{aligned}
$$

Multiple linear regression



Normal linear regression

How does each Y_i vary around $\mathrm{E}[Y_i|\boldsymbol{\beta}, \mathbf{x}_i]$?

Assumption of normal errors:

$$
\begin{aligned}
\epsilon_1, \ldots, \epsilon_n &\sim \text{i.i.d. normal}(0, \sigma^2) \\
Y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i \,.
\end{aligned}
$$

This completely specifies the probability density of the data:

$$
\begin{aligned}
& p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) \\
&= \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \\
&= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2\}.
\end{aligned}
$$

Matrix form

- Let $\mathbf{y}$ be the n-dimensional column vector $(y_1, \ldots, y_n)^T$;
- Let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i$.

Then the normal regression model is that

$$\{\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \text{ multivariate normal } (\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}_1 \rightarrow \\ \mathbf{x}_2 \rightarrow \\ \vdots \\ \mathbf{x}_n \rightarrow \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} \mathrm{E}[Y_1|\boldsymbol{\beta}, \mathbf{x}_1] \\ \vdots \\ \mathrm{E}[Y_n|\boldsymbol{\beta}, \mathbf{x}_n] \end{pmatrix}.$$

Ordinary least squares estimation

What values of $\boldsymbol{\beta}$ are consistent with our data $\mathbf{y}, \mathbf{X}$?

Recall

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2\}.$$

This is big when $\mathrm{SSR}(\boldsymbol{\beta}) = \sum (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2$ is small.

$$\begin{aligned} \mathrm{SSR}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta}\,. \end{aligned}$$

What value of $\boldsymbol{\beta}$ makes this the smallest?

Calculus

Recall from calculus that

1. a minimum of a function $g(z)$ occurs at a value z such that $\frac{d}{dz}g(z) = 0$;
2. the derivative of $g(z) = az$ is a and the derivative of $g(z) = bz^2$ is $2bz$.

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}}\text{SSR}(\boldsymbol{\beta}) &= \frac{d}{d\boldsymbol{\beta}}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) \\ &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\ ,\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}}\text{SSR}(\boldsymbol{\beta}) = 0 \quad &\Leftrightarrow \quad -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0 \\ &\Leftrightarrow \quad \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \\ &\Leftrightarrow \quad \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,.\end{aligned}$$

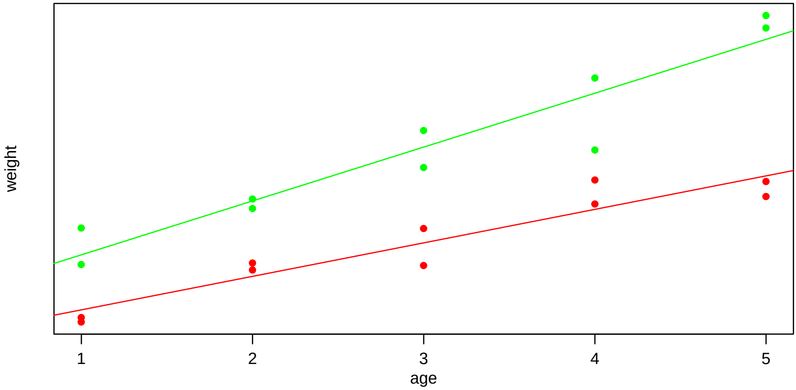
$\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the *OLS estimator* of $\boldsymbol{\beta}$.

OLS estimation in R

```
### OLS estimate
beta.ols<- solve( t(X)%*%X )%*%t(X)%*%y

c(beta.ols)

## [1] -0.06822  2.94485  2.84421  1.72948
```

```
### using lm
fit.ols<-lm(y~ X[,2] + X[,3] +X[,4] )

summary(fit.ols)$coef

##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -0.06822     1.4223 -0.04796 9.623e-01
## X[, 2]       2.94485     2.0114  1.46406 1.625e-01
## X[, 3]       2.84421     0.4288  6.63235 5.761e-06
## X[, 4]       1.72948     0.6065  2.85171 1.154e-02
```

OLS estimation




```
summary(fit.ols)$coef

##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -0.06822     1.4223 -0.04796 9.623e-01
## X[, 2]       2.94485     2.0114  1.46406 1.625e-01
## X[, 3]       2.84421     0.4288  6.63235 5.761e-06
## X[, 4]       1.72948     0.6065  2.85171 1.154e-02
```

Bayesian inference for regression models

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

Motivation:

- Posterior probability statements: $\Pr(\beta_j > 0 | \mathbf{y}, \mathbf{X})$
- OLS tends to overfit when p is large, Bayes more conservative.
- Model selection and averaging

Prior and posterior distribution

$$
\begin{array}{llcl}
\text{prior} & \boldsymbol{\beta} & \sim & \text{mvn}(\boldsymbol{\beta}_0, \Sigma_0) \\
\text{sampling model} & \mathbf{y} & \sim & \text{mvn}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \\
\text{posterior} & \boldsymbol{\beta}|\mathbf{y},\mathbf{X} & \sim & \text{mvn}(\boldsymbol{\beta}_n, \Sigma_n)
\end{array}
$$

where

$$
\begin{aligned}
\Sigma_n = \mathrm{Var}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] &= (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1} \\
\boldsymbol{\beta}_n = \mathrm{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] &= (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{y}/\sigma^2).
\end{aligned}
$$

Notice:

- If $\Sigma_0^{-1} \ll \mathbf{X}^T\mathbf{X}/\sigma^2$, then $\boldsymbol{\beta}_n \approx \hat{\boldsymbol{\beta}}_{\text{ols}}$
- If $\Sigma_0^{-1} \gg \mathbf{X}^T\mathbf{X}/\sigma^2$, then $\boldsymbol{\beta}_n \approx \boldsymbol{\beta}_0$

The g-prior

How to pick $\boldsymbol{\beta}_0, \Sigma_0$?

g-prior:

$$\boldsymbol{\beta} \sim \text{mvn}(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Idea: The variance of the OLS estimate $\hat{\boldsymbol{\beta}}_{\text{ols}}$ is

$$\text{Var}[\hat{\boldsymbol{\beta}}_{\text{ols}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{n}(\mathbf{X}^T\mathbf{X}/n)^{-1}$$

This is roughly the uncertainty in $\boldsymbol{\beta}$ from n observations.

$$\text{Var}[\boldsymbol{\beta}]_{\text{gprior}} = g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{n/g}(\mathbf{X}^T\mathbf{X}/n)^{-1}$$

The g-prior can roughly be viewed as the uncertainty from n/g observations.

For example, $g = n$ means the prior has the same amount of info as 1 obs.

Posterior distributions under the g-prior

$$\{\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2\} \sim \text{mvn}(\boldsymbol{\beta}_n, \boldsymbol{\Sigma}_n)$$

$$\begin{aligned} \Sigma_n = \mathrm{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\ \boldsymbol{\beta}_n = \mathrm{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

Notes:

- The posterior mean estimate $\boldsymbol{\beta}_n$ is simply $\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}}$.
- The posterior variance of $\boldsymbol{\beta}$ is simply $\frac{g}{g+1}\mathrm{Var}[\hat{\boldsymbol{\beta}}_{\text{ols}}]$.
- g shrinks the coefficients and can prevent overfitting to the data
- If $g = n$, then as n increases, inference approximates that using $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

Monte Carlo simulation

What about the error variance σ^2?

$$
\begin{array}{lrcl}
\text{prior} & 1/\sigma^2 & \sim & \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\
\text{sampling model} & \mathbf{y} & \sim & \text{mvn}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \\
\text{posterior} & 1/\sigma^2|\mathbf{y}, \mathbf{X} & \sim & \text{gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2)
\end{array}
$$

where SSR_g is somewhat complicated.

Simulating the joint posterior distribution:

$$
\begin{array}{lrcl}
\text{joint distribution} & p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) & = & p(\sigma^2|\mathbf{y}, \mathbf{X}) \times p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2) \\
\text{simulation} & \{\sigma^2, \boldsymbol{\beta}\} \sim p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) & \Leftrightarrow & \sigma^2 \sim p(\sigma^2|\mathbf{y}, \mathbf{X}), \boldsymbol{\beta} \sim p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2)
\end{array}
$$

To simulate $\{\sigma^2, \boldsymbol{\beta}\} \sim p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$,

1. First simulate σ^2 from $p(\sigma^2|\mathbf{y}, \mathbf{X})$
2. Use this σ^2 to simulate $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2)$

Repeat 1000's of times to obtain MC samples: $\{\sigma^2, \boldsymbol{\beta}\}^{(1)}, \ldots, \{\sigma^2, \boldsymbol{\beta}\}^{(S)}$.

FTO example

Priors:

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(1/2, 3.6781/2) \\ \boldsymbol{\beta}|\sigma^2 &\sim \text{mvn}(\mathbf{0}, g \times \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

Posteriors:

$$\begin{aligned} \{1/\sigma^2|\mathbf{y}, \mathbf{X}\} &\sim \text{gamma}((1+20)/2, (3.6781 + 251.7753)/2) \\ \{\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \sigma^2\} &\sim \text{mvn}(.952 \times \hat{\boldsymbol{\beta}}_{\text{ols}}, .952 \times \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

where

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.55 & -0.55 & -0.15 & 0.15 \\ -0.55 & 1.10 & 0.15 & -0.30 \\ -0.15 & 0.15 & 0.05 & -0.05 \\ 0.15 & -0.30 & -0.05 & 0.10 \end{pmatrix} \quad \hat{\boldsymbol{\beta}}_{\text{ols}} = \begin{pmatrix} -0.0682 \\ 2.9449 \\ 2.8442 \\ 1.7295 \end{pmatrix}$$

R-code

```
##  data dimensions
n<-dim(X)[1] ; p<-dim(X)[2]

## prior parameters
nu0<-1
s20<-summary(lm(y~-1+X))$sigma^2
g<-n

## posterior calculations
Hg<- (g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X)
SSRg<- t(y)%*%( diag(1,nrow=n)  - Hg ) %*%y

Vbeta<- g*solve(t(X)%*%X)/(g+1)
Ebeta<- Vbeta%*%t(X)%*%y

## simulate sigma^2  and beta
s2.post<-beta.post<-NULL
for(s in 1:5000)
{
  s2.post<-c(s2.post,1/rgamma(1, (nu0+n)/2, (nu0*s20+SSRg)/2 ) )
  beta.post<-rbind(beta.post, rmvnorm(1,Ebeta,s2.post[s]*Vbeta))
}
```

MC approximation to posterior

```
s2.post[1:5]

## [1]  9.737 13.002 15.284 14.528 14.818
```

```
beta.post[1:5,]

##        [,1]    [,2]  [,3]  [,4]
## [1,]  1.701  1.2066 1.649 2.841
## [2,] -1.868  1.2554 3.216 1.975
## [3,]  1.032  1.5555 1.909 2.338
## [4,]  3.351 -1.3819 2.401 2.364
## [5,]  1.486 -0.6652 2.032 2.977
```

MC approximation to posterior

```
quantile(s2.post,probs=c(.025,.5,.975))

##   2.5%    50%  97.5%
##  7.163 12.554 24.774

quantile(sqrt(s2.post),probs=c(.025,.5,.975))

##  2.5%   50% 97.5%
## 2.676 3.543 4.977
```

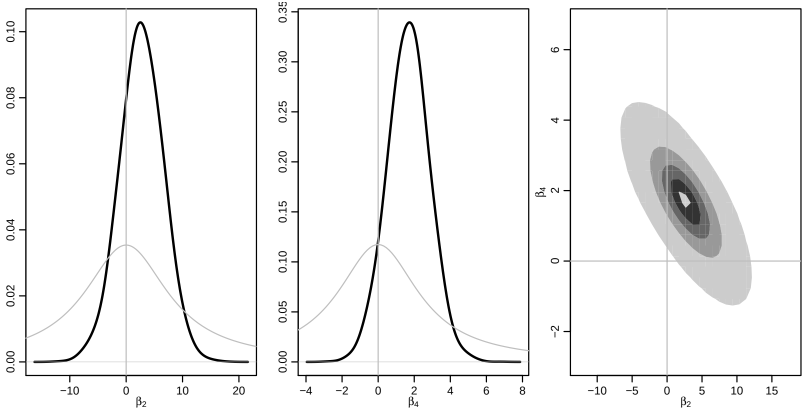
```
apply(beta.post,2,quantile,probs=c(.025,.5,.975))

##           [,1]   [,2]  [,3]    [,4]
## 2.5%  -5.26996 -4.840 1.065 -0.5929
## 50%   -0.01051  2.698 2.678  1.6786
## 97.5%  5.20650  9.992 4.270  3.9071
```

OLS/Bayes comparison

```
apply(beta.post,2,mean)

## [1] 0.0133 2.7080 2.6796 1.6736

apply(beta.post,2,sd)

## [1] 2.6637 3.7726 0.8055 1.1429
```

```
summary(fit.ols)$coef

##      Estimate Std. Error  t value  Pr(>|t|)
## X   -0.06822     1.4223 -0.04796 9.623e-01
## Xxg  2.94485     2.0114  1.46406 1.625e-01
## Xxa  2.84421     0.4288  6.63235 5.761e-06
## X    1.72948     0.6065  2.85171 1.154e-02
```

Posterior distributions



Summarizing the genetic effect

$$
\begin{aligned}
\text{Genetic effect} &= \mathrm{E}[y|age, +/-] - \mathrm{E}[y|age, -/-] \\
&= [(\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age}] - [\beta_1 + \beta_3 \times \text{age}] \\
&= \beta_2 + \beta_4 \times \text{age}
\end{aligned}
$$