Module 17: Bayesian Statistics for Genetics
Lecture 4: Linear regression

Ken Rice

Department of Biostatistics
University of Washington

Outline

Regression models

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

- What are the effect sizes?
- What is the effect of x_1, in observations that have the same $x_2, x_3, \ldots x_p$ (a.k.a. "keeping these covariates constant")?
- 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:

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

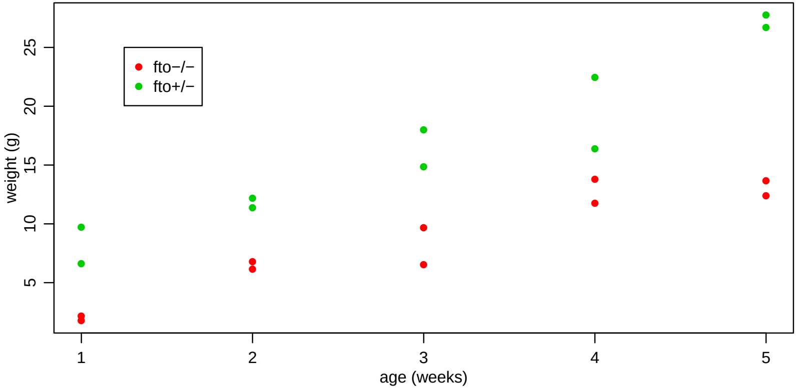
FTO experiment

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 = indicator of 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: 10 parameters – 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", i.e. several covariates multiplied by corresponding α and added.

A more complex model might assume e.g.

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

– but this is still a linear regression model, even with age^2, age^3 terms.

Multiple linear regression

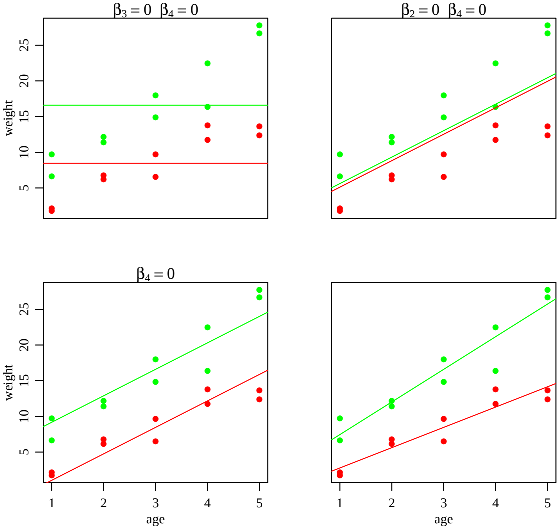
With enough variables, we can describe the regressions for both groups simultaneously:

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

Note that 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 its mean $\mathrm{E}[Y_i|\boldsymbol{\beta}, \mathbf{x}_i]$?

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

This assumption of Normal errors completely specifies the likelihood:

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

Note: in larger sample sizes, analysis is "robust" to the Normality assumption—but we are relying on the mean being linear in the **x**'s, and on the ϵ_i's variance being constant with respect to **x**.

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

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

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 *Ordinary Least Squares (OLS) estimator* of $\boldsymbol{\beta}$.

No Calculus

The calculus-free, algebra-heavy version – which relies on knowing the answer in advance!

Writing $\Pi = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and noting that $\mathbf{X} = \Pi\mathbf{x}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} = \Pi\mathbf{y}$;

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) &= (\mathbf{y} - \Pi\mathbf{y} + \Pi\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \Pi\mathbf{y} + \Pi\mathbf{y} - \mathbf{X}\beta) \\ &= ((I - \Pi)\mathbf{y} + \Pi(\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}))^T((I - \Pi)\mathbf{y} + \Pi(\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta})) \\ &= \mathbf{y}^T(I - \Pi)\mathbf{y} + (\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta})^T\Pi(\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}),\end{aligned}$$

because all the 'cross terms' with Π and $I - \Pi$ are zero.

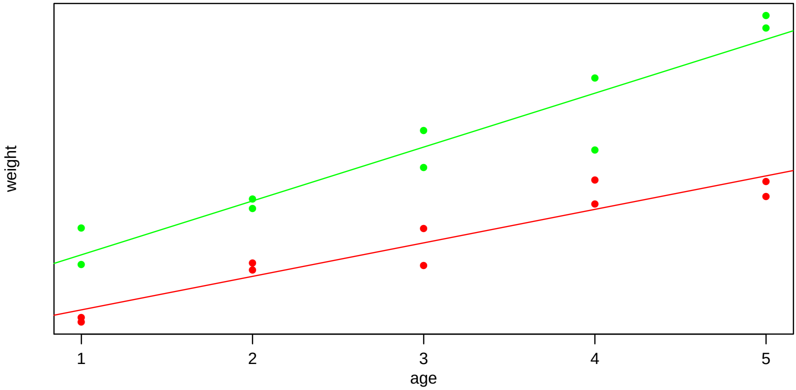
Hence the value of $\boldsymbol{\beta}$ that minimizes the SSR – for a given set of data – is $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

OLS estimation in R

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

c(beta.ols)

## [1] -0.06821632  2.94485495  2.84420729  1.72947648
```

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

summary(fit.ols)$coef

##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) -0.06821632  1.4222970 -0.04796208 9.623401e-01
## X[, 2]       2.94485495  2.0114316  1.46405917 1.625482e-01
## X[, 3]       2.84420729  0.4288387  6.63234803 5.760923e-06
## X[, 4]       1.72947648  0.6064695  2.85171239 1.154001e-02
```

OLS estimation




```
summary(fit.ols)$coef

##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) -0.06821632  1.4222970 -0.04796208 9.623401e-01
## X[, 2]       2.94485495  2.0114316  1.46405917 1.625482e-01
## X[, 3]       2.84420729  0.4288387  6.63234803 5.760923e-06
## X[, 4]       1.72947648  0.6064695  2.85171239 1.154001e-02
```

Bayesian inference for regression models

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

Motivation:

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

Prior and posterior distribution

prior	$\boldsymbol{\beta}$	$\sim$	$\text{mvn}(\boldsymbol{\beta}_0, \Sigma_0)$
sampling model	$\mathbf{y}$	$\sim$	$\text{mvn}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
posterior	$\boldsymbol{\beta}\vert\mathbf{y}, \mathbf{X}$	$\sim$	$\text{mvn}(\boldsymbol{\beta}_n, \Sigma_n)$

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 towards**0** 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:

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(1/2, 3.678/2) \\ \boldsymbol{\beta}|\sigma^2 &\sim \text{mvn}(\mathbf{0}, g \times \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

Posteriors:

$$\begin{aligned} \{1/\sigma^2|\mathbf{y}, \mathbf{X}\} &\sim \text{gamma}((1+20)/2, (3.678+251.775)/2) \\ \{\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \sigma^2\} &\sim \text{mvn}(.952 \times \hat{\boldsymbol{\beta}}_{\text{ols}}, .952 \times \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

where

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.55 & -0.55 & -0.15 & 0.15 \\ -0.55 & 1.10 & 0.15 & -0.30 \\ -0.15 & 0.15 & 0.05 & -0.05 \\ 0.15 & -0.30 & -0.05 & 0.10 \end{pmatrix} \quad \hat{\boldsymbol{\beta}}_{\text{ols}} = \begin{pmatrix} -0.068 \\ 2.945 \\ 2.844 \\ 1.729 \end{pmatrix}$$

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]  9.737351 13.002432 15.283947 14.527585 14.818471
```

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

##            [,1]       [,2]     [,3]     [,4]
## [1,]  1.701434  1.2066217 1.649404 2.840527
## [2,] -1.868185  1.2553571 3.216233 1.974885
## [3,]  1.031936  1.5554807 1.908681 2.337766
## [4,]  3.350976 -1.3819152 2.400596 2.364326
## [5,]  1.485922 -0.6651715 2.032383 2.977433
```

MC approximation to posterior

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

##      2.5%       50%     97.5%
##  7.162945 12.554219 24.773727

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

##     2.5%      50%    97.5%
## 2.676368 3.543193 4.977321
```

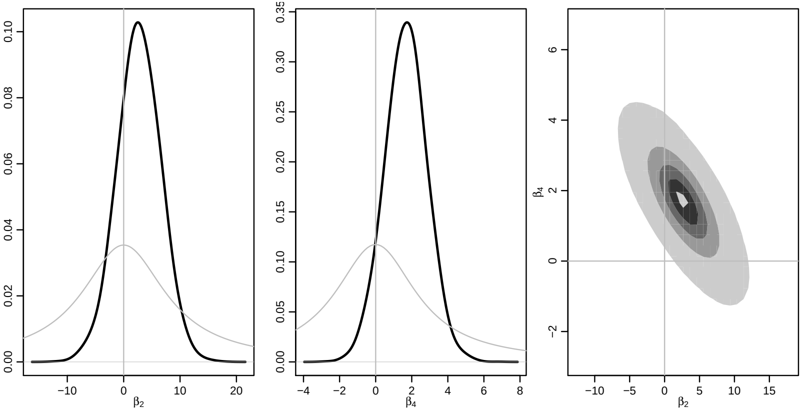
```
apply(beta.post,2,quantile,probs=c(.025,.5,.975))

##              [,1]      [,2]     [,3]       [,4]
## 2.5%  -5.26995978 -4.839650 1.064610 -0.5928799
## 50%   -0.01050552  2.697659 2.677907  1.6786014
## 97.5%  5.20649638  9.992408 4.270029  3.9070770
```

OLS/Bayes comparison

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

## [1] 0.01330163 2.70795309 2.67964282 1.67363273

apply(beta.post,2,sd)

## [1] 2.6637246 3.7725596 0.8054542 1.1429453
```

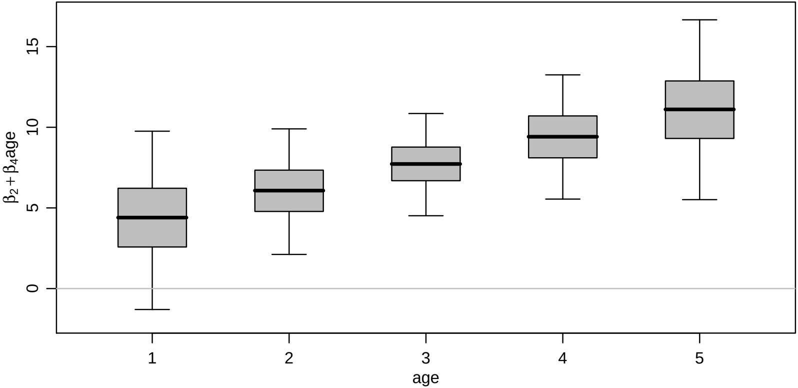
```
summary(fit.ols)$coef

##        Estimate Std. Error     t value     Pr(>|t|)
## X   -0.06821632  1.4222970 -0.04796208 9.623401e-01
## Xxg  2.94485495  2.0114316  1.46405917 1.625482e-01
## Xxa  2.84420729  0.4288387  6.63234803 5.760923e-06
## X    1.72947648  0.6064695  2.85171239 1.154001e-02
```

Posterior distributions



Summarizing the genetic effect

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



What if the model's wrong?

Different types of violation—in decreasing order of how much they typically matter in practice

- Just have the wrong data (!) i.e. not the data you claim to have
- Observations are not independent, e.g. repeated measures on same mouse over time
- Mean model is incorrect
- Error terms do not have constant variance
- Error terms are not Normally distributed

Dependent observations

- Observations from the same mouse are more likely to be similar than those from different mice (even if they have same age and genotype)
- SBP from subjects (even with same age, genotype etc) in the same family are more likely to be similar than those in different familes – perhaps unmeasured common diet?
- Spatial and temporal relationships also tend to induce correlation

If the pattern of relationship is known, can allow for it – typically in "random effects modes" – see later session.
If not, treat results with caution! Precision is likely over-stated.

Wrong mean model

Even when the scientific background is highly informative about the variables of interest (e.g. we want to know about the association of Y with $\mathbf{x}_1$, adjusting for $\mathbf{x}_2$, $\mathbf{x}_3$...) there is rarely strong information about the form of the model

- Does mean weight increase with age? age^2? age^3?
- Could the effect of genotype also change non-linearly with age?

Including quadratic terms is a common approach – but quadratics are sensitive to the tails. Instead, including "spline" representations of covariates allows the model to capture many patterns.

Including interaction terms (as we did with $x_{i,2} \times x_{i,3}$) lets one covariate's effect vary with another.
(Deciding which covariates to use is addressed in the Model Choice session.)

Non-constant variance

This is plausible in many situations; perhaps e.g. young mice are harder to measure, i.e. more variables. Or perhaps the FTO variant affects weight regulation — again, more variance.

- Having different variances at different covariate values is known as *heteroskedasticity*
- Unaddressed, it can result in over- or under-statement of precision

The most obvious approach is to model the variance, i.e.

$$
\begin{aligned}
Y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i, \\
\epsilon_i &\sim \text{Normal}(0, \sigma_i^2),
\end{aligned}
$$

where σ_i depends on covariates, e.g. σ_{homozy} and $\sigma_{heterozy}$ for the two genotypes. Of course, these parameters need priors. Constraining variances to be positive also makes choosing a model difficult in practice.

Robust standard errors (in Bayes)

In linear regression, some robustness to model-misspecification and/or non-constant variance is available – but it relies on interest in linear 'trends'. Formally, we can define parameter

$$\boldsymbol{\theta} = \mathrm{argmin} E_{y,x}\left[\left(E_y[y|x] - \mathbf{x}^t\boldsymbol{\theta}\right)^2\right],$$

i.e. the straight line that best-captures random-sampling, in a least-squares sense.

- This 'trend' can capture important features in how the mean y varies at different x
- Fitting extremely flexible Bayesian models, we get a posterior for $\boldsymbol{\theta}$
- The posterior mean approaches $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, in large samples
- The posterior variance approaches the 'robust' *sandwich estimate*, in large samples (details in Szpiro et al, 2011)

Robust standard errors

The OLS estimator can be written as $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \sum_{i=1}^n c_i y_i$, for appropriate c_i.

$$\begin{array}{rrcl}
\textbf{True}\text{ variance} & \mathrm{Var}[\,\hat{\beta}\,] & = & \sum_{i=1}^n c_i^2 \mathrm{Var}[\,Y_i\,] \\
\textbf{Robust}\text{ estimate} & \widehat{\mathrm{Var}}_R[\,\hat{\beta}\,] & = & \sum_{i=1}^n c_i^2 e_i^2 \\
\textbf{Model-based}\text{ estimate} & \widehat{\mathrm{Var}}_M[\,\hat{\beta}\,] & = & \mathrm{Mean}(e_i^2) \sum_{i=1}^n c_i^2,
\end{array}$$

where $e_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\text{ols}}$, the residuals from fitting a linear model.

Non-Bayesian sandwich estimates are available through R's `sandwich` package – much quicker than Bayes with a very-flexible model. For correlated outcomes, see the `GEE` package for generalizations.

This is not a big problem for learning about population parameters;

- The variance statements/estimates we just saw don't rely on Normality
- The *central limit theorem* means that $\hat{\beta}$ ends up Normal anyway, in large samples
- In small samples, expect to have limited power to detect non-Normality
- ... except, perhaps, for extreme outliers (data errors?)

For prediction – where we assume that outcomes do follow a Normal distibution – this assumption is more important.

Summary

- Linear regressions are of great applied interest
- Corresponding models are easy to fit, particularly with judicious prior choices
- Assumptions are made — but a well-chosen linear regression usually tells us **something** of interest, even if the assumptions are (mildly) incorrect