2014 SISG Module 4: Bayesian Statistics for Genetics
Lecture 8: Generalized Linear Modeling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction

- In this lecture we will discuss Bayesian modeling in the context of Generalized Linear Models (GLMs).
- This discussion will include the addition of random effects, i.e. the class of Generalized Linear Mixed Models (GLMMs).
- Estimation via the quick INLA technique will be demonstrated, along with its R implementation.
- An approximation technique that is useful in the context of Genome Wide Association Studies (GWAS) (in which the number of tests is large) will also be introduced.

Motivating Example I: Logistic Regression

- We consider case-control data for the disease Leber Hereditary Optic Neuropathy (LHON) disease with genotype data for marker rs6767450:

	CC $x = 0$	CT $x = 1$	TT $x = 2$	Total
Cases	6	8	75	89
Controls	10	66	163	239
Total	16	74	238	328

- Let $x = 0, 1, 2$ represent the number of T alleles, and $p(x)$ the probability of being a case, given x copies of the T allele.

Motivating Example I: Logistic Regression

- For such case-control data one may fit the multiplicative odds model:

$$\frac{p(x)}{1-p(x)} = \exp(\alpha) \times \exp(\theta x),$$

  with a binomial likelihood.
- Interpretation:
  - $\exp(\alpha)$ is of little interest given the case-control sampling.
  - $\exp(\theta)$ is the odds ratio describing the multiplicative change in risk for one T allele versus zero T alleles.
  - $\exp(2\theta)$ is the odds ratio describing the multiplicative change in risk for two T alleles versus zero T alleles.
  - Odds ratios approximate the relative risk for a rare disease.

R code for Logistic Regression Estimation via MLE

```
> x <- c(0,1,2)
# Case data for CC CT TT
> y <- c(6,8,75)
# Control data for CC CT TT
> z <- c(10,66,163)
#
# Fit logistic regression model, the default choice for
# binomial data
#
> logitmod <- glm(cbind(y,z)~x,family="binomial")
> thetahat <- logitmod$coeff[2]          # Log odds ratio
thetahat
        x
0.4787428
> exp(thetahat) # Odds ratio
       x       # An extra T allele is associated with an increase
1.614044       # of 61% in risk
> V <- vcov(logitmod)[2,2]               # standard error^2
# Asymptotic confidence interval for odds ratio
> exp(thetahat-1.96*sqrt(V))
       x
0.987916
> exp(thetahat+1.96*sqrt(V))
       x
2.637004
# So 95% interval (just) contains 1.
```

R code for Logistic Regression Hypothesis Testing via a LRT

```
#
# Let's look at a likelihood ratio test of H0: theta=0
#
> logitmod
Call:  glm(formula = cbind(y, z) ~ x, family = "binomial")
Coefficients:
(Intercept)            x
    -1.8077       0.4787
Degrees of Freedom: 2 Total (i.e. Null);  1 Residual
Null Deviance:      15.01
Residual Deviance: 10.99   AIC: 27.79
> pchisq(15.01-10.99,1,lower.tail=F)
[1] 0.04496
# So just significant at the 5% level.
```

- So for these data both estimation and testing point towards borderline significance, at conventional levels.

Motivating Example II: FTO Data Revisited

Linear Model Example

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

We examine the fit of the model

$$E[Y|x_g, x_a] = \beta_0 + \beta_g x_g + \beta_a x_a + \beta_{int} x_g x_a.$$

```
> fto<-
  read.table("http://www.stat.washington.edu/~hoff/SISG/fto_data.txt",
                    header=TRUE)
> liny <- fto$y
> linxg <- fto$xg
> linxa <- fto$xa
> linxint <- fto$xgxa
> ftodf <- list(liny=liny,linxg=linxg,linxa=linxa,linxint=linxint)
> ols.fit <- lm(liny~linxg+linxa+linxint,data=ftodf)
> summary(ols.fit)
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -0.06822    1.42230   -0.048    0.9623
linxg        2.94485    2.01143    1.464    0.1625
linxa        2.84421    0.42884    6.632  5.76e-06 ***
linxint      1.72948    0.60647    2.852    0.0115 *
```

Motivating Example III: RNA Seq with Replicates

- We report an experiment carried out in a collaboration with Caitlin Connelly and Josh Akey (UW Genome Sciences), see Connelly *et al.* (2014) for further details.
- Start with two haploid yeast strains (individuals).
- From these we obtain RNA-Seq data, where we isolate RNA from the two individuals, fragment and sequence it using next-generation sequencing, and map the sequencing reads back to the genome to generate RNA levels in the form of counts of the number of sequencing reads mapping at each gene.
- Also mate the two haploid yeast strains together to form a diploid hybrid. We again isolate RNA, fragment, and sequence it.
- Then take advantage of polymorphisms between the two strains in order to map reads to either of the two haploid individuals, giving us counts for the number of reads mapping to either one of the parental genomes in the diploid hybrid for each gene.

Motivating Example III: RNA Seq with Replicates

- We are interested in two questions from this data. First, we want to look for evidence of trans effects at each gene; in biological terms, this means that polymorphisms located far from the gene are responsible for differences in RNA levels.
- To detect this, look for genes where the difference between RNA levels in the haploids differs from the difference between RNA levels for the two parental strains in the diploid.
- Also interested in looking for cis effects, meaning polymorphisms near the gene itself are responsible for differences in RNA levels. We can detect cis effects as a difference in the count of reads mapping to each of the parental strains in the diploid at a gene.

Motivating Example III: RNA Seq Data, Statistical Model

- There are two replicates and so for each of N genes we obtain two sets of counts.
- For the diploid hybrid let Y_{ij} be the number of A alleles for gene i and replicate j, and N_{ij} is the total number of counts, so that $N_{ij} - Y_{ij}$ is the number of T alleles $j = 1, 2$.
- We fit a hierarchical logistic regression model starting with first stage:

$$Y_{ij}|N_{ij}, p_{ij} \sim \text{binomial}(N_{ij}, p_{ij})$$

  so that p_{ij} is the probability of seeing an A read for gene i and replicate j.
- At the second stage:

$$\text{logit } p_{ij} \quad = \quad \theta_i + \epsilon_{ij}$$

  where $\epsilon_{ij} \sim \text{normal}(0, \sigma^2)$ represent random effects that allow for excess-binomial variation.
- In the model θ_i is a parameter of interest – if a (say) 95% posterior interval estimate contains 0 then we have evidence of cis effects.

Generalized Linear Models

- Generalized Linear Models (GLMs) provide a very useful extension to the linear model class.
- GLMs have three elements:
  1. The responses follow an exponential family.
  2. The mean model is linear in the covariates on some scale.
  3. A link function relates the mean of the data to the covariates.
- In a GLM the response y_i are independently distributed and follow an exponential family[1], $i = 1, \ldots, n$.
- Examples: Normal, Poisson, binomial.

[1] so that the distribution is of the form $p(y_i|\theta_i, \alpha) = \exp(\{y_i\theta_i - b(\theta_i)\}/\alpha + c(y_i, \alpha))$, where θ_i and α are scalars

Generalized Linear Models

- The link function $g(\cdot)$ provides the connection between the mean $\mu = \mathrm{E}[Y]$ and the linear predictor $\mathbf{x}\boldsymbol{\beta}$, via

$$g(\mu) = \mathbf{x}\boldsymbol{\beta},$$

  where $\mathbf{x}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of regression parameters.
- For normal data, the usual link is the identity

$$g(\mu) = \mu = \mathbf{x}\boldsymbol{\beta}.$$

- For binary data, a common link is the logistic

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{x}\boldsymbol{\beta}.$$

- For Poisson data, a common link is the log

$$g(\mu) = \log(\mu) = \mathbf{x}\boldsymbol{\beta}.$$

Bayesian Modeling with GLMs

- For a generic GLM, with regression parameters $\boldsymbol{\beta}$ and a scale parameter α, the posterior is

$$p(\boldsymbol{\beta}, \alpha|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \alpha) \times p(\boldsymbol{\beta}, \alpha).$$

- An immediate question is: How to specify a prior distribution $p(\boldsymbol{\beta}, \alpha)$?
- How to perform the computations required to summarize the posterior distribution (including the calculation of Bayes factors)?

Bayesian Computation

Various approaches to computation are available:

- Conjugate analysis — the prior combines with likelihood in such a way as to provide analytic tractability (at least for some parameters).
- Analytical Approximations — asymptotic arguments used (e.g. Laplace).
- Numerical integration.
- Direct (Monte Carlo) sampling from the posterior, as we have already seen.
- Markov chain Monte Carlo — very complex models can be implemented, for example within the free software `WinBUGS`.
- Integrated nested Laplace approximation (INLA). Cleverly combines analytical approximations and numerical integration: we illustrate the use of this method in some detail.

Integrated Nested Laplace Approximation (INLA)

- To download INLA:

```
> source("http://www.math.ntnu.no/inla/givemeINLA.R")
> inla.upgrade()
```

- Alternatively, on a mac you can type

```
> install.packages("INLA.tgz", repos=NULL, type="source")
```

- The homepage of the software is here:
  http://www.r-inla.org/home
- There are also lots of example links at this website.
- The fitting of many common models is described here:
  http://www.r-inla.org/models/likelihoods
- INLA can fit GLMs, GLMMs and many other useful model classes.

INLA for the Linear Model

- We first fit a linear model to the FTO data with the default prior settings.

```
> liny <- fto$y
> linxg <- fto$xg
> linxa <- fto$xa
> linxint <- fto$xgxa
#
# Data should be input to INLA as either a list or a dataframe
#
> ftodf <-list(liny=liny,linxg=linxg,linxa=linxa,linxint=linxint)
> formula <- liny~linxg+linxa+linxint
> lin.mod <- inla(formula,data=ftodf,family=''gaussian'')
```

- We might wonder, where are the priors?
- We didn't specify any...but INLA has default choices (more on this later).

INLA for the Linear Model

```
> summary(lin.mod)
Fixed effects:
               mean     sd 0.025quant 0.5quant 0.975quant kld
(Intercept) -0.0609 1.3709    -2.7681  -0.0611     2.6496   0
linxg        2.9323 1.9367    -0.8943   2.9326     6.7593   0
linxa        2.8423 0.4134     2.0255   2.8424     3.6593   0
linxint      1.7329 0.5841     0.5795   1.7328     2.8878   0
Model hyperparameters:
                              mean   sd     0.025quant 0.5quant
Precision for the Gaus obsers 0.3055 0.1018 0.1457     0.2930
                                     0.975quant
Precision for the Gaus observations 0.5402
> summary(ols.fit) # From before!
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06822    1.42230  -0.048   0.9623
linxg        2.94485    2.01143   1.464   0.1625
linxa        2.84421    0.42884   6.632 5.76e-06 ***
linxint      1.72948    0.60647   2.852   0.0115 *
Residual standard error: 1.917 on 16 degrees of freedom
```

- The posterior means and standard deviations are in very close agreement with the OLS fits presented earlier.

INLA for the Linear Model

```
> summary(lin.mod)
Fixed effects:
                 mean     sd 0.025quant 0.5quant 0.975quant
(Intercept) -0.0609 1.3709     -2.7681  -0.0611      2.6496
linxg        2.9323 1.9367     -0.8943   2.9326      6.7593
linxa        2.8423 0.4134      2.0255   2.8424      3.6593
linxint      1.7329 0.5841      0.5795   1.7328      2.8878
Model hyperparameters:
                                  mean    sd      0.025quant 0.5quant 0.975quant
Precision for Gaus obsers 0.3055 0.1018 0.1457      0.2930   0.5402
```

- The model is

$$Y = \mathrm{E}[Y|x_g, x_a] = \beta_0 + \beta_g x_g + \beta_a x_a + \beta_{\text{int}} x_g x_a + \epsilon$$

  where $\epsilon|\sigma^2 \sim_{iid} N(0, \sigma^2)$.
- The four fixed effects are $\beta_0, \beta_g, \beta_a, \beta_{\text{int}}$ and for each the posterior mean and standard deviation are given along with the 2.5%, 50% and 97.5% quantiles of the posterior.
- The model hyperparameter is the precision σ^{-2}.
- Note that posterior quantiles are invariant to transformation so, for example, the posterior median for σ is $1/\sqrt{0.2930} = 1.85$ (compare with 1.92 from OLS fit).

R Code for Marginal Distributions

- It is straightforward to create plots of the marginal distributions using INLA.
- The code below sends the output to a file, `plot(lin.mod)` sends to a separate window.

```
> plot(lin.mod, pdf=T, prefix="linmod")
```

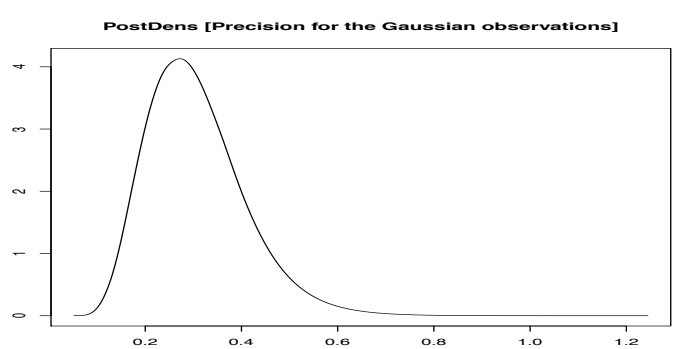

Figure 1 : Marginal distribution of the error precision.

FTO Posterior Marginal Distributions

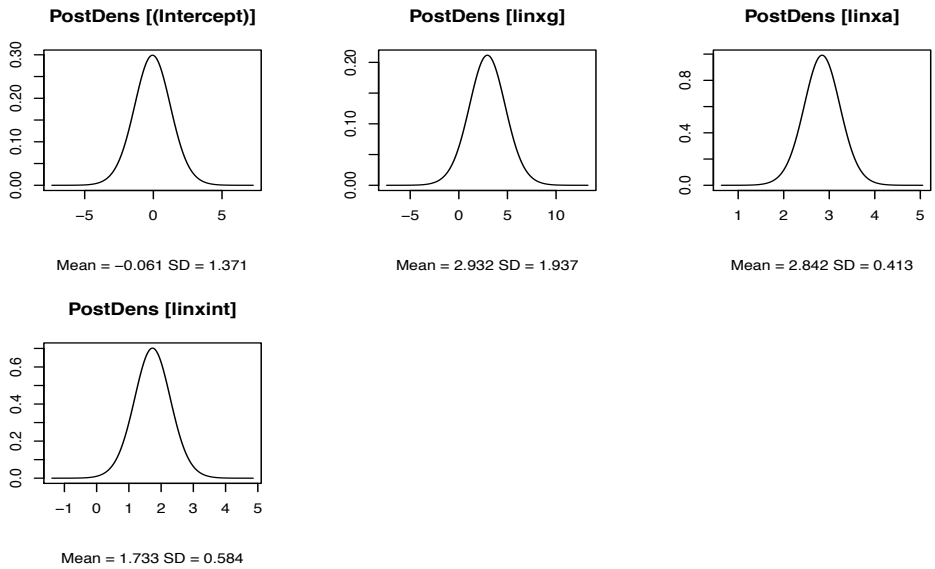

Figure 2 : Marginal distributions of the regression coefficients.

FTO Extended Analysis

- In order to carry out model checking we rerun the analysis, but now switch on a flag to obtain fitted values.

```
> lin.mod <- inla(liny~linxg+linxa+linxint,data=ftodf,
      family="gaussian",control.predictor=list(compute=TRUE))
> names(lin.mod) # Type this to see what can be extracted
> fitted <- lin.mod$summary.fitted.values[,1]
#
# Now extract the posterior median of the measurement error sd
> sigmamed <- 1/sqrt(lin.mod$summary.hyperpar[,4])
```

- With the fitted values we can examine the fit of the model. In particular:
  - Normality of the errors (sample size is relatively small).
  - Errors have constant variance (and are uncorrelated).
  - Linear model is adequate.

Assessing the Model

- The code below forms residuals and then forms
  1. a QQ plot to assess normality,
  2. a plot of residuals versus age, to assess linearity,
  3. a plot of residuals versus fitted values, to see if an unmodeled mean-variance relationship) and
  4. a plot of fitted versus observed for an overall assessment of fit.

```
> residuals <- (liny-fitted)/sigmamed
> par(mfrow=c(2,2))
> qqnorm(residuals,main="")
> title("(a)")
> abline(0,1,lty=2,col="red")
> plot(residuals~linxa,ylab="Residuals",xlab="Age")
> title("(b)")
> abline(h=0,lty=2,col="red")
> plot(residuals~fitted,ylab="Residuals",xlab="Fitted")
> title("(c)")
> abline(h=0,lty=2,col="red")
> plot(fitted~liny,xlab="Observed",ylab="Fitted")
> title("(d)")
> abline(0,1,lty=2,col="red")
```


FTO Diagnostic Plots

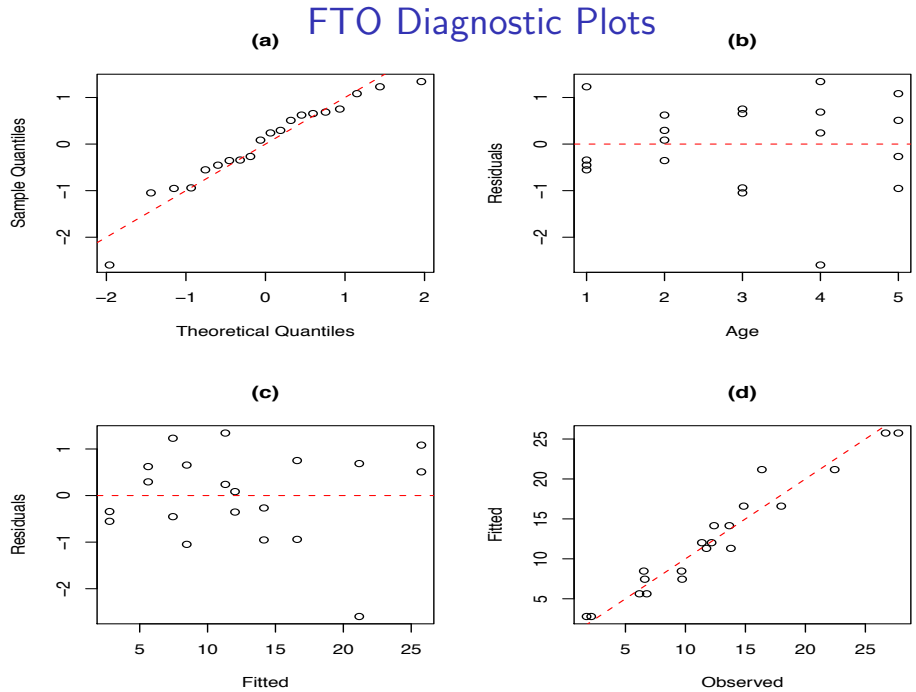

Figure 3 : Plots to assess model adequacy: (a) Normal QQ plot, (b) residuals versus age, (c) residuals versus fitted, (d) fitted versus observed.

Bayes Logistic Regression

- The likelihood is

$$Y(x)|p(x) \sim \text{Binomial}(\, N(x), p(x)\,), \qquad x = 0, 1, 2.$$

- Logistic link:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \theta x$$

- The prior is

$$p(\alpha, \theta) = p(\alpha) \times p(\theta)$$

  with
  - $\alpha \sim \text{normal}(\mu_\alpha, \sigma_\alpha)$ and
  - $\theta \sim \text{normal}(\mu_\theta, \sigma_\theta)$. where $\mu_\alpha, \sigma_\alpha, \mu_\theta, \sigma_\theta$ are constant that are specified to reflect prior beliefs.

Bayes Logistic Regression

- We perform two analyses.
- The first analysis uses the default priors in INLA (which are relatively flat).

```
> x <- c(0,1,2)
> y <- c(6,8,75)
> z <- c(10,66,163)
> cc.dat <- as.data.frame(rbind(y,z,x))
> cc.mod <- inla(y~x,family="binomial",data=cc.dat,Ntrials=y+z)
> summary(cc.mod)
Fixed effects:
              mean     sd 0.025quant 0.5quant 0.975quant
(Intercept) -1.807 0.4554    -2.7487  -1.7903    -0.9593
x            0.480 0.2505     0.0088   0.4726     0.9930
```

Prior Choice for Poisson-Lognormal Models

- It is convenient to specify lognormal priors for a positive parameter $\exp(\beta)$, since one may specify two quantiles of the distribution, and directly solve for the two parameters of the lognormal.
- Denote by $\text{LogNormal}(\mu, \sigma)$ the lognormal distribution for a generic parameter θ with $\text{E}[\theta] = \mu$ and $\text{var}(\log\theta) = \sigma^2$, and let θ_1 and θ_2 be the q_1 and q_2 quantiles of this prior.
- In our example, $\theta = \exp(\beta)$.
- Then it is straightforward to show that

$$\mu = \log(\theta_1)\left(\frac{z_{q_2}}{z_{q_2} - z_{q_1}}\right) - \log(\theta_2)\left(\frac{z_{q_1}}{z_{q_2} - z_{q_1}}\right), \ \sigma = \frac{\log(\theta_1) - \log(\theta_2)}{z_{q_1} - z_{q_2}}. \tag{1}$$

Prior Choice for Poisson-Lognormal Models

- As an example, suppose that for the odds ratio e^{β} we believe there is a 50% chance that the odds ratio is less than 1 and a 95% chance that it is less than 5; with $q_1 = 0.5, \theta_1 = 1.0$ and $q_2 = 0.95, \theta_2 = 5.0$, we obtain lognormal parameters $\mu = 0$ and $\sigma = (\log 5)/1.645 = 0.98$.
- There is a function in the `SpatialEpi` R package to find the parameters, as we illustrate.

```
library(SpatialEpi}
lnprior <- LogNormalPriorCh(1,5,0.5,0.95)
lnprior
$mu
[1] 0
$sigma
[1] 0.9784688
plot(seq(0,7,.1),dlnorm(seq(0,7,.1),meanlog=lnprior$mu,
 sdlog=lnprior$sigma),type="l",xlab="x",ylab="LogNormal Density")
```

The density is shown in Figure 4.

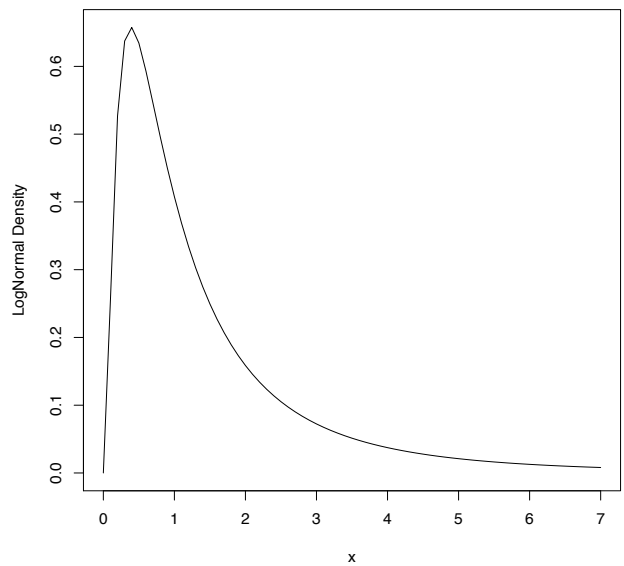

Figure 4 : Lognormal density with 50% point 1 and 95% point 5.

R Code for Logistic Regression Example

- In the second analysis we specify

$$\begin{aligned} \alpha &\sim \text{normal}(0, 1/0.1) \\ \theta &\sim \text{normal}(0, W) \end{aligned}$$

where W is such that the 97.5% point of the prior is log(1.5), i.e. we believe the odds ratio lies between 2/3 and 3/2 with probability 0.95.

```
# Now with informative priors
> W <- LogNormalPriorCh(1,1.5,0.5,0.975)$sigma^2
> cc.mod2 <- inla(y~x,family="binomial",data=cc.dat,Ntrials=y+z,
    control.fixed=list(mean.intercept=c(0),prec.intercept=c(.1),
                       mean=c(0),prec=c(1/W)))
> summary(cc.mod2)
Fixed effects:
                mean     sd 0.025quant 0.5quant 0.975quant
(Intercept) -1.3227 0.2896    -1.9005  -1.3194    -0.7641
x            0.1986 0.1536    -0.0999   0.1977     0.5027
plot(cc.mod2,pdf=T,prefix=''logistic'')
```

The quantiles for θ can be translated to odds ratios by exponentiating.

Logistic Marginal Plots

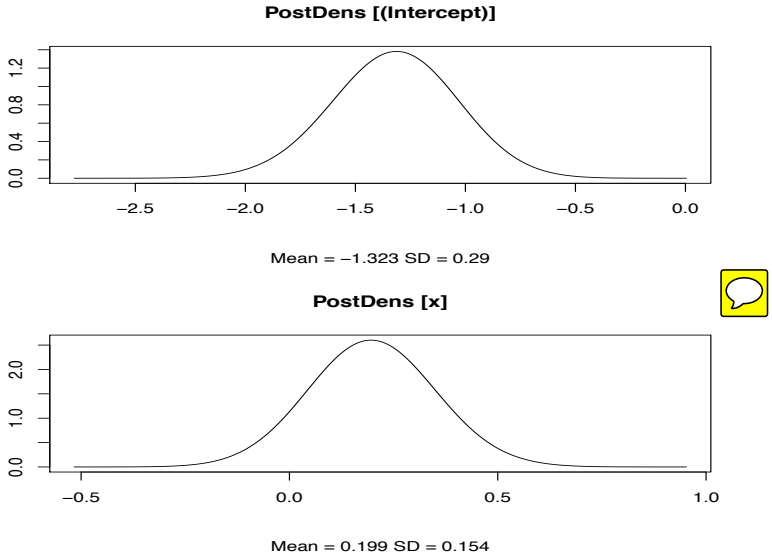

Figure 5 : Posterior marginals for the intercept α and the log odds ratio θ.

A Simple ANOVA Example

- We begin with simulated data from the simple one-way ANOVA model example:

$$\begin{aligned} Y_{ij}|\beta_0, b_i &= \beta_0 + b_i + \epsilon_{ij} \\ \epsilon_{ij}|\sigma_\epsilon^2 &\sim_{iid} \text{normal}(0, \sigma_\epsilon^2) \\ b_i|\sigma_b^2 &\sim_{iid} \text{normal}(0, \sigma_b^2) \end{aligned}$$

  $i = 1, ..., 10; j = 1, ..., 5$, with $\beta_0 = 0.5$, $\sigma_\epsilon^2 = 0.2^2$ and $\sigma_b^2 = 0.3^2$.
- b_i are random effects and ϵ_{ij} are measurement errors and there are two variances to estimate, σ_ϵ^2 and σ_b^2.
- In a fixed effects Bayesian model, the variance σ_b^2 would be fixed in advance.
- Simulation:

```
> sigma.b <- 0.3
> sigma.e <- 0.2
> m <- 10
> ni <- 5
> beta0 <- 0.5
> b <- rnorm(m, mean=0, sd=sigma.b)
> e <- rnorm(m*ni, mean=0, sd=sigma.e)
> Yvec <- beta0 + rep(b, each=ni) + e
> simdata <- data.frame(y=Yvec, ind=rep(1:m, each=ni))
```

A Simple ANOVA Example

- We fit the one-way ANOVA model and see reasonable recovery of the true values that were used to simulate the data.
- Not a big surprise, since we have fitted the model that was used to simulate the data!

```
> result <- inla(y ~ f(ind, model="iid"), data = simdata)
> summary(result)
Fixed effects:
                 mean          sd 0.025quant  0.5quant   0.975quant
(Intercept) 0.3780418 0.09710096 0.1828018   0.3780368  0.5734883
Model hyperparameters:
                      mean   sd      0.025quant 0.5quant 0.975quant
Precision for Gaus obs 22.816 4.943  14.439     22.389   33.755
Precision for ind      13.983 6.857  4.784      12.632   31.164
> sigma.est <- 1/sqrt(result$summary.hyperpar[,4])
> sigma.est
Gaus obs    ind          # Extract the posterior medians
0.2119460   0.2795877    # of the precision and invert
```

- `sigma.est` correspond to the posterior medians of σ_ϵ and σ_b, respectively.

The RNA-Seq Data: INLA Analysis

- Recall there are two replicates and so for each of N genes we obtain two sets of counts.
- For the diploid hybrid, let Y_{ij} be the number of A alleles for gene i and replicate j, and N_{ij} is the total number of counts, $j = 1, 2$.
- We fit a hierarchical logistic regression model starting with first stage:

$$Y_{ij}|N_{ij}, p_{ij} \sim \text{binomial}(N_{ij}, p_{ij})$$

  so that p_{ij} is the probability of seeing an A read for gene i and replicate j.
- At the second stage:

$$\text{logit } p_{ij} \quad = \quad \theta_i + \epsilon_{ij}$$

  where $\epsilon_{ij}|\sigma^2 \sim \text{normal}(0, \sigma^2)$ represent random effects that allow for excess-binomial variation; there are a pair for each gene.
- The θ_i parameters are taken as fixed effects with a relatively flat prior (the default choice in `INLA`).
- $\exp(\theta_i)$ is the odds of seeing an A read for gene i.

INLA Code for RNA-Seq Data: Data Setup

- Rows 1 and 2 represent the two replicates for gene 1, rows 3 and 4 represent the two replicates for gene 2, etc...
- `rep1` is the variable that defines the random effects.
- `xvar` is the gene number, there are 10 genes in this dataset.

```
repdat <- read.table("RNA-SeqN10.txt",header=T)
repdat
      y     n rep1 xvar
1  1963  7617    1    1
2  3676 10413    2    1
3   249   308    3    2
4   110   114    4    2
...
17  397   810   17    9
18  480   928   18    9
19  242   466   19   10
20  174   313   20   10
```


INLA Code for RNA-Seq Data: Model Fitting

- Below is the code for fitting the random effects model.
- The `-1` in the model specification removes the intercept, so that the factor levels are defined with one level for each gene.

```
> RNAdat <- data.frame(repdat)
> RNAfit <- inla(y~as.factor(xvar)-1+f(rep1,model="iid"),
          family="binomial",data=RNAdat,Ntrials=n)
> summary(RNAfit)
Fixed effects:
                      mean     sd 0.025quant 0.5quant 0.975quant
as.factor(xvar)1   -0.8316 0.2298    -1.2923  -0.8315    -0.3712
as.factor(xvar)2    2.0222 0.2932     1.4748   2.0105     2.6364
as.factor(xvar)3    0.1730 0.2559    -0.3351   0.1728     0.6822
as.factor(xvar)4    0.4302 0.2317    -0.0342   0.4303     0.8938
as.factor(xvar)5    0.5964 0.2304     0.1348   0.5964     1.0580
as.factor(xvar)6    0.5456 0.2347     0.0760   0.5456     1.0150
as.factor(xvar)7    1.3631 0.2826     0.8126   1.3599     1.9316
as.factor(xvar)8    0.0280 0.2308    -0.4345   0.0280     0.4903
as.factor(xvar)9    0.0149 0.2342    -0.4537   0.0150     0.4834
as.factor(xvar)10   0.1497 0.2406    -0.3302   0.1496     0.6303
```

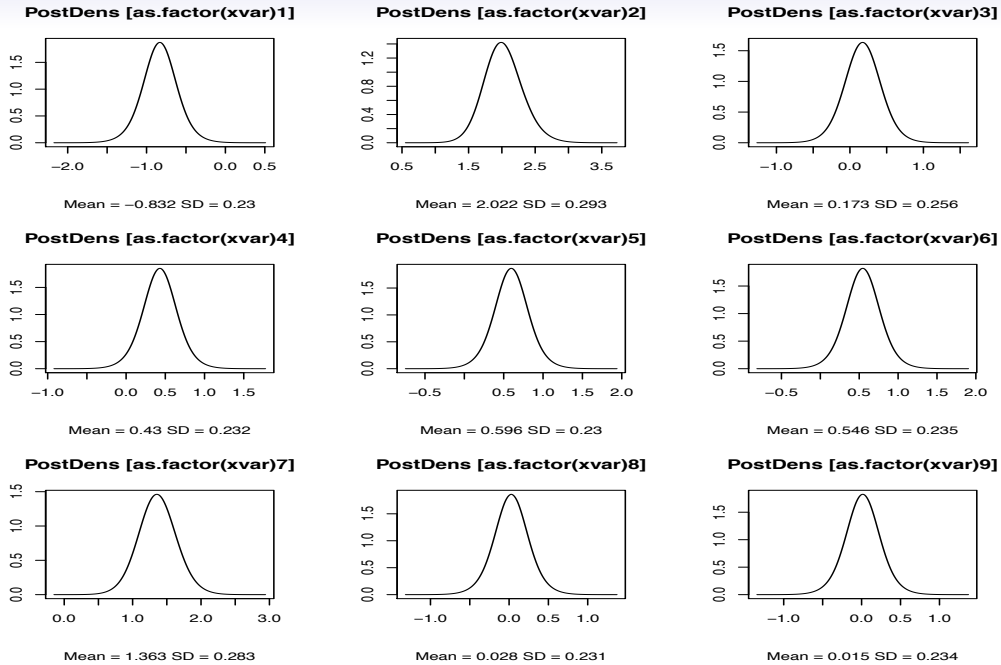

Figure 6 : Posterior marginals for the first 9 gene effects (compare with zero for evidence of cis effects). We plot 9 rather than all 10 for display purposes.

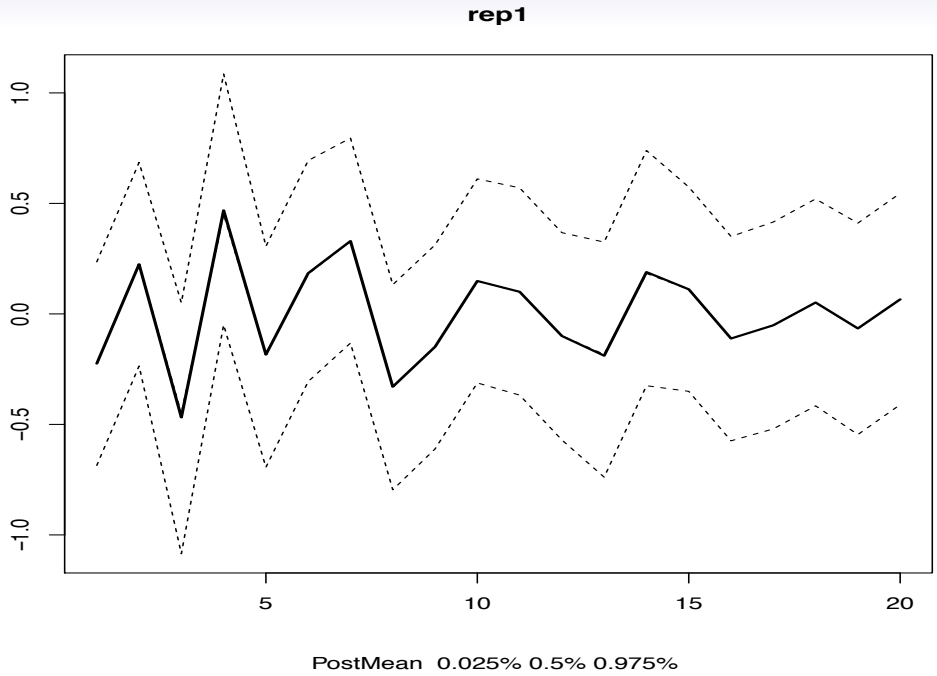

Figure 7 : Posterior quantiles for 20 random effects, which allow excess-binomial variation.

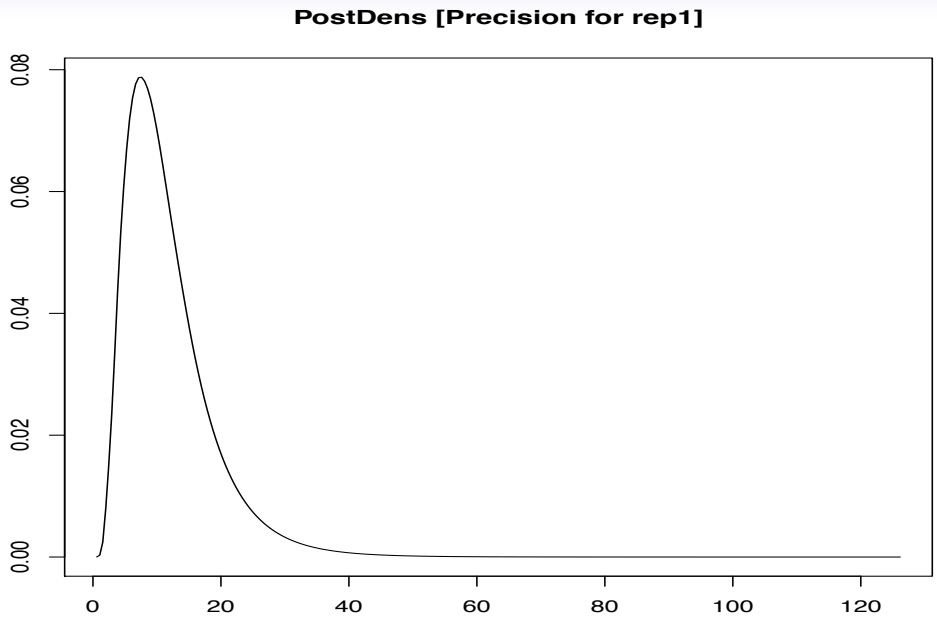

Figure 8 : Posterior marginal for precision of random effects.

An Informative Summary for the RNA-Seq Data

- We extract the 95% intervals and posterior medians for the log odds of being an A allele.
- Comparison with 0 (in Figure 9) gives an indication of cis effects.
- Genes 1, 2, 5, 6, 7 show evidence of cis effects.

```
thetasum <- RNAfit$summary.fixed[,3:5]
par(mfrow=c(1,1))
plot(thetasum[,2],seq(1,10),xlim=c(min(thetasum),max(thetasum)),
      ylab="Gene",xlab="Log Odds")
for (i in 1:10){
   lines(x=c(thetasum[i,1],thetasum[i,3]),y=c(i,i))
}
abline(v=0,col="red") # Intervals to the left/right of this line?
```

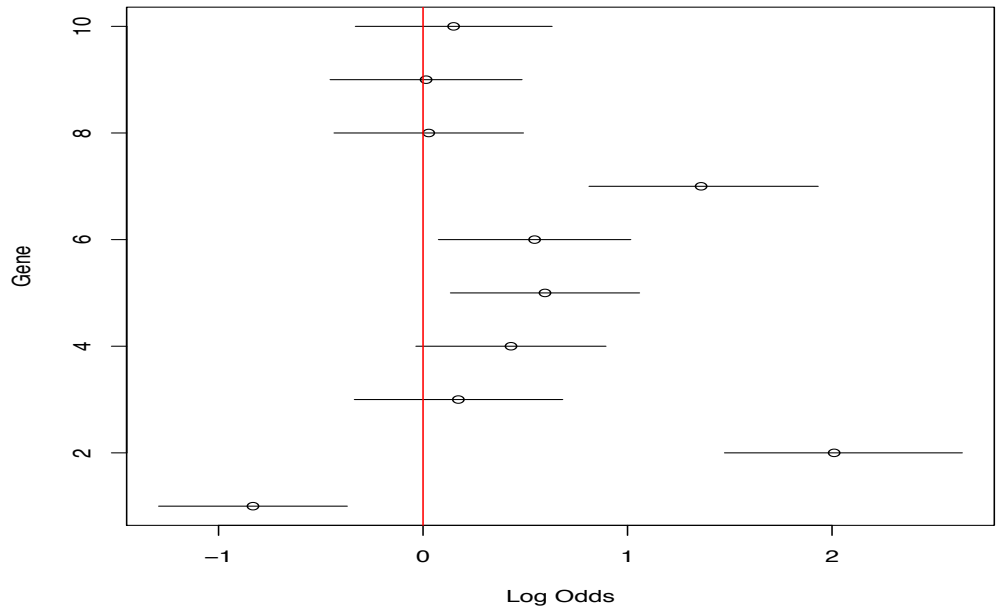

Figure 9 : Posterior marginal intervals for posterior of interest. Genes with posterior intervals that do not include zero, show evidence of cis effects.

Approximate Bayes Inference

- Particularly in the context of a large number of experiments, a quick and accurate model is desirable.
- We describe such a model in the context of a GWAS.
- This model is relevant when the sample size in each experiment is large.
- We first recap the normal-normal Bayes model.
- Subsequently, we describe the approximation and provide an example.

Recall: The Normal-Normal Model

- For the model
  - Prior: $\theta \sim$ normal(μ_0, τ_0^2) and
  - Likelihood: $Y_1, ..., Y_n|\theta \sim$ normal(θ, σ^2).
- Posterior

$$\theta|y_1, ..., y_n \sim \text{normal}(\mu, \tau^2)$$

where

$$\begin{aligned} \text{var}(\theta|y_1, ..., y_n) = \tau^2 &= [1/\tau_0^2 + n/\sigma^2]^{-1} \\ \text{Precision} = 1/\tau^2 &= 1/\tau_0^2 + n/\sigma \end{aligned}$$

and

$$\begin{aligned} \text{E}[\theta|y_1, ..., y_n] = \mu &= \frac{\mu_0/\tau_0^2 + \bar{y}n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \\ &= \mu_0 \left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2} \right) + \bar{y} \left(\frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \right) \end{aligned}$$

A Normal-Normal Approximate Bayes Model

- Consider again the logistic regression model

$$\text{logit } p_i = \alpha + x_i\theta$$

with interest focusing on θ.

- We require priors for α, θ, and some numerical/analytical technique for estimation/Bayes factor calculation.
- As discussed in Lecture 6 Wakefield (2007, 2009) considered replacing the likelihood by the approximation

$$p(\theta|\widehat{\theta}) \propto p(\widehat{\theta}|\theta)p(\theta)$$

where
  - $\widehat{\theta}|\theta \sim$ normal(θ, V) – the asymptotic distribution of the MLE,
  - $\theta \sim$ normal$(0, W)$ – the prior on the log RR. Can choose W so that 95% of relative risks lie in some range, e.g. [2/3,1.5].

Posterior Distribution

- Under the alternative, the posterior distribution for the log odds ratio θ is

$$\theta|\widehat{\theta} \sim \text{normal}(r\widehat{\theta}, rV)$$

where

$$r = \frac{W}{V+W}.$$

- Hence, we have shrinkage to the prior mean of 0.
- The posterior median for the odds ratio is $\exp(r\widehat{\theta})$ and a 95% credible interval is

$$\exp(r\widehat{\theta} \pm 1.96\sqrt{rV}).$$

- Note that as $W \to \infty$ and/or $V \to 0$ (which occurs as we gather more data) the non-Bayesian point and interval estimates are recovered (since $r \to 1$).

A Normal-Normal Approximate Bayes Model

- We are interested in the hypotheses: $H_0 : \theta = 0$, $H_1 : \theta \neq 0$ and evaluation of the Bayes factor

$$\text{BF} = \frac{p(\widehat{\theta}|H_0)}{p(\widehat{\theta}|H_1)}.$$

- Using the approximate likelihood and normal prior we obtain:

$$\text{Approximate Bayes Factor } = \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2} r\right),$$

with $Z = \frac{\widehat{\theta}}{\sqrt{V}}$, $r = \frac{W}{V+W}$.

A Normal-Normal Approximate Bayes Model

- The approximation can be combined with a Prior Odds $= \pi_0/(1-\pi_0)$ to give
$$\text{Posterior Odds on } H_0 = \frac{\text{BFDP}}{1-\text{BFDP}} = \text{ABF} \times \text{Prior Odds}$$
where BFDP is the Bayesian False Discovery Probability.
- BFDP depends on the power, through r.
- For implementation, all that we need from the data is the Z-score and the standard error $\sqrt{V}$, or a confidence interval.
- Hence, published results that report confidence intervals can be converted into Bayes factors for interpretation (see later lecture).
- The approximation relies on large sample sizes, so the normal distribution of the estimator provides a good summary of the information in the data.

Bayesian Logistic Regression: Estimation

- We return to the example presented at the start of the lecture.
- Case-control data for the disease Leber Hereditary Optic Neuropathy (LHON) disease with genotype data for marker rs6767450:

	CC $x=0$	CT $x=1$	TT $x=2$	Total
Cases	6	8	75	89
Controls	10	66	163	239
Total	16	74	238	328

```
> x <- c(0,1,2)
> y <- c(6,8,75)
> z <- c(10,66,163)
```

Bayesian Logistic Regression: Estimation

- Below we construct the posterior "by hand".

```
> logitmod <- glm(cbind(y,z)~x,family="binomial")
> thetahat <- logitmod$coef[2]
> V <- vcov(logitmod)[2,2]
#
# 97.5 point of prior is log(1.5) so that we with prob
# 0.95 we think theta lies in (2/3,1.5)
#
> W <- LogNormalPriorCh(1,1.5,0.5,0.975)$sigma^2
> r <- W/(V+W)
> r
[1] 0.405545  # Not so much data here, so weight on prior is high.
# Bayesian posterior median
> exp(r*thetahat)
        x
1.214281   # Shrunk towards prior median of 1
# Note: INLA estimate (with same prior) is 1.22 and approximate
# posterior SD here is sqrt(rV)=0.159, INLA version is 0.154.
# Bayesian approximate 95% credible interval
> exp(r*thetahat-1.96*sqrt(r*V))
        x
0.888
> exp(r*thetahat+1.96*sqrt(r*V))
        x
1.660
```


Bayesian Logistic Regression: Hypothesis Testing

- Now we turn to testing using Bayes factors.
- We examine the sensitivity to the prior on the alternative, π_1.

```
> pi1 <- c(1/2,1/100,1/1000,1/10000,1/100000) # 5 prior probs on the null
> source("http://faculty.washington.edu/jonno/BFDP.R")
> BFcall <- BFDPfunV(thetahat,V,W,pi1)
> BFcall
$BF
        x
0.5110967
$pH0
        x
0.256323
$pH1
        x
0.5015156
$BFDP # Corresponding 5 posterior probs of the null
[1] 0.3382290 0.9806196 0.9980453 0.9998044 0.9999804
```

- So data are twice as likely under the alternative (0.502) as compared to the null (0.256).
- Apart from under the 0.5 prior, under these priors the overall evidence is of no association.

Combination of Data Across Studies

- Suppose we wish to combine data from two studies where we assume a common log odds ratio θ.
- The estimates from the two studies are $\widehat{\theta}_1, \widehat{\theta}_2$ with standard errors $\sqrt{V_1}$ and $\sqrt{V_2}$.
- The Bayes factor is
$$\frac{p(\widehat{\theta}_1, \widehat{\theta}_2 | H_0)}{p(\widehat{\theta}_1, \widehat{\theta}_2 | H_1)}.$$
- The approximate Bayes factor is
$$\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2) = \mathrm{ABF}(\widehat{\theta}_1) \times \mathrm{ABF}(\widehat{\theta}_2 | \widehat{\theta}_1) \tag{2}$$
where
$$\mathrm{ABF}(\widehat{\theta}_2 | \widehat{\theta}_1) = \frac{p(\widehat{\theta}_2 | H_0)}{p(\widehat{\theta}_2 | \widehat{\theta}_1, H_1)}$$
and
$$p(\widehat{\theta}_2 | \widehat{\theta}_1, H_1) = \mathrm{E}_{\theta | \widehat{\theta}_1} \left[p(\widehat{\theta}_2 | \theta) \right]$$
so that the density is averaged with respect to the posterior for θ.
- Important Point: The Bayes factors are not independent.

Combination of Data Across Studies

- This leads to an approximate Bayes factor (which summarizes the data from the two studies) of
$$\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2) = \sqrt{\frac{W}{RV_1V_2}} \exp\left\{ -\frac{1}{2} \left(Z_1^2 RV_2 + 2Z_1Z_2R\sqrt{V_1V_2} + Z_2^2 RV_1 \right) \right\}$$
where
  - $R = W/(V_1W + V_2W + V_1V_2)$
  - $Z_1 = \frac{\widehat{\theta}_1}{\sqrt{V_1}}$ and
  - $Z_2 = \frac{\widehat{\theta}_2}{\sqrt{V_2}}$ are the usual Z statistics.
- The ABF will be small (evidence for H_1) when the absolute values of Z_1 and Z_2 are large and they are of the same sign.

Combination of Data Across Studies: The General Case

- Suppose we have K studies with estimates $\widehat{\theta}_k$ and asymptotic variances V_k, $k = 1, ..., K$.
- Assume a common underlying parameter θ.
- The Bayes factor is given by

$$\begin{aligned}
\mathrm{BF}_K &= \frac{p(\widehat{\theta}_1, ..., \widehat{\theta}_K | H_0)}{p(\widehat{\theta}_1, ..., \widehat{\theta}_K | H_1)} \\
&= \frac{\prod_{k=1}^{K} (2\pi V_k)^{-1/2} \exp\left(-\frac{\widehat{\theta}_k^2}{2V_k}\right)}{\int \prod_{k=1}^{K} (2\pi V_k)^{-1/2} \exp\left(-\frac{(\widehat{\theta}_k^2 - \theta)^2}{2V_k}\right) (2\pi W)^{-1/2} \exp\left(-\frac{\theta^2}{2V_k}\right) d\theta} \\
&= \sqrt{W\left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)} \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{K} \frac{\widehat{\theta}_k}{V_k}\right)^2 \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}\right]
\end{aligned}$$

Combination of Studies: The General Case

- The posterior is given by

$$\theta | \widehat{\theta}_1, ..., \widehat{\theta}_K \sim \mathrm{normal}(\mu, \sigma^2)$$

where

$$\begin{aligned}
\mu &= \left(\sum_{k=1}^{K} \frac{\widehat{\theta}_k}{V_k}\right) \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1} \\
\sigma^2 &= \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}
\end{aligned}$$

Example of Combination of Studies in a GWAS

- We illustrate how reported confidence intervals can be converted to Bayesian summaries.
- Frayling *et al.* (2007) report a GWAS for Type II diabetes.
- For SNP rs9939609:

Stage	Estimate (CI)	p-value	$-\log_{10}$ BF	$\Pr(H_0\|\text{data})$ with prior: 1/5,000	1/50,000
1st	1.27 (1.16–1.37)	6.4×10^{-10}	7.28	0.00026	0.0026
2nd	1.15 (1.09–1.23)	4.6×10^{-5}	2.72	0.905	0.990
Combined	–	–	13.8	8×10^{-11}	8×10^{-10}

- Combined evidence is stronger than each separately since the point estimates are in agreement.
- For summarizing inference the (5%, 50%, 95%) points for the RR are:

Prior	1.00 (0.67–1.50)
First Stage	1.26 (1.17–1.36)
Combined	1.21 (1.15–1.27)

Conclusions

- Computationally GLMs and GLMMs can now be fitted in a relatively straightforward way.
- `INLA` is very convenient and is being constantly improved.
- As with all analyses, it is crucial to check modeling assumptions (and there are usually more in a Bayesian analysis).
- For binary observations `INLA` can produce inaccurate estimates.
- Markov chain Monte Carlo provides an alternative for computation. `WinBUGS` is one popular implementation.
- Other MCMC possibilities include: `JAGS`, `BayesX`, Stan.

References

Connelly, C., Wakefield, J., and Akey, J. (2014). Evolution and architecture of chromatin accessibility and function in yeast. *PLoS Genetics*. To appear.

Frayling, T., Timpson, N., Weedon, M., Zeggini, E., Freathy, R., and et al., C. L. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889–894.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, **81**, 208–227.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with *p*-values. *Genetic Epidemiology*, **33**, 79–86.