Module 7, part III: Bayesian Hierarchical Regression

BIOS 526

Reading

- R help file for *MCMCregress ( )* and *MCMChregress ( )*.

Concepts

- Priors for Bayesian hierarchical model.
- Comparison between Bayesian and Frequentist estimates.
- Prior sensitivity analysis for random effects.

Random Effect Model

Let index $i = 1, \ldots, n$ denote group ID, and let index $j = 1, \ldots, r_i$ denote observation within group i.

We will first consider the random intercept model.

$$y_{ij} = \beta_0 + \theta_i + \boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \epsilon_{ij}$$

$$\theta_i \overset{iid}{\sim} N(0, \tau^2) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

where $\boldsymbol{x}_{ij} \in \mathbb{R}^p$ and $\boldsymbol{\gamma}$ is the corresponding vector of regression coefficients. Also let $N = \sum_i r_i$

The above can be written in an equivalent hierarchical model:

$$y_{ij} = \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \epsilon_{ij}$$

$$\beta_{0i} \overset{iid}{\sim} N(\mu, \tau^2) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

where the random effects are assumed to be centered around an unknown population average μ, where $\mu = \beta_0$.

Bayesian Random Intercept Model

A Bayesian version of the random intercept model is complete with prior distributions assigned to all the unknown parameters.

$$y_{ij} = \beta_{0i} + \boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \epsilon_{ij}$$

$$\beta_{0i} \overset{iid}{\sim} N(\mu, \tau^2) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

Priors:

$$\begin{aligned}
\sigma^2 &\sim \text{Inv-Gamma}(\,\nu,\, \delta\,) \\
\boldsymbol{\gamma} &\sim N_p(\,\mathbf{0}, v_\gamma^2 \mathbf{I}_{p\times p}\,) \\
\mu &\sim N(\,0,\, v_\mu^2\,) \\
\tau^2 &\sim \text{Inv-Gamma}(\,\alpha_0,\, \beta_0\,).
\end{aligned}$$

In `MCMChregress`, we will parameterize τ^2 using an inverse Wishart:

$$\tau^2 \sim \text{Inv-Wishart}(r, rR).$$

where $r = 2\alpha_0$ and $R = 2\beta_0/r$.
To reflect uninformative priors, we often set v_γ^2 and v_μ^2 to be large (e.g. 1000^2) and r, ν, δ to be small.

BHM

We can write the ingredients of a Bayesian Hierarchical Model as a Data model, Process model, and Parameter model. Let $\boldsymbol{\theta} = [\sigma^2, \boldsymbol{\gamma}, \mu, \tau^2]$:

1. Data model: $[Y|Z, \boldsymbol{\theta}]$. This usually is a measurement error model, where the parameters are nuisance parameters.
Here, $y_{ij} \sim N(z_{ij}, \sigma^2)$.

2. Process model: $[Z|\boldsymbol{\theta}]$ This is the "scientific process," where the parameters have biological meaning.
Here, $z_{ij} \sim N(\mu + \boldsymbol{x}'_{ij}\boldsymbol{\gamma}, \tau^2)$, $Cov(z_{ij}, z_{ij'}|\mu, \boldsymbol{\gamma}) = \tau^2$.

3. Parameter model: $[\boldsymbol{\theta}]$
Here,

$$\begin{aligned} \sigma^2 &\sim \text{Inv-Gamma}(\,\nu,\, \delta\,) \\ \boldsymbol{\gamma} &\sim N_p(\,\mathbf{0}, v_\gamma^2 \mathbf{I}_{p\times p}\,) \\ \mu &\sim N(\,0,\, v_\mu^2\,) \\ \tau^2 &\sim \text{Inv-Gamma}(\,\alpha_0,\, \beta_0\,) \end{aligned}$$

Bayesian Random Slope Model

For random intercept and a random slope on covariate x_{ij}:

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma} + \epsilon_{ij}$$

$$\begin{bmatrix}\beta_{0i}\\ \beta_{1i}\end{bmatrix} \sim N\left(\boldsymbol{\mu} = \begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \Sigma_{2\times 2}\right) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

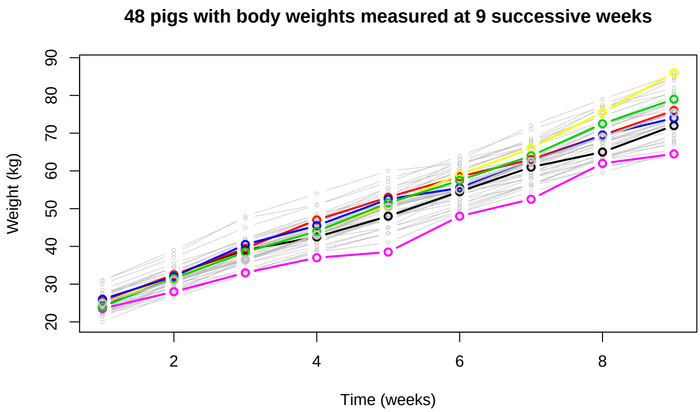
Priors:

$$\begin{aligned}
\boldsymbol{\gamma} &\sim N_p(\,\mathbf{0}, v_\gamma^2 \mathbf{I}_{p\times p}\,)\\
\boldsymbol{\mu} &\sim N(\,\mathbf{0},\, v_\mu^2 \mathbf{I}_{2\times 2}\,)\\
\sigma^2 &\sim \text{Inv-Gamma}(\,\nu,\, \delta\,)\\
\Sigma &\sim \text{Inv-Wishart}(\,r,\, r\mathbf{R}_{2\times 2}\,)
\end{aligned}$$

We treat the random coefficients as realizations from a bivariate normal distribution with population mean $\boldsymbol{\mu}$ and population covariance Σ.

It is straightforward to extend the model to include multiple random slopes.

Pig Weight Data



Let y_{ij} be the weight (kg) at the j^{th} week for the i^{th} pig.

MCMChregress in library MCMCpack

```
> library (MCMCpack)
> MCMChregress
function (fixed, random, group, data, burnin = 1000, mcmc = 10000,
    thin = 10, verbose = 1, seed = NA, beta.start = NA, sigma2.start = NA,
    Vb.start = NA, mubeta = 0, Vbeta = 1e+06, r, R, nu = 0.001,
    delta = 0.001, ...)
```

The MCMChregress function fits a Bayesian hierarchical model. It estimates an intercept and assumes random effects have mean zero.

- fixed: a formula with form Y$\sim$ x1 + x2 + ... for fixed effects
- random: a formula with form $\sim$ z1 + z2 + ... for random effects
- group: the variable name that indexes group ID
- mu.beta: prior mean for the fixed effects
- Vbeta: prior variance of the fixed effects
- r, R: hyper-parameters for the random effect heterogeneity.
- nu, delta: hyper-parameters of the residual variance

MCMChregress in library MCMCpack

A warning in the R help file:

"NOTE: We do not provide default parameters for the priors on the precision matrix for the random effects. When fitting one of these models, it is of utmost importance to choose a prior that reflects your prior beliefs about the random effects."

Why?

- We usually have a smaller sample size to estimate heterogeneity compared to estimating residual error variance. When the number of groups is small, the prior will have a greater impact.
- The multilevel Normal model assumes that

$$Var(y_{ij}) = Var(\beta_{0i}) + Var(\epsilon_{ij}) = \tau^2 + \sigma^2.$$

  We wish to partition the total variance into two components. These two parameters can be highly correlated: larger $\tau^2 \rightarrow$ smaller σ^2, and vice versa.

The Inverse-Wishart Distribution

The MCMChregress function uses the inverse-Wishart prior for the random effects covariance.

The inverse-Wishart distribution:

- is a multivariate generalization of the inverse-Gamma distribution.
- is a probability distribution for random covariance matrices! This is tricky because a covariance matrix Σ needs to be
  - symmetric
  - positive definite: $a'\Sigma a > 0$ for all vector $a \neq 0$.
- Parameterizations differ...
- if $\Sigma \sim \text{Inv-Wishart}(r, r\mathbf{R})$, then for $r > q + 1$,

$$E[\,\Sigma\,] = \frac{r}{r - q - 1}\mathbf{R}$$

where $\mathbf{R}$ is a $q \times q$ positive definite matrix. MCMChregress, $r \geq q$.

The Inverse-Wishart Distribution

The inverse-Wishart distribution is useful for making inference about a population covariance matrix. Specifically, let $\boldsymbol{\theta}_i = (\theta_{0i}, \theta_{1i})$, for i in $1, \ldots, n$. Let's assume

$$\boldsymbol{\theta}_i \sim N(0, \Sigma)$$

$$\Sigma \sim \text{Inv-Wishart}(r, r\mathbf{R}).$$

Then let $\mathbf{S}$ be the sample covariance matrix from data generated from a zero-mean multivariate normal distribution. The posterior distribution of $[\Sigma|\text{data}]$ is

$$[\Sigma|\text{data}] \sim \text{Inv-Wishart}(r+n, r\mathbf{R} + n\mathbf{S}),$$

which has mean

$$E[\Sigma|\text{data}] = \frac{r\mathbf{R} + n\mathbf{S}}{r+n-p-1}.$$

where p is the dimension of $\boldsymbol{\theta}_i$. Note that for large n the above is approximately $\frac{n\mathbf{S}}{n} = \mathbf{S}$.

Therefore, for prior $\Sigma \sim \text{Inv-Wishart}(r, r\mathbf{R})$, we can think of

- $\mathbf{R}$ as the prior assumption on the random effect covariance matrix;
- r as the sample size that $\mathbf{R}$ is based on.

The Inverse-Wishart Distribution: Examples

Some realizations from Inv-Wishart $\left(r = 2,\ r\mathbf{R} = 2 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$.

```
> riwish (2, 2*diag(2))
         [,1]     [,2]
[1,] 7.993478 1.056120
[2,] 1.056120 1.184883
```

```
> riwish (2, 2*diag(2))
          [,1]     [,2]
[1,] 0.9417932 1.434751
[2,] 1.4347506 6.704089
```

```
> riwish (2, 2*diag(2))
          [,1]      [,2]
[1,] 153.62756 -68.26434
[2,] -68.26434  30.74045
```

Some realizations from Inv-Wishart $\left(r = 20,\ r\mathbf{R} = 20 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$.

```
> riwish (20, 20*diag(2))
           [,1]       [,2]
[1,]  1.0477453 -0.4198504
[2,] -0.4198504  0.8217627
```

```
> riwish (20, 20*diag(2))
          [,1]      [,2]
[1,] 0.7991031 0.1561055
[2,] 0.1561055 1.1227496
```

```
> riwish (20, 20*diag(2))
            [,1]        [,2]
[1,]  1.09554075 -0.09640348
[2,] -0.09640348  0.94267139
```

Some realizations from Inv-Wishart $\left(r = 200,\ r\mathbf{R} = 200 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$.

```
> riwish (200, 200*diag(2))
            [,1]        [,2]
[1,]  0.92784643 -0.04097586
[2,] -0.04097586  0.90093856
```

```
> riwish (200, 200*diag(2))
           [,1]       [,2]
[1,] 0.97532764 0.05100576
[2,] 0.05100576 1.06659492
```

```
> riwish (200, 200*diag(2))
            [,1]        [,2]
[1,]  0.99540455 -0.05506508
[2,] -0.05506508  1.10315592
```

Pig Weight Example: lmer

Start with a random intercept model with lmer ( ).

```
> library (lme4)
> dat[1:3,]
  id weeks weight
1  1     1     24
2  1     2     32
3  1     3     39

> fit = lmer (weight~weeks + (1|id), data = dat)
> summary (fit)
Linear mixed model fit by REML
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 15.1418  3.8913
 Residual              4.3947  2.0964
Number of obs: 432, groups: id, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) 19.35561    0.60311   32.09
weeks        6.20990    0.03906  158.97
```

The REML estimate of τ^2 is around 15.1. One common practice is to use this as the mean of our prior distribution.

Pig Weight Example (MCMChregress)

Assume $\tau^2 \sim \text{Inv-Wishart}(r = 1, R = 15.1)$.

```
####### Fit a Bayesian random effect model #######
> fit = MCMChregress (fixed = weight~weeks, random = ~ 1, group = "id",
      r = 1, R=15.1,  data = dat)

## The fitted object has two items, the "mcmc" samples and the fitted values
> names (fit)
> [1] "mcmc"   "Y.pred"


## Default is 10000 iterations, 1000 burn-in, and save every 10th sample
> summary (fit$mcmc)
Iterations = 1001:10991
Thinning interval = 10
Number of chains = 1
Sample size per chain = 1000


## Let's extract the mcmc samples which is 1000 by 53
## 52 parameters = 1 overall intercept + 1 slope for weeks
      + 48 random intercepts + 2 variances
> post.samp = as.data.frame (fit$mcmc)
> dim (post.samp)
[1] 1000   53
```

Pig Weight Example

```
#### Print the first 3 posterior samples of the fixed intercept, fixed weeks slope,
####    and random intercept for pig ID 1 and 10

> post.samp[1:3, 1:4]
beta.(Intercept) beta.weeks b.(Intercept).1 b.(Intercept).10
1         19.60016   6.166843       -2.513519        2.4223889
2         18.28803   6.328646       -1.197773        0.9636971
3         20.72839   6.194585       -3.413064       -0.1510965


#### Print the first 3 posterior samples of the random intercept for pig ID 8 and 9,
####    heterogeneity, residual variance, and deviance (-2loglik)

> post.samp[1:3, 49:53]
  b.(Intercept).8 b.(Intercept).9 VCV.(Intercept).(Intercept)   sigma2 Deviance
1       -1.450973       -6.990467                   15.10369 4.435315 1883.749
2       -1.588611       -6.344787                   15.43770 4.478911 1876.732
3       -2.147836       -6.563484                   16.43365 4.517250 1877.318

#The names are pretty bad, but we can change them manually if desired.
```

Pig Weight Example

Recall our model:

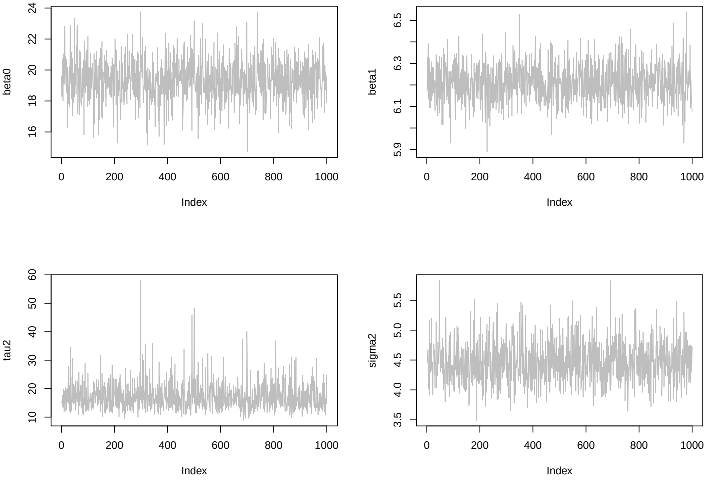
$$y_{ij} = \beta_0 + \theta_i + \beta_1 Week_{ij} + \epsilon_{ij}$$

$$\theta_i \sim N(0, \tau^2) \qquad \epsilon_{ij} \sim N(0, \sigma^2),$$

Extract useful summary statistics. Note the double square brackets!

```
> beta0 = post.samp[["beta.(Intercept)"]]
> beta1 = post.samp[["beta.weeks"]]
> tau2 = post.samp[["VCV.(Intercept).(Intercept)"]]
> sigma2 = post.samp[["sigma2"]]

### Trace plot
> plot (tau2, type = "l")

### Histogram
> hist (tau2)

### Summary statistics
> mean(tau2)
7.27796
> quantile (tau2, c(0.025, 0.5, 0.975))
    2.5%      50%    97.5%
10.94793 16.29694 29.38310
```

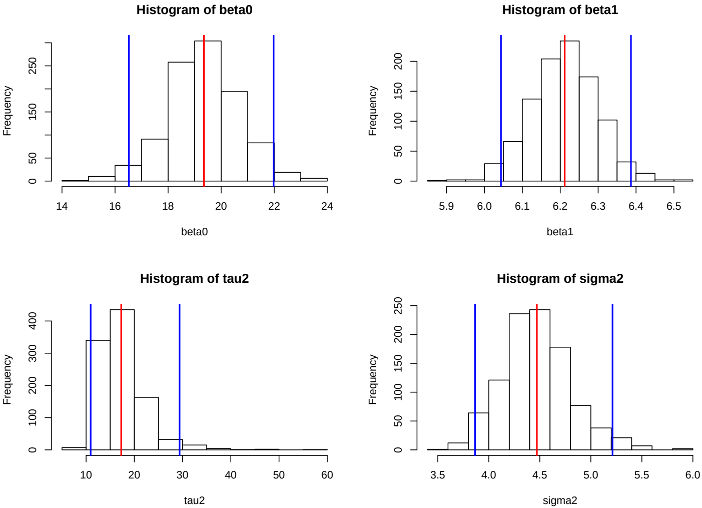
Trace Plots: Check for Convergence



Histograms: Visualize Posterior Distributions

Red line = Posterior mean

Blue lines = Posterior 2.5% and 97.5% quantiles



Compare to Estimates from *lmer( )*

	lmer			MCMChregress		
Parameter	Estimate	SE	95% C.I.	Estimate	SE	95% P.I.
β_0	19.3	0.60	(18.1, 20.5)	19.3	1.33	(16.5, 22.0)
β_1	6.21	0.04	(6.12, 6.29)	6.21	0.09	(6.04, 6.39)
τ^2	15.1	NA	NA	17.3	4.85	(10.9, 29.4)
σ^2	4.4	NA	NA	4.5	0.34	(3.9, 5.2)

- The point estimates of the fixed effects are nearly identical.
- The standard errors of the fixed effects are larger for the Bayesian analysis. This can be attributed to the uncertainties in τ^2, σ^2, and all of the random effects θ_{0i}!
- The estimates of τ^2 show the greatest disagreement, possibly due to the large right-skewness in the posterior distribution of τ^2.

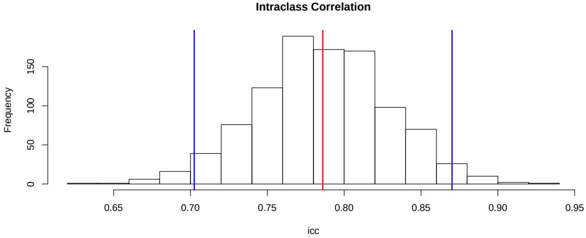
Posterior Inference: ICC

One advantage of the Bayesian approach via MCMC is the ability to utilize posterior samples of the parameters to quantify uncertainties.

Ex. recall the intraclass correlation is given by:

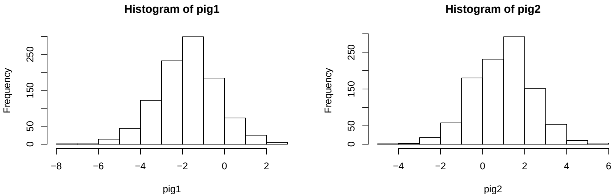
$$\frac{\tau^2}{\tau^2 + \sigma^2}.$$

```
#### Estimate intraclass correlation with posterior samples
> icc = tau2/(tau2+sigma2)
> quantile (icc, c(0.025, .5, .975))
     2.5%       50%     97.5%
0.7024624 0.7852280 0.8701815
> hist (icc)
```



Posterior Inference: Random Effects

We also have posterior samples of every random effect θ_i.

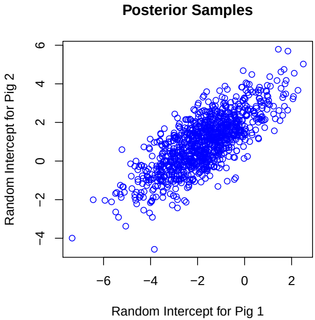


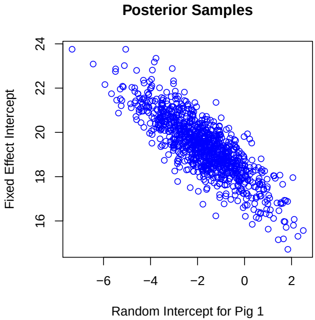
Evidence that the baseline weight of pig 2 is larger than pig 1?

```
### Random intercept for pig ID = 1
> pig1 = post.samp[["b.(Intercept).1"]]
> quantile (pig1, c(0.025, .5, .975))
     2.5%        50%      97.5%
-4.593144 -1.722720  1.101687

### Random intercept for pig ID = 2
> pig2 = post.samp[["b.(Intercept).2"]]
> quantile (pig2, c(0.025, .5, .975))
     2.5%        50%      97.5%
-1.883674  1.050871  3.669131

> table (pig2 > pig1)
FALSE  TRUE
    5   995
```

Posterior Inference: Random Effects

Note that the estimated random effects of θ_1 and θ_2 are highly positively correlated. This is because the random intercepts tend to be negatively correlated with the overall intercept.





When we directly use posterior samples to make inference, the complex correlations between estimates are taken into account!

Prior Sensitivity Analysis

Since Bayesian inference uses prior distributions, it is important to conduct sensitivity analyses by varying hyper-parameter values. Here we refit the model with a more informative distribution for τ^2 by setting $r = 10, 100, 1000$ (prior "sample size" of τ^2).

	r=10		r=100		r=1000	
Parameter	Estimate	SE	Estimate.	SE	Estimate	SE
β_0	19.3	1.32	19.4	1.29	19.3	1.33
β_1	6.21	0.08	6.21	0.09	6.21	0.08
τ^2	16.8	3.85	15.7	1.89	15.2	0.67
σ^2	4.5	0.34	4.47	0.33	4.45	0.33
θ_{17}	7.1	1.43	7.0	1.39	7.0	1.40
θ_{11}	0.6	1.38	0.5	1.38	0.53	1.37
θ_{25}	-5.7	1.38	-5.9	1.39	-5.8	1.39

We chose θ_{17}, θ_{11}, and θ_{25} in the table to look at how small, large, and typical θ_i changes. Overall, the τ^2 prior's impact on the fixed and random effects are minimal. But note we set `R` equal to the REML estimate.

Prior Sensitivity Analysis

However, inference does change if we change both the prior precision r, and prior heterogeneity value **R**.

	r=1, **R** $= 15.1$		r=1, **R** $= 5$		r=1000, **R** $= 5$	
Parameter	Estimate	SE	Estimate.	SE	Estimate	SE
β_0	19.3	1.33	19.3	1.32	19.3	0.89
β_1	6.21	0.08	6.21	0.09	6.21	0.08
τ^2	17.2	4.85	17.0	4.78	5.4	0.24
σ^2	4.47	0.34	4.47	0.34	4.53	0.34
θ_{17}	7.0	1.41	7.0	1.40	6.65	0.98
θ_{11}	0.5	1.38	0.5	1.38	0.50	0.96
θ_{25}	-5.9	1.47	-5.9	1.47	-4.43	0.96

Note that the estimated random intercepts become less extreme when we assign an informative prior for τ^2. However, the overall intercept (β_0) and slope β_1 are robust.

Prior Sensitivity Analysis

Finally, let's assume a very informative prior for σ^2 centered at about $1/2$ of the estimate from *lmer ( )*.

	Uninformative $[\sigma^2]$		Informative $[\sigma^2]$	
Parameter	Estimate	SE	Estimate.	SE
β_0	19.3	1.33	19.3	0.87
β_1	6.21	0.08	6.21	0.04
τ^2	17.2	4.85	17.0	3.82
σ^2	4.47	0.34	2.20	0.007
θ_{17}	7.0	1.41	7.14	0.98
θ_{11}	0.5	1.38	0.53	0.99
θ_{25}	-5.9	1.47	-5.9	1.00

All standard error estimates are smaller!

Random Slope Model: lmer

Start with *lmer ( )*.

```
> fit.lmer = lmer (weight~weeks + (weeks |id), data = dat)

> summary (fit.lmer)

Linear mixed model fit by REML
Formula: weight ~ weeks + (weeks | id)
   Data: dat
  AIC  BIC logLik deviance REMLdev
 1753 1777 -870.4     1738    1741
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 6.9865   2.64319
          weeks       0.3800   0.61644  -0.063
 Residual             1.5968   1.26366
Number of obs: 432, groups: id, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) 19.35561    0.40387   47.93
weeks        6.20990    0.09204   67.47
```

We will center the prior of the heterogeneity covariance at $\begin{bmatrix} 7.0 & 0 \\ 0 & 0.4 \end{bmatrix}$.

Random Slope Model: MCMChregress

```
### Define prior covariance mean
> R.prior = matrix ( c(7, 0, 0, 0.4), ncol = 2)
> R.prior
     [,1] [,2]
[1,]    7  0.0
[2,]    0  0.4

### Fit model
> fit = MCMChregress (fixed = weight~weeks, random = ~ 1+weeks, group = "id",
    r = 2, R=R.prior, data = dat)

> post.samp = as.data.frame (fit$mcmc)
> dim (post.samp)
[1] 1000  104

> summary (fit$mcmc) ## Print summary statistics for all parameters
> names (post.samp)  ## Print variable names. Useful if you want to look at specific parame
```

Note that here we set $r = 2$. When fitting a Bayesian hierarchical model, r needs to be $\geq$ the dimension of R. Intuitively, you need at least 2 data points to estimate a 2×2 covariance matrix.

Also, we have 102 posterior parameters! 48 random intercepts + 48 random slopes + 2 fixed effects + 1 residual variance + 3 parameters of heterogeneity matrix.

Random Slope Model: MCMChregress

Some summary results:

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                                 Mean      SD Naive SE Time-series SE
beta.(Intercept)            1.939e+01 0.52623 0.016641       0.015067
beta.weeks                  6.209e+00 0.12768 0.004038       0.003813
VCV.(Intercept).(Intercept) 7.440e+00 1.72988 0.054704       0.044541
VCV.weeks.(Intercept)      -1.087e-01 0.29393 0.009295       0.008187
VCV.weeks.weeks             4.166e-01 0.09724 0.003075       0.002882
sigma2                      1.610e+00 0.12187 0.003854       0.003524


2. Quantiles for each variable:

                                 2.5%        25%        50%      75%     97.5%
beta.(Intercept)            1.835e+01  1.904e+01  1.938e+01 19.73871 2.045e+01
beta.weeks                  5.949e+00  6.124e+00  6.211e+00  6.29395 6.440e+00
VCV.(Intercept).(Intercept) 4.744e+00  6.274e+00  7.262e+00  8.39873 1.184e+01
VCV.weeks.(Intercept)      -7.343e-01 -2.847e-01 -9.296e-02  0.07974 4.462e-01
VCV.weeks.weeks             2.703e-01  3.481e-01  4.031e-01  0.46358 6.440e-01
sigma2                      1.379e+00  1.527e+00  1.605e+00  1.68838 1.866e+00
```

Bayesian Hierarchical Model Summary

For linear mixed models, results from Frequentist and Bayesian analysis are typically similar. Consider a Bayesian analysis when you:

- Have only a small sample size.
- Want to make inference on random effect estimates and the heterogeneity parameters.
- Want to quantify and reflect "all" sources of uncertainties.
- Need to relax model assumptions.
- Want to incorporate *a priori* information through prior distributions.

Some limitations for Bayesian inference

- A lot more computationally intensive.
- Sometimes priors need to be carefully chosen, especially for complex models.

With great power (flexibility) comes great responsibility!!!