Hierarchical linear models

Dr. Jarad Niemi

STAT 544 - Iowa State University

April 20, 2016

Outline

- Mixed effect models
- Seedling weight example
- Non-Bayesian analysis (missing pvalues/CI method)
- Bayesian analysis in Stan
- Compute posterior probabilities and CIs

Notation

Standard notation for mixed-effect models:

$$y = X\beta + Zu + e$$

where

- y is an $n \times 1$ response vector
- X is an $n \times p$ design matrix for fixed effects
- β is a $p \times 1$ unknown fixed effect parameter vector
- Z is an $n \times q$ design matrix for random effects
- u is a $q \times 1$ unknown random effect parameter vector
- e is an $n \times 1$ unknown error vector

Assumptions

$$y = X\beta + Zu + e$$

Typically assume

- $E[u] = E[e] = 0$
- $V[u] = \Omega$ and $V[e] = \Lambda$
- $Cov[u, e] = 0$

These assumptions imply

- $E[y|\beta, \Omega, \Lambda] = X\beta$
- $V[y|\beta, \Omega, \Lambda] = Z\Omega Z' + \Lambda = \Sigma_y$

Common addition assumptions

- $V[e] = \Lambda = \sigma_e^2 \mathrm{I}$,
- $V[u] = \Omega = \mathrm{diag}\{\sigma^2_{u,\cdot}\}$, (or $V[u] = \Omega = \sigma_u^2 \mathrm{I}$ for single source), and
- u and e are normally distributed.

Rewrite as a standard linear regression model

We can rewrite

$$y = X\beta + Zu + e$$

as

$$y = \tilde{X}\tilde{\beta} + e$$

where $\tilde{X}$ is $n \times (p + q)$ with

$$\tilde{X} = [X \; Z]$$

and $\tilde{\beta}$ is a $(p + q) \times 1$ vector with

$$\tilde{\beta} = \begin{bmatrix} \beta \\ u \end{bmatrix}.$$

The fixed and random effects have been concatenated into the same vector.

Hierarchical linear model

Assume $y \sim N(\tilde{X}\tilde{\beta}, \Lambda)$. A Bayesian analysis proceeds by assigning prior distributions to $\tilde{\beta}$ and Λ. In constructing the prior for $\tilde{\beta}$, consider the components β and u separately. Assume

$$\beta \sim N(\beta_0, \Sigma_\beta) \qquad \text{and} \qquad u \sim N(0, \Omega)$$

independently.

For the

- **fixed** effects β, we select β_0 and Σ_β while for the
- **random** effects u, we assign a prior for Ω.

Therefore we have created a hierarchical model for the random effects and thus refer to this as a *hierarchical linear model*.

Summary

These models are referred to as

- mixed-effect models,
- hierarchical linear models, or
- multi-level models.

The parameters for the prior distribution for the

- fixed effects are not learned and
- random effects are learned.

This corresponds to a non-Bayesian analysis learning a variance parameter for random effects.

Seedling weight example

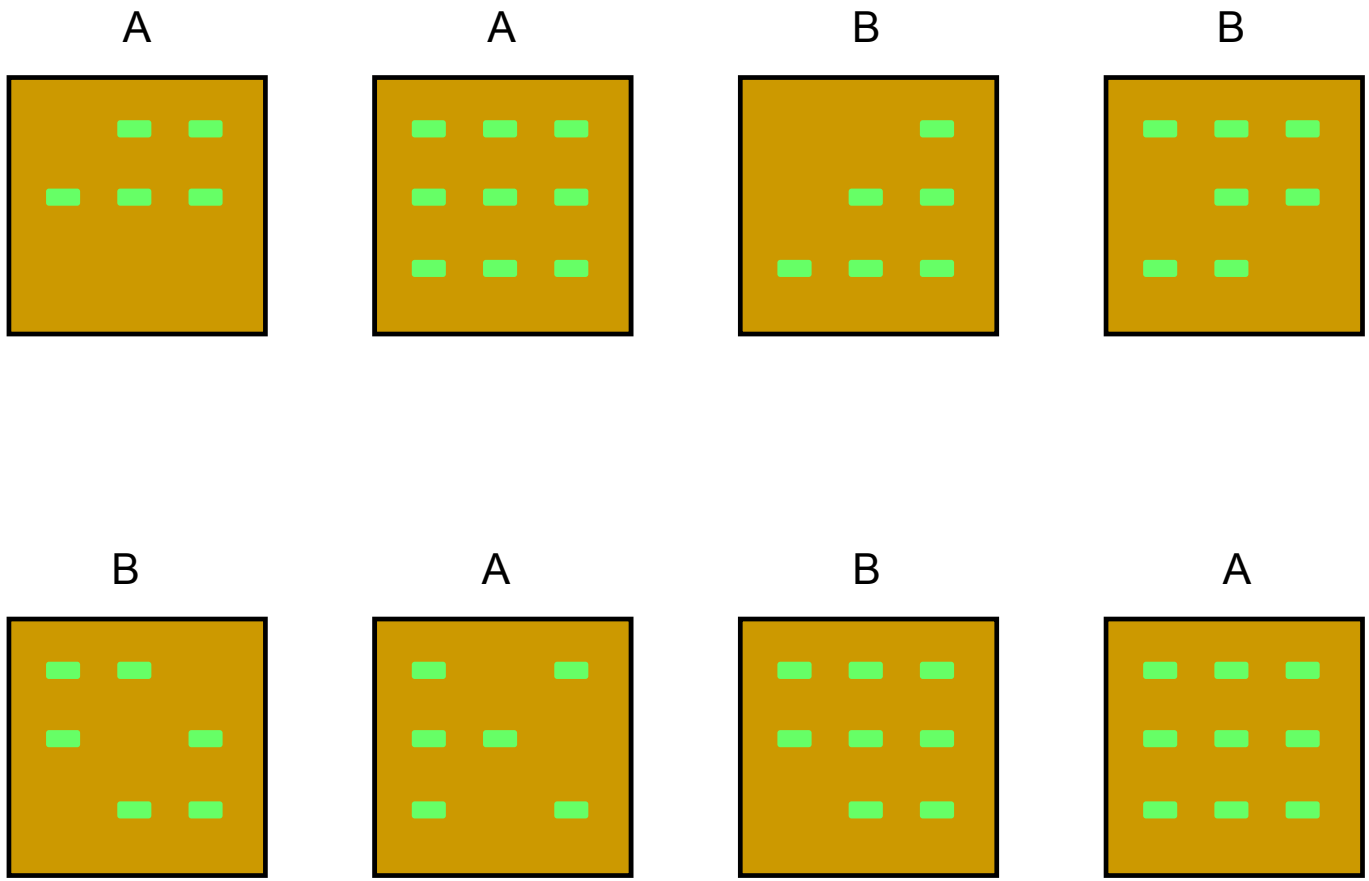
Example taken from Dan Nettleton:

> *Researchers were interested in comparing the dry weight of maize seedlings from two different genotypes (A and B). For each genotype, nine seeds were planted in each of four trays. The eight trays in total were randomly positioned in a growth chamber. Three weeks after the emergence of the first seedling, emerged seedlings were harvested from each tray and, after drying, weighed.*

Assume the missing data (emergence) mechanism is ignorable.

Data: `http://www.public.iastate.edu/~dnett/S511/SeedlingDryWeight2.txt`

A picture



A mixed effect model for seedling weight

Let y_{ijk} be the seedling weight of the

- i^{th} genotype with $i = 1, 2$,
- j^{th} tray $j = 1, 2, 3, 4$ of the i^{th} genotype, and
- k^{th} seedling with $k = 1, \ldots, n_{ij}$.

Then, we assume

$$y_{ijk} = \gamma_i + \tau_{ij} + e_{ijk}$$

where

- $\tau_{ij} \overset{ind}{\sim} N(0, \sigma_\tau^2)$ and, independently,
- $e_{ijk} \overset{ind}{\sim} N(0, \sigma_e^2)$.

The main quantity of interest is the difference in mean seedling weight: $\gamma_2 - \gamma_1$.

As a general mixed effects model

Let X have the following 2 columns

- col1: all ones (intercept) $[\gamma_1]$
- col2: ones if genotype B and zeros otherwise $[\gamma_2 - \gamma_1]$

Let Z have the following 8 columns

- col1: ones if genotype 1, tray 1 and zeros otherwise $[\tau_{11}]$
- col2: ones if genotype 1, tray 2 and zeros otherwise $[\tau_{12}]$
- $\vdots$
- col8: ones if genotype 2, tray 4 and zeros otherwise $[\tau_{24}]$

Then

$$y = X\beta + Zu + e$$

with $u \sim N(0, \sigma^2_\tau \mathrm{I})$ and, independently, $e \sim N(0, \sigma^2_e \mathrm{I})$.

Seedling weight data

```
head(d)
```

```
  Genotype Tray SeedlingWeight
1        A    1              8
2        A    1              9
3        A    1             11
4        A    1             12
5        A    1             10
6        A    2             17
```

```
summary(d)
```

```
 Genotype      Tray       SeedlingWeight
 A:29     Min.   :1.000   Min.   : 6.00
 B:27     1st Qu.:2.750   1st Qu.:10.00
          Median :4.000   Median :14.00
          Mean   :4.554   Mean   :13.88
          3rd Qu.:6.250   3rd Qu.:17.00
          Max.   :8.000   Max.   :24.00
```

```
with(d, table(Genotype, Tray))
```

```
        Tray
Genotype 1 2 3 4 5 6 7 8
       A 5 9 6 9 0 0 0 0
       B 0 0 0 0 6 7 6 8
```

Non-Bayesian analysis

```
m1 = lmer(SeedlingWeight ~ Genotype + (1|Tray), d); summary(m1)


Linear mixed model fit by REML ['lmerMod']
Formula: SeedlingWeight ~ Genotype + (1 | Tray)
   Data: d

REML criterion at convergence: 247.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0928 -0.5697  0.0470  0.5146  3.2347

Random effects:
 Groups   Name        Variance Std.Dev.
 Tray     (Intercept) 11.661   3.415
 Residual              3.543   1.882
Number of obs: 56, groups:  Tray, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   15.289      1.745   8.761
GenotypeB     -3.550      2.469  -1.438

Correlation of Fixed Effects:
          (Intr)
GenotypeB -0.707
```

Why no pvalues?

From https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html (19 May 2006):

> *Users are often surprised and alarmed that the summary of a linear mixed model fit by lmer provides estimates of the fixed-effects parameters, standard errors for these parameters and a t-ratio but no p-values.*
>
> *...*
>
> *Most of the research on tests for the fixed-effects specification in a mixed model begin with the assumption that these statistics will have an F distribution with a known numerator degrees of freedom and the only purpose of the research is to decide how to obtain an approximate denominator degrees of freedom. I don't agree.*
>
> *...*
>
> *For the time being, I would recommend using a Markov Chain Monte Carlo sample (function mcmcsamp) to evaluate the properties of individual coefficients (use HPDinterval or just summary from the "coda" package).*
>
> *Dr. Douglas Bates*

```
confint(m1, method="profile")
```

```
                2.5 %    97.5 %
.sig01       1.837050  5.379221
.sigma       1.560415  2.332764
(Intercept) 11.926526 18.637543
GenotypeB   -8.287734  1.204894
```

```
confint(m1, method="Wald")
```

```
                2.5 %    97.5 %
.sig01             NA        NA
.sigma             NA        NA
(Intercept) 11.868527 18.709148
GenotypeB   -8.388448  1.288046
```

```
confint(m1, method="boot")
```

```
                2.5 %    97.5 %
.sig01       1.529722  5.404522
.sigma       1.542917  2.195049
(Intercept) 11.907640 19.013466
GenotypeB   -8.758632  1.066519
```

Bayesian model

An alternative notation convenient for programming in Stan is

- y_i is the weight for seedling i with $i = 1, \ldots, n$
- $g[i] \in \{1, 2\}$ is the genotype for seedling i
- $t[i] \in \{1, 2, \ldots, 8\}$ is the **unique** tray id for seedling i

Then the model is

$$y_i = \gamma_{g[i]} + \tau_{t[i]} + e_i$$

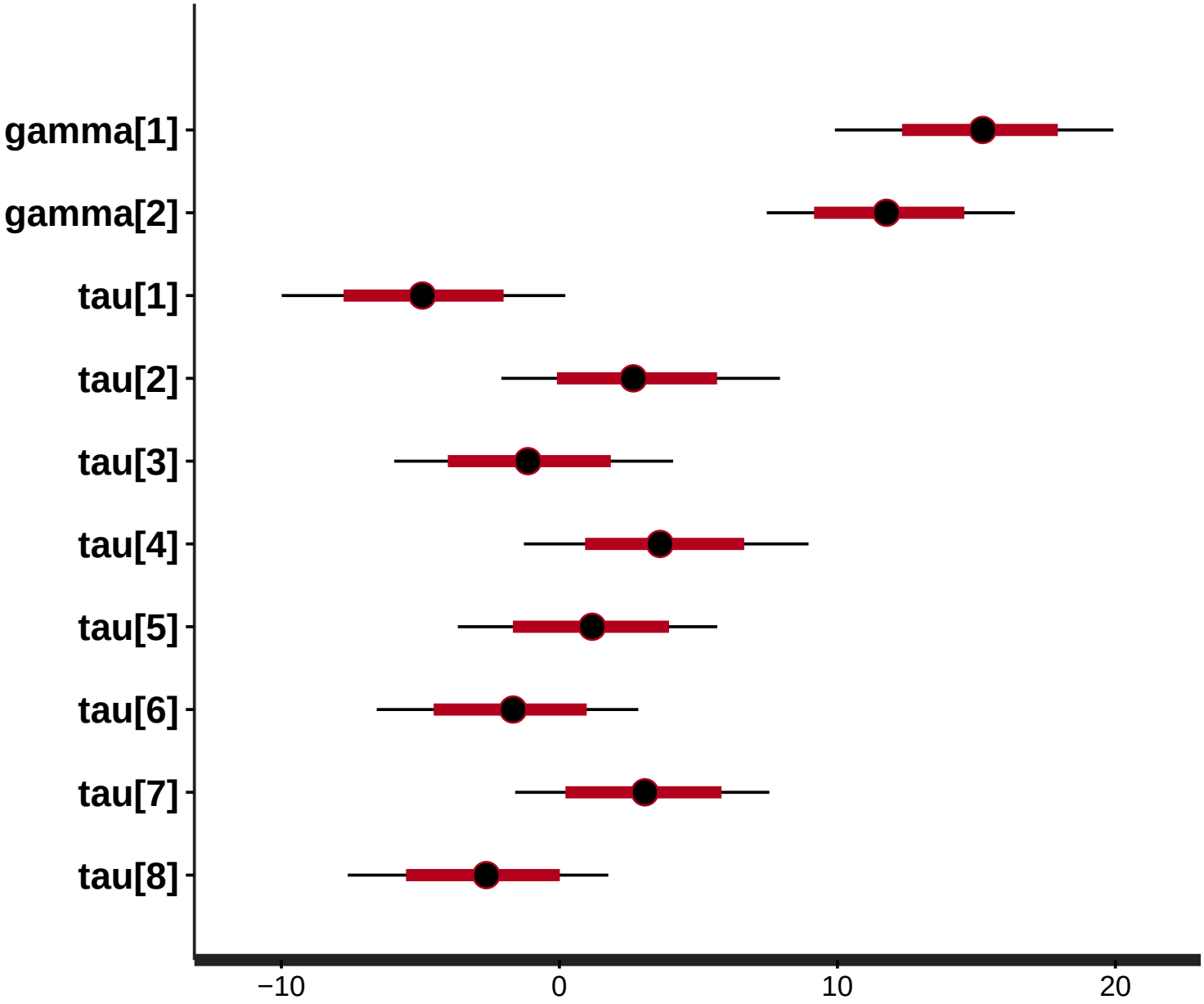
with $e_i \overset{ind}{\sim} N(0, \sigma_e^2)$ and, independently, $\tau_t \overset{ind}{\sim} N(0, \sigma_\tau^2)$ with $t = 1, \ldots, 8$.

Prior: $p(\gamma_1, \gamma_2, \sigma_e, \sigma_u) \propto Ca^+(\sigma_e; 0, 10) Ca^+(\sigma_u; 0, 10)$.

```
stan_model = "
data {
  int<lower=1> n;
  int<lower=1> n_genotypes;
  int<lower=1> n_trays;

  real y[n];
  int genotype[n];
  int tray[n];
}
parameters {
  real gamma[n_genotypes]; // Implicit prior over whole real line
  real tau[n_trays];
  real<lower=0> sigma_e;   // Implicit prior over positive reals
  real<lower=0> sigma_tau; // Implicit prior over positive reals
}

model {
  sigma_e   ~ cauchy(0,10);
  sigma_tau ~ cauchy(0,10);

  tau ~ normal(0,sigma_tau);

  for (i in 1:n) y[i] ~ normal(gamma[genotype[i]]+tau[tray[i]], sigma_e);
}

generated quantities {
  real delta;
  delta <- gamma[2] - gamma[1];
}
"
```

```
m = stan_model(model_code=stan_model)

r = sampling(m,
             list(n = nrow(d),
                  n_genotypes = nlevels(d$Genotype),
                  n_trays     = max(d$Tray),
                  genotype    = as.numeric(d$Genotype),
                  tray        = d$Tray,
                  y           = d$SeedlingWeight),
             c("gamma","tau","sigma_e","sigma_tau","delta"))



SAMPLING FOR MODEL 'd1a96af22c0a2816d3028d7c573a3908' NOW (CHAIN 1).

Chain 1, Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1, Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1, Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1, Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1, Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1, Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1, Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1, Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1, Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1, Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1, Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1, Iteration: 2000 / 2000 [100%]  (Sampling)#
#  Elapsed Time: 0.235 seconds (Warm-up)
#                0.204 seconds (Sampling)
#                0.439 seconds (Total)
#

SAMPLING FOR MODEL 'd1a96af22c0a2816d3028d7c573a3908' NOW (CHAIN 2).

Chain 2, Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2, Iteration:  200 / 2000 [ 10%]  (Warmup)
```

```
Inference for Stan model: d1a96af22c0a2816d3028d7c573a3908.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

           mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
gamma[1]  15.18    0.08 2.40   9.91  13.79  15.23  16.63  19.92   820 1.01
gamma[2]  11.79    0.09 2.37   7.46  10.40  11.76  13.12  16.38   742 1.01
tau[1]    -4.90    0.08 2.48  -9.99  -6.39  -4.93  -3.35   0.21   887 1.00
tau[2]     2.73    0.08 2.43  -2.09   1.24   2.66   4.12   7.93   834 1.01
tau[3]    -1.13    0.08 2.47  -5.94  -2.66  -1.13   0.33   4.09   860 1.00
tau[4]     3.71    0.08 2.45  -1.27   2.21   3.62   5.13   8.96   864 1.00
tau[5]     1.14    0.09 2.43  -3.65  -0.24   1.18   2.55   5.68   770 1.01
tau[6]    -1.72    0.09 2.44  -6.57  -3.09  -1.66  -0.28   2.84   769 1.01
tau[7]     3.06    0.08 2.42  -1.59   1.64   3.07   4.46   7.55   813 1.01
tau[8]    -2.71    0.09 2.43  -7.61  -4.05  -2.63  -1.27   1.76   772 1.01
sigma_e    1.93    0.01 0.20   1.59   1.79   1.91   2.06   2.37  1369 1.00
sigma_tau  4.28    0.07 1.62   2.25   3.21   3.91   4.92   8.37   553 1.01
delta     -3.38    0.14 3.52 -10.09  -5.31  -3.40  -1.51   3.88   667 1.01
lp__     -77.08    0.11 2.78 -83.59 -78.70 -76.66 -75.02 -72.90   615 1.01

Samples were drawn using NUTS(diag_e) at Wed Apr 20 20:47:00 2016.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```


gamma[1]
gamma[2]
tau[1]
tau[2]
tau[3]
tau[4]
tau[5]
tau[6]
tau[7]
tau[8]
−10
0
10
20

Probability that genotype B has greater mean seedling weight than genotype A.

Given our prior, i.e.

$$p(\gamma_1, \gamma_2, \sigma_e, \sigma_u) \propto Ca^+(\sigma_e; 0, 10) Ca^+(\sigma_u; 0, 10),$$

Our posterior probability that genotype B has greater mean seedling weight than genotype A is

$$P(\gamma_2 > \gamma_1 | y) = P(\delta > 0 | y) = E[\mathrm{I}(\delta > 0) | y] = E[\mathrm{I}(\gamma_2 > \gamma_1) | y].$$

If $\delta^{(k)}$ are MCMC samples from $p(\delta|y)$, then

$$\frac{1}{K} \sum_{k=1}^{K} \mathrm{I}(\delta^{(k)} > 0) \stackrel{a.s.}{\to} P(\gamma_2 > \gamma_1 | y)$$

and (if the regularity conditions hold)

$$\frac{1}{K} \sum_{k=1}^{K} \mathrm{I}(\delta^{(k)} > 0) \stackrel{d}{\to} N(P(\gamma_2 > \gamma_1 | y), \sigma^2/K).$$

Probability that genotype B has greater mean seedling weight than genotype A.

The probability is estimated to be

```
library(mcmcse)
delta = extract(r, "delta")$delta
as.data.frame(mcse(delta>0))
```

```
      est          se
1 0.12825 0.006601652
```

A point estimate (posterior median) and a 95% credible interval are calculated below:

```
      q        est         se
1 0.025 -10.112134 0.23055408
2 0.500  -3.396796 0.06398783
3 0.975   3.878877 0.43900686
```

Prediction for a new comparison

The real question is whether this idea generalizes, i.e. is true for other representatives of these genotypes. Let $\tilde{y}_A$ and $\tilde{y}_B$ be some future observation of seedling weight (on the same tray) for genotype A and B, respectively. We might be interested in

$$P(\tilde{y}_B > \tilde{y}_A|y) = P(\tilde{\delta} > 0|y) = E[\mathrm{I}(\tilde{\delta} > 0)|y]$$

where $\tilde{\delta} = \tilde{y}_B - \tilde{y}_A$. If $\tilde{\delta}^{(k)} = \tilde{y}_B^{(k)} - \tilde{y}_A^{(k)}$ is a sample from the posterior predictive distribution, then we can estimate this probability via

$$\frac{1}{K}\sum_{k=1}^{K} \mathrm{I}(\tilde{\delta}^{(k)} > 0)$$

and have a similar LLN and CLT (if regularity conditions hold).

Prediction for a new comparison

Assuming $\tilde{y}_A^{(k)}$ and $\tilde{y}_B^{(k)}$ are independent conditional on γ_1, γ_2, and σ_e, then

$$\tilde{\delta} = \tilde{y}_B - \tilde{y}_A \sim N(\gamma_2 - \gamma_1, 2\sigma_e^2)$$

and

$$p(\tilde{\delta}|y) \int N(\tilde{\delta}; \gamma_2 - \gamma_1, 2\sigma_e^2) p(\gamma_1, \gamma_2, \sigma_e|y) d\gamma_1 d\gamma_2 d\sigma_e$$

```
samps = extract(r, c("gamma","sigma_e"))
gamma1 = samps['gamma']$gamma[,1]
gamma2 = samps['gamma']$gamma[,2]
sigmae = samps['sigma_e']$sigma_e
tilde_delta = rnorm(length(gamma1), gamma2-gamma1, sqrt(2)*sigmae)
as.data.frame(mcse(tilde_delta>0))
```

```
     est          se
1 0.2035 0.007689082
```

```
ddply(data.frame(q=c(.025,.5,.975)), .(q), function(x) as.data.frame(mcse.q(tilde_delta, q=x$q)))
```

```
      q        est         se
1 0.025 -11.915533 0.19823587
2 0.500  -3.422386 0.08704035
3 0.975   5.259084 0.27868462
```

Extensions

Consider the model

$$y_i = \gamma_{g[i]} + \tau_{t[i]} + e_i$$

and the following modeling assumptions:

- $\gamma_g \overset{ind}{\sim} N(\mu, \sigma^2_\gamma)$ and learn μ, σ_γ
- $\tau_t \overset{ind}{\sim} La(0, \sigma^2_\tau)$
- $\gamma_g \overset{ind}{\sim} La(\mu, \sigma^2_\gamma)$
- $e_i \overset{ind}{\sim} La(0, \sigma^2_e)$
- $e_i \overset{ind}{\sim} t_\nu(0, \sigma^2_e)$

From a Bayesian perspective these changes do not affect the approach to inference.