

# 19. Maximum Likelihood Estimation for the General Linear Model

# Likelihood Functions

- Suppose  $f(\mathbf{y}|\boldsymbol{\theta})$  is the probability density function (*pdf*) or probability mass function (*pmf*) of a random vector  $\mathbf{y}$ , where  $\boldsymbol{\theta}$  is a  $k \times 1$  vector of parameters.
- Given a value of the parameter vector  $\boldsymbol{\theta}$ ,  $f(\mathbf{y}|\boldsymbol{\theta})$  is a real-valued function of  $\mathbf{y}$ .
- The likelihood function  $L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})$  is a real-valued function of  $\boldsymbol{\theta}$  for a given value of  $\mathbf{y}$ .

## Example

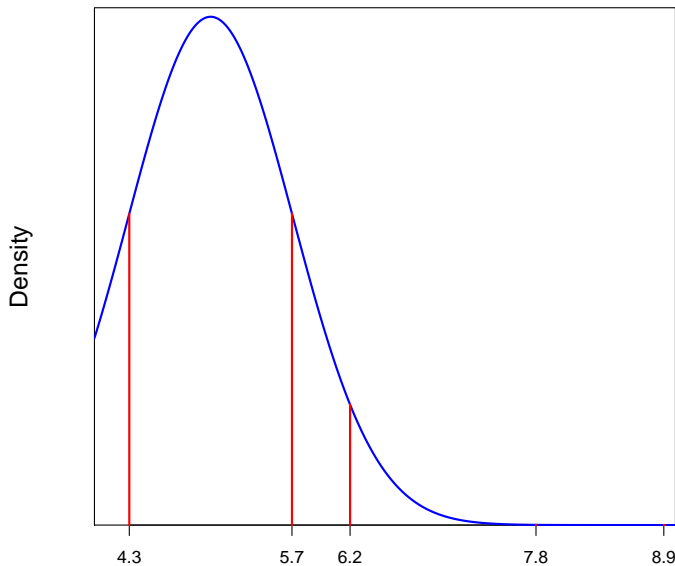
Suppose  $\mathbf{y} = [y_1, y_2, y_3, y_4, y_5]'$   $\sim N(\mu \mathbf{1}_{5 \times 1}, \sigma^2 \mathbf{I}_{5 \times 5})$ .

Let  $\boldsymbol{\theta} = [\mu, \sigma^2]'$ .

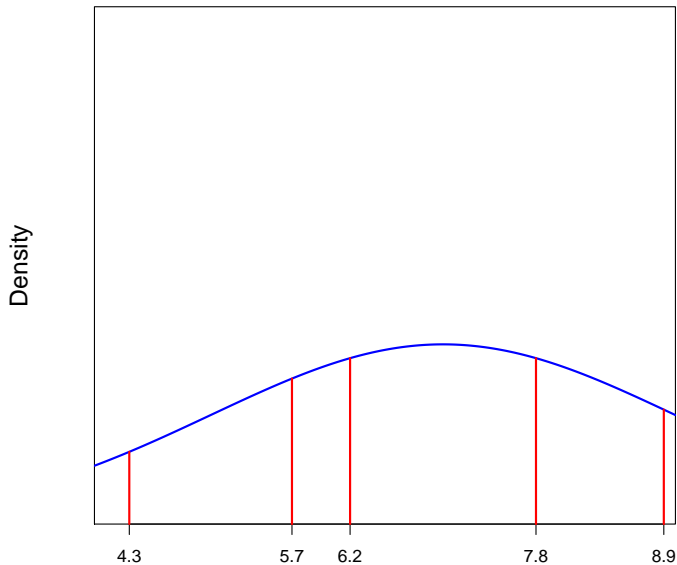
Then  $L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^5 \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\}$ .

Suppose  $\mathbf{y} = [7.8, 5.7, 6.2, 8.9, 4.3]'$ .

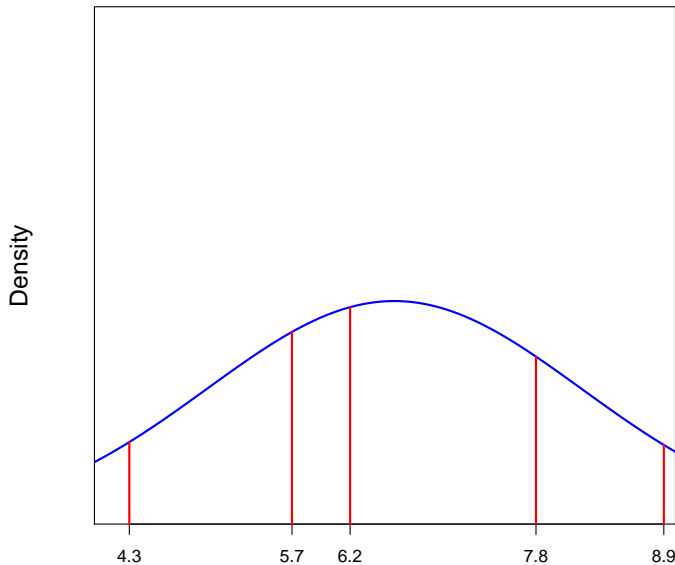
$$L([5, 0.5]'|\mathbf{y}) = 0.00000000000005$$



$$L([7, 4]'|\mathbf{y}) = 0.000056$$



$$L([6.58, 2.5976]'|\mathbf{y}) = 0.000076$$



# Maximum Likelihood Estimators

- For any potential observed vector of values  $\mathbf{y}$ , define  $\hat{\theta}(\mathbf{y})$  to be a parameter value at which  $L(\theta|\mathbf{y})$  attains its maximum value.
- If  $\mathbf{y}$  is a random vector distributed according to  $f(\mathbf{y}|\theta)$ , then the random variable  $\hat{\theta}(\mathbf{y})$  is called a

*Maximum Likelihood Estimator (MLE) of  $\theta$ .*

## Invariance Property of MLEs

The MLE of a function of  $\theta$ , say  $g(\theta)$ , is the function evaluated at the MLE of  $\theta$ :

$$\widehat{g(\theta)} = g(\hat{\theta}).$$



# Log Likelihood Functions

- It is often more convenient to work with the log likelihood function  $\ell(\boldsymbol{\theta}|\mathbf{y}) = \log L(\boldsymbol{\theta}|\mathbf{y})$ .
- The maximizers of  $\ell(\boldsymbol{\theta}|\mathbf{y})$  and  $L(\boldsymbol{\theta}|\mathbf{y})$  are the same because  $u < v \iff \log(u) < \log(v)$  for  $u, v > 0$ .

# The Score Function

- If  $\ell(\boldsymbol{\theta}|\mathbf{y})$  is differentiable, the score function is

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \theta_k} \end{bmatrix}.$$

# The Score Equations

- The score equations are

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0} \iff \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \theta_j} = 0 \quad \forall j = 1, \dots, k.$$

- One strategy for obtaining an MLE is to find a solution or solutions to the score equations and verify that at least one such solution maximizes  $\ell(\boldsymbol{\theta}|\mathbf{y})$  over the parameter space.

# Gauss-Markov Linear Model with Normal Errors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{bmatrix}$$

$$f(\mathbf{y}|\boldsymbol{\theta}) = \frac{\exp\left\{\frac{-1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}}{(2\pi)^{n/2}|\sigma^2\mathbf{I}|^{1/2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

The score function is

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} (\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \\ \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^4} - \frac{n}{2\sigma^2} \end{bmatrix}.$$

The score equations are

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0} \iff \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \quad \sigma^2 = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n}.$$

A solution to the score equations is

$$\left[ \begin{array}{c} \hat{\beta} \\ \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} \end{array} \right],$$

where  $\hat{\beta}$  is any solution to the normal equations.

For such a solution to the score equations to be an MLE, we need to show that the likelihood is maximized at such a solution.

We already know that any solution to the normal equations minimizes  $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$  over  $\mathbf{b} \in \mathbb{R}^p$ . Thus,

$$\forall \sigma^2 > 0, \ell \left( \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \sigma^2 \end{bmatrix} \middle| \mathbf{y} \right) \geq \ell \left( \begin{bmatrix} \mathbf{b} \\ \sigma^2 \end{bmatrix} \middle| \mathbf{y} \right) \quad \forall \mathbf{b} \in \mathbb{R}^p .$$

To see that  $\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n}$  is the MLE of  $\sigma^2$ , note that

$$\frac{\partial \ell \left( \begin{bmatrix} \hat{\beta} \\ \sigma^2 \end{bmatrix} \middle| y \right)}{\partial \sigma^2} = 0$$

has  $\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n}$  as its only solution. Furthermore,

$$\frac{\partial^2 \ell \left( \begin{bmatrix} \hat{\beta} \\ \sigma^2 \end{bmatrix} \middle| y \right)}{(\partial \sigma^2)^2} \bigg|_{\sigma^2 = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n}} < 0.$$



We have shown that  $\sigma^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n}$  is the only extreme point in the interior of the parameter space and that a local maximum occurs at this point.

Could the likelihood increase without bound as  $\sigma^2$  approaches a boundary of the parameter space (0 or  $\infty$ )?

No because if so, there would have to be a local minimum somewhere in the interior of the parameter space.

It follows that  $\frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n}$  is the global maximizer of

$$\ell \left( \left[ \begin{array}{c} \hat{\beta} \\ \sigma^2 \end{array} \right] \middle| y \right).$$

We have established that

$$\begin{bmatrix} \hat{\beta} \\ \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n} \end{bmatrix} \text{ is an MLE of } \boldsymbol{\theta} = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}.$$

Thus, if  $\mathbf{C}\beta$  is estimable, the MLE of  $\mathbf{C}\beta$  is  $\mathbf{C}\hat{\beta}$  (by the Invariance Property of MLEs), which is the BLUE of  $\mathbf{C}\beta$ .

Note that the MLE of  $\sigma^2$  is not the unbiased estimator we have been using.

$$E \left[ \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} \right] = E \left( \frac{SSE}{n} \right) = \frac{n-r}{n} \sigma^2 < \sigma^2.$$

Thus, the MLE of  $\sigma^2$  underestimates  $\sigma^2$  on average.

Now consider the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where  $\boldsymbol{\Sigma}$  is a positive definite covariance matrix whose entries depend on unknown parameters in some vector  $\boldsymbol{\gamma}$ .

For example, suppose  $\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$ ,  $\boldsymbol{\gamma} = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}$ ,

where  $\sigma^2 > 0$  and  $\rho \in (-1, 1)$ .

In general, we have

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \gamma \end{bmatrix}, \quad f(\mathbf{y}|\boldsymbol{\theta}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}},$$

and

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{n}{2}\log(2\pi).$$

We know that for any positive definite covariance matrix  $\Sigma$ ,

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

is minimized over  $\boldsymbol{\beta} \in \mathbb{R}^p$  by

$$\hat{\boldsymbol{\beta}}_{\Sigma} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{y}.$$

Thus, for any  $\gamma$  such that  $\Sigma$  is a positive definite covariance matrix,

$$\ell \left( \left[ \begin{array}{c} \hat{\boldsymbol{\beta}}_{\Sigma} \\ \gamma \end{array} \right] \middle| \mathbf{y} \right) \geq \ell \left( \left[ \begin{array}{c} \boldsymbol{\beta} \\ \gamma \end{array} \right] \middle| \mathbf{y} \right) \quad \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

## Profile Log Likelihood

We define the *profile log likelihood* for  $\gamma$  to be

$$\ell^*(\gamma \mid \mathbf{y}) = \ell \left( \left[ \begin{array}{c} \hat{\beta}_{\Sigma} \\ \gamma \end{array} \right] \middle| \mathbf{y} \right).$$

The MLE of  $\theta$  is

$$\hat{\theta} = \left[ \begin{array}{c} \hat{\beta}_{\hat{\Sigma}} \\ \hat{\gamma} \end{array} \right]$$

where  $\hat{\gamma}$  is a maximizer of  $\ell^*(\gamma \mid \mathbf{y})$  and  $\hat{\Sigma}$  is obtained by replacing  $\gamma$  in  $\Sigma$  with  $\hat{\gamma}$ .

In general, numerical methods are required to find  $\hat{\gamma}$ , a maximizer of  $\ell^*(\gamma \mid \mathbf{y})$ .

Details of numerical maximization techniques are discussed in STAT520.

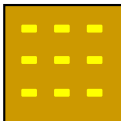


## An Example

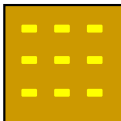
Researchers were interested in comparing the dry weight of maize seedlings from two different genotypes. For each genotype, nine seeds were planted in each of four trays. The eight trays in total were randomly positioned in a growth chamber. Three weeks after the emergence of the first seedling, emerged seedlings were harvested from each tray and individually weighed after drying to obtain one dry weight for each seedling. Although nine seeds were planted in each tray, fewer than nine seedlings emerged in many of the trays.

# Planted Seeds

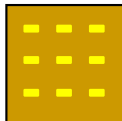
Genotype 1



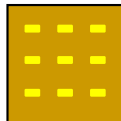
Genotype 1



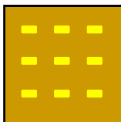
Genotype 2



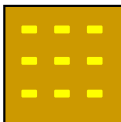
Genotype 2



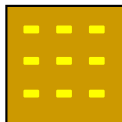
Genotype 2



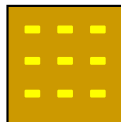
Genotype 1



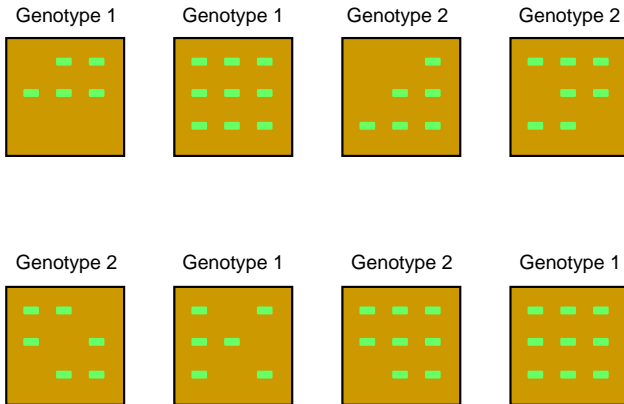
Genotype 2



Genotype 1



# Emergenced Seedlings



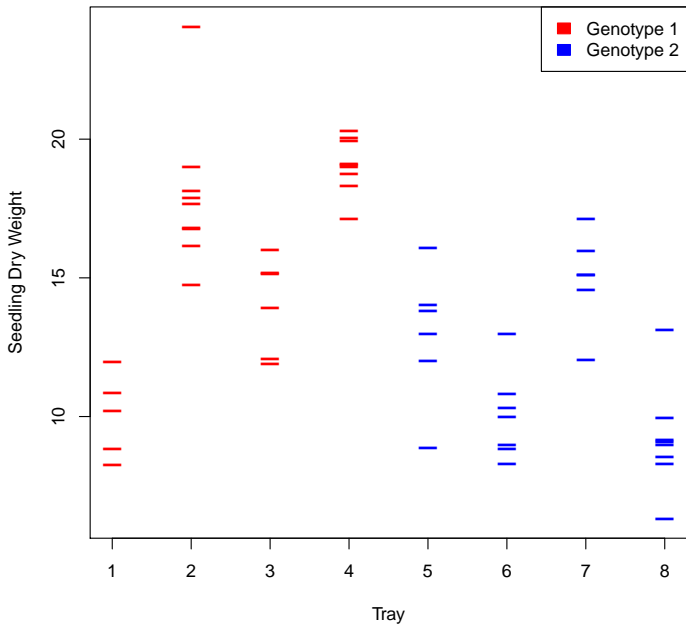
```
> d=read.delim(  
+ "https://dnett.github.io/S510/SeedlingDryWeight2.txt")  
> d
```

	Genotype	Tray	Seedling	SeedlingWeight
1	1	1	1	8
2	1	1	2	9
3	1	1	3	11
4	1	1	4	12
5	1	1	5	10
6	1	2	1	17
7	1	2	2	17
8	1	2	3	16
9	1	2	4	15
10	1	2	5	19
11	1	2	6	18
12	1	2	7	18
13	1	2	8	18
14	1	2	9	24
15	1	3	1	12

16	1	3	2	12
17	1	3	3	16
18	1	3	4	15
19	1	3	5	15
20	1	3	6	14
21	1	4	1	17
22	1	4	2	20
23	1	4	3	20
24	1	4	4	19
25	1	4	5	19
26	1	4	6	18
27	1	4	7	20
28	1	4	8	19
29	1	4	9	19
30	2	5	1	9
31	2	5	2	12
32	2	5	3	13
33	2	5	4	16
34	2	5	5	14

35	2	5	6	14
36	2	6	1	10
37	2	6	2	10
38	2	6	3	9
39	2	6	4	8
40	2	6	5	13
41	2	6	6	9
42	2	6	7	11
43	2	7	1	12
44	2	7	2	16
45	2	7	3	17
46	2	7	4	15
47	2	7	5	15
48	2	7	6	15
49	2	8	1	9
50	2	8	2	6
51	2	8	3	8
52	2	8	4	8
53	2	8	5	13
54	2	8	6	9
55	2	8	7	9
56	2	8	8	10

```
> plot(d[,2],d[,4]+rnorm(56,0,.2),  
+      xlab="Tray",ylab="Seedling Dry Weight",  
+      col=2*d[,1],pch="-",cex=2)  
> legend("topright",c("Genotype 1","Genotype 2"),  
+      fill=c(2,4),border=c(2,4))
```





## A Model for the Seedling Dry Weights

Let  $y_{ijk}$  be the dry weight of the  $k$ th seedling in the  $j$ th tray for genotype  $i$ .

Suppose

$$y_{ijk} = \mu_i + t_{ij} + e_{ijk},$$

where  $\mu_1$  and  $\mu_2$  are unknown constants,

$$t_{ij} \sim N(0, \sigma_t^2), \quad e_{ijk} \sim N(0, \sigma_e^2),$$

and all random terms are independent.

```
> d$Genotype=factor(d$Genotype)
>
> library(nlme)
>
> lme(SeedlingWeight~Genotype,random=~1|Tray,method="ML",
+     data=d)
```

Linear mixed-effects model fit by maximum likelihood

Data: d

Log-likelihood: -126.3709

Fixed: SeedlingWeight ~ Genotype

(Intercept)      Genotype2

15.301832      -3.567017

Random effects:

Formula: ~1 | Tray

(Intercept) Residual

StdDev:      2.932294    1.88247

Number of Observations: 56

Number of Groups: 8

```
> library(lme4)
>
> lmer(SeedlingWeight ~ Genotype + (1 | Tray), REML=F, data=d)
```

Linear mixed model fit by maximum likelihood ['lmerMod']

Formula: SeedlingWeight ~ Genotype + (1 | Tray)

Data: d

AIC	BIC	logLik	deviance
260.7418	268.8432	-126.3709	252.7418

Random effects:

Groups	Name	Std.Dev.
Tray	(Intercept)	2.932
Residual		1.882

Number of obs: 56, groups: Tray, 8

Fixed Effects:

(Intercept)	Genotype2
15.302	-3.567