Maximum Likelihood Estimation for the General Linear Model

Likelihood Functions

- Suppose $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the probability density function (pdf) or probability mass function (pmf) of a random vector $\boldsymbol{y}$, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters.

- Given a value of the parameter vector $\boldsymbol{\theta}$, $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a real-valued function of $\boldsymbol{y}$.

- The likelihood function $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ is a real-valued function of $\boldsymbol{\theta}$ for a given value of $\boldsymbol{y}$.

Example

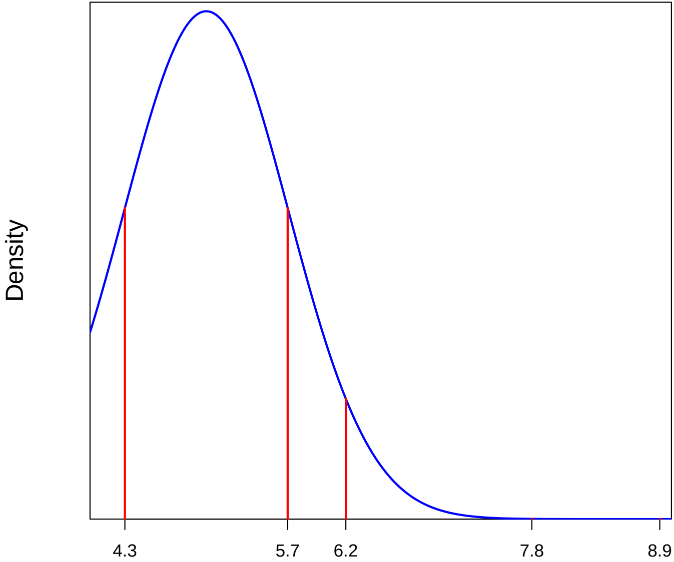
Suppose $\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5]' \sim N(\mu \mathbf{1}_{5\times 1}, \sigma^2 \boldsymbol{I}_{5\times 5})$.

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

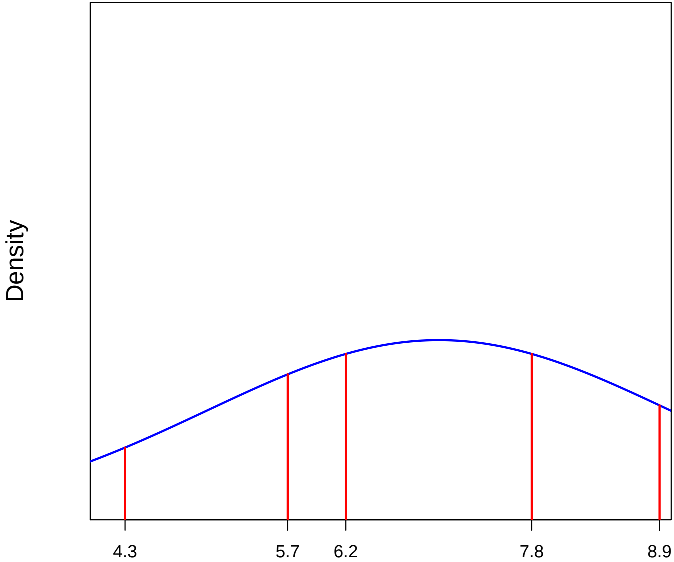
Then $L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\}$.

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

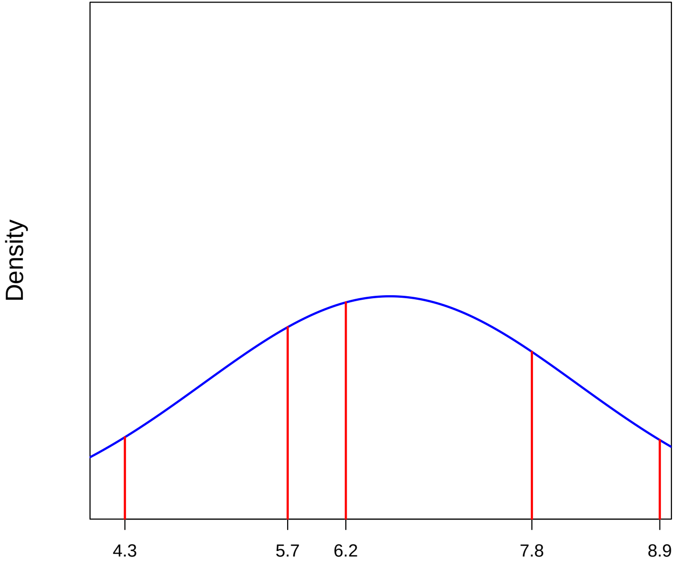
$L([5, 0.5]'|\boldsymbol{y}) = 0.000000000005$



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



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



Maximum Likelihood Estimators

- For any potential observed vector of values $\boldsymbol{y}$, define $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ to be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum value.

- If $\boldsymbol{y}$ is a random vector distributed according to $f(\boldsymbol{y}|\boldsymbol{\theta})$, then the random variable $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ is called a

  Maximum Likelihood Estimator (MLE) of $\boldsymbol{\theta}$.

Invariance Property of MLEs

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

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

Log Likelihood Functions

- It is often more convenient to work with the log likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log L(\boldsymbol{\theta}|\boldsymbol{y})$.

- The maximizers of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ and $L(\boldsymbol{\theta}|\boldsymbol{y})$ are the same because $u < v \iff \log(u) < \log(v)$ for $u, v > 0$.

The Score Function

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

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

The Score Equations

- The score equations are

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

- 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 $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ over the parameter space.

Gauss-Markov Linear Model with Normal Errors

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

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

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

The score function is

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

The score equations are

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

A solution to the score equations is

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n} \end{bmatrix},$$

where $\hat{\boldsymbol{\beta}}$ 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 $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})$ over $\boldsymbol{b} \in \mathbb{R}^p$. Thus,

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

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

$$\frac{\partial \ell\left(\left[\begin{array}{c}\hat{\boldsymbol{\beta}}\\ \sigma^2\end{array}\right] \,\middle|\, \boldsymbol{y}\right)}{\partial \sigma^2} = 0$$

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

$$\left.\frac{\partial^2 \ell\left(\left[\begin{array}{c}\hat{\boldsymbol{\beta}}\\ \sigma^2\end{array}\right] \,\middle|\, \boldsymbol{y}\right)}{(\partial \sigma^2)^2}\right|_{\sigma^2=\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}} < 0.$$

We have shown that $\sigma^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$ 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 σ^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 $\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$ is the global maximizer of

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

We have established that

$$\left[\begin{array}{c} \hat{\boldsymbol{\beta}} \\ \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n} \end{array}\right] \text{ is an MLE of } \boldsymbol{\theta} = \left[\begin{array}{c} \boldsymbol{\beta} \\ \sigma^2 \end{array}\right].$$

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

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

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

Thus, the MLE of σ^2 underestimates σ^2 on average.

Now consider the general linear model

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

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

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

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

and

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

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

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

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

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

Thus, for any γ such that $\boldsymbol{\Sigma}$ is a positive definite covariance matrix,

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

Profile Log Likelihood

We define the *profile log likelihood* for γ to be

$$\ell^*(\gamma \mid \boldsymbol{y}) = \ell\left(\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} \\ \gamma \end{bmatrix} \middle| \, \boldsymbol{y} \right).$$

The MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}} \\ \hat{\gamma} \end{bmatrix}$$

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

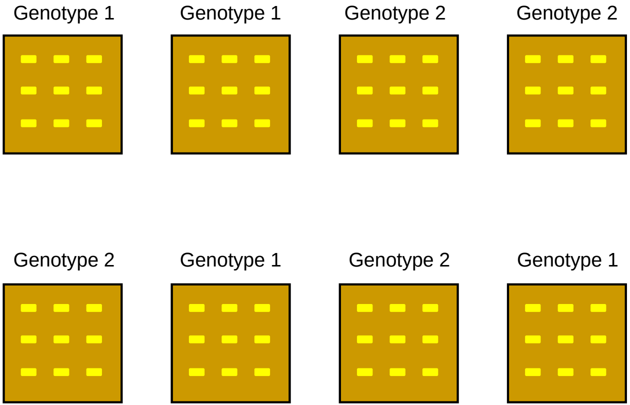
In general, numerical methods are required to find $\hat{\gamma}$, a maximizer of $\ell^*(\boldsymbol{\gamma} \mid \boldsymbol{y})$.

Details of numerical maximization techniques are discussed in STAT520.

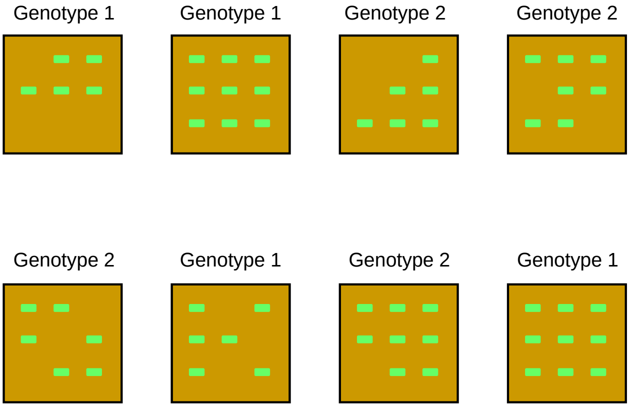
An Example

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



Emerged Seedlings

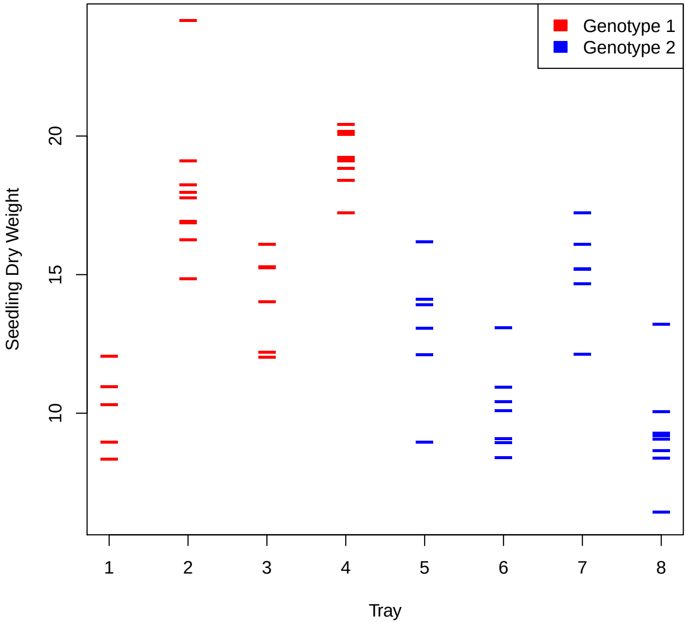


```
> d=read.delim(
+ "https://dnett.github.io/S510/SeedlingDryWeight2.txt")
> d
   Genotype Tray Seedling SeedlingWeight
1         1    1        1              8
2         1    1        2              9
3         1    1        3             11
4         1    1        4             12
5         1    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))
```



Genotype 1
Genotype 2
Seedling Dry Weight
10
15
20
1
2
3
4
5
6
7
8
Tray

A Model for the Seedling Dry Weights

Let y_{ijk} be the dry weight of the kth seedling in the jth tray for genotype i.

Suppose

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

where μ_1 and μ_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
```