

Overhead Lecture 4

Peter von Rohr

2020-10-16

Import Data in R

Reading Data into R:

`read.csv(...)`
`read.csv2(...)`

CSV - comma separated values

semi-colon ';' as separator of columns
comma is used as decimal delimiter

- comma as separator for columns
- point is used as decimal delimiter

US: $10^{-1} = \frac{1}{10} = 0.1$

Option 'stringsAsFactors' in `read.csv2(...)`

- In R versions newer than 4.0, the default for the option 'stringsAsFactors' is set to FALSE.
- If the option 'stringsAsFactors' = TRUE, then columns with string data are converted to a datatype that is called 'FACTOR'. Factors are important when using data in a linear model.

Recap Genetic Model

is called **FACTOR**. Factors are implemented using data in a linear model.

- In R versions older than 4.0 it was the other way around
⇒ default: `stringsAsFactors = TRUE`

Recap:

- Genetic Model: $y = \mu + u + e$
to decompose phenotypic observation y .
- Fixed Linear Effect Model (Regression Model)
 $y = \mu + \beta + e$ → random error term
fixed effect

Example: What different factors influence clinical mastitis (CM) in sows?

Dataset:

Sows	CM	Length of gestating interval
1	yes	
2	no	
3	no	
4	yes	
⋮	⋮	
N	yes	

Fixed Linear Effects Model

Dataset:	Sows	CM	Length of gestation interval	...
	1	yes		
	2	no		
	3	no		
	4	yes		
	⋮	⋮		
	N	yes		

response variable
 y

influence factors
independent variables
or predictors.

$$y = \mu + \beta_{\text{length gestation}} + \beta_{\text{parmia}} + \dots + \beta_{\text{treatment}} + e$$

insert the ~~numbers~~ information from the dataset, with that the unknown parameters β can be estimated

↳ R: `lm(...)`

Limitation with Fixed Linear Effect Models (FLEM) is that factors cannot be "random". A FLEM contains only fixed effects, except for the random residual e .

Mixed Linear Effects Model

to R. LmL ...)

- Limitation with Fixed Linear Effect Models (FLEM) is that factors cannot be "random". A FLEM contains only fixed effects, except for the random residual term (e).
- In Genetic Model, breeding values (u) have to be treated as random effects, because we had seen that they are defined as deviations and they have a pre-defined variance-covariance structure.

⇒ Use a Mixed Linear Effect Model (MLEM).

MLEM can accommodate additional random effects, besides the random residual error term.

• Model: $y = X\beta + Zu + e$

$X\beta$ → fixed effects (sex, herd, season...)
 Zu → random breeding values
 e → random residuals error terms

! In a MLEM, the expected values and the variance-covariance structure of the random terms must be specified

Model Specification

(sex, herd, season...)

- In a MLEH, the expected values and the variance-covariance structure of the random terms must be specified
- Random terms:
 - u → breeding values
 - e → random error terms
 - y → random observations

$$\begin{cases} E[u] = E \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{bmatrix} = \begin{bmatrix} E[u_1] \\ E[u_2] \\ \vdots \\ E[u_q] \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_q \rightarrow \text{vector of length } q \text{ with all zeros} \\ E[e] = \mathbf{0}_n \\ E[y] = E[X\beta + Zu + e] = E[X\beta] + E[Zu] + E[e] = X\beta \end{cases}$$

Remember, u stands for a vector of length q ,
 e stands for a vector of length n and y is a vector of length n .

$$E \begin{bmatrix} y \\ Zu \\ e \end{bmatrix} = \begin{bmatrix} X\beta \\ \mathbf{0}_n \\ \mathbf{0}_n \end{bmatrix}$$

... Variance-Covariance Structure ... (end of lecture)

Given: Example Data Set on WVG
 ... Mixed Linear Effect Model for

Model Components

Variance-Covariance Structure ... (end of lecture)

- Given: Example Data Set on WVG
- Goal: Specify Mixed Linear Effect Model for the given data-set. (Data set is only used for explanatory purposes, for real analyses, we need larger datasets!)

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

vector of observations,
trait, response variable
⇒ Example WVG

① $y = \begin{bmatrix} 45 \\ 2.9 \\ 10.9 \\ 13.5 \end{bmatrix}$ is known from the data set.

② vector β corresponds to the vector of fixed effects. ⇒ Example: herd as fixed effect.
In the data set there are two herds, hence the vector β is of length 2.
 $\beta = [\text{Herd}] \rightarrow$ 'average' influence of herd 1 on ...

Fixed and Random Effects

(2) vector β corresponds to the vector of fixed effects. \rightarrow Example: herd as fixed effect.
In the data set there are two herds, hence the vector β is of length 2

$$\beta = \begin{bmatrix} \beta_{\text{herd 1}} \\ \beta_{\text{herd 2}} \end{bmatrix}$$

$\beta_{\text{herd 1}}$ \rightarrow 'average' influence of herd 1 on WWS
 $\beta_{\text{herd 2}}$ \rightarrow 'average' influence of herd 2 on WWS

unknown and has to be estimated from the data. What is known from the data set is which animal is in which herd. And this information will be used to construct the design matrix X

(3) Vector u of breeding values. The vector u contains breeding values for all animals given in the data set. Hence, also for those animals without observations but with offspring in the data set.

Example:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}$$

(Example NA stands for not available and is the encoding for a missing data point)

Combining Components

vector u of breeding values is unknown and must be predicted from the data. The connection between which breeding value is associated with which observation is given by the design matrix Z .

(+) vector e of random error terms. The vector e is of length n which is the same length as the vector of observations y .

$$e = \begin{bmatrix} e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

Combining (1) - (+) :

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

Diagram illustrating the combination of components:

The vector y is calculated as $y = X\beta + Zu + e$.

The vector y is $\begin{bmatrix} 4.5 \\ 2.9 \\ 3.9 \\ 3.5 \end{bmatrix}$.

The matrix X is $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$. The vector β is $\begin{bmatrix} \mu \\ \mu \end{bmatrix}$. The matrix X is labeled "Plants".

The matrix Z is $\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$. The vector u is $\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}$. The matrix Z is labeled "Plants".

The vector e is $\begin{bmatrix} e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$.

The equation $0 \cdot u_1 + 0 \cdot u_2 + 1 \cdot u_3 + 0 \cdot u_4 + 0 \cdot u_5 + 0 \cdot u_6$ is shown above the first row of Z .

Numerical Example

Combining (1) - (4):

$$\begin{bmatrix} 4.5 \\ 2.9 \\ 3.9 \\ 3.5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_{lead1} \\ \beta_{lead2} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

$y = X\beta + Zu + e$

$y_3 = 4.5 = 1 \cdot \beta_{lead1} + 0 \cdot \beta_{lead2} + u_6 + e_3$
 calf number 3 is in lead 1

$y_4 = 2.9 = 0 \cdot \beta_{lead1} + 1 \cdot \beta_{lead2} + u_4 + e_4$

- Genetic Model: $y_i = \mu + u_i + e_i$

.. vectors β and u contained unknowns, and we want to estimate β and to predict u using properties described by BLUP, we get

Solutions

the Model: $y_i = \mu + u_i + e_i$

- Vectors β and u contained unknowns, and we want to estimate β and to predict u
- Using properties described by BLUP, we get to estimates $\hat{\beta}$ for the unknowns β :

$$\left. \begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \hat{u} &= G Z^T V^{-1} (y - X \hat{\beta}) \end{aligned} \right\} \begin{array}{l} \text{theory, but not usable} \\ \text{in practice} \end{array}$$

- Mixed Model Equations to get results for $\hat{\beta}$ and \hat{u}

General:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

X, Z are given design matrices

$R = \text{var}(e)$ variance-covariance matrix of e

$G = \text{var}(u)$ variance-covariance matrix of u

- Variance Structure of BLUE

$\text{var}(\hat{\beta}) = (X^T R^{-1} X)^{-1}$ $\text{var}(\hat{u}) = G - G Z^T (X^T R^{-1} X + Z^T R^{-1} Z + G^{-1})^{-1} X^T R^{-1} X G$

Variance Structure

$G = \text{var}(u)$ variance-covariance matrix of u

- Variance Structure of MLEM
 - $R = \text{var}(e) = I_n \cdot \sigma_e^2$
 - I_n : $n \times n$ identity matrix
 - σ_e^2 : scalar number "error variance"
 - $$I_n = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$
 - multiplying any matrix M with I_n does change M : $I_n \cdot M = M$
- Matrix G : $G = A \cdot \sigma_u^2$
 - σ_u^2 : genetic additive variance (scalar number)
 - A : numerator relationship matrix (Verwandtschaftsmatrix)
- Meaning of the notation "var(u)" when u is a vector of length q :
 - $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{bmatrix}$; ~~$\text{var}(u) = \begin{bmatrix} \text{var}(u_1) \\ \text{var}(u_2) \\ \text{var}(u_3) \\ \text{var}(u_q) \end{bmatrix}$~~
- Meaning of $\text{var}(x)$, if x is a scalar random

Genetic Variance

Meaning of $\text{var}(x)$, if x is a scalar random variable

For a vector u of random variables:

$\text{var}(u)$ is a $q \times q$ variance-covariance matrix

$$G = \text{var}(u) = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \text{cov}(u_1, u_3) & \dots & \text{cov}(u_1, u_q) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \text{cov}(u_2, u_3) & \dots & \text{cov}(u_2, u_q) \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}(u_q, u_1) & \text{cov}(u_q, u_2) & \dots & \dots & \text{var}(u_q) \end{bmatrix}$$

What are the single components of matrix G :

$(G)_{11} = \text{var}(u_1) = 1 \cdot \sigma_{u_1}^2$ (because animal 1 has unknown parents and genetic is not inbred)
additive variance

$(G)_{12} = \text{cov}(u_1, u_2) = 0$
Animal 2 has unknown parents, hence animals 1 and 2 are assumed to be unrelated

$(G)_{13} = \text{cov}(u_1, u_3) = \text{cov}(u_1, [\frac{1}{2}(u_1 + u_2) + u_3])$
decomposition u_3 into two

Numerator Relationship Matrix

$$(G)_{13} = \text{cov}(u_1, u_3) = \text{cov}(u_1, \underbrace{\left[\frac{1}{2}(u_1 + u_2) + m_3 \right]}_{\text{decomposition of } u_3 \text{ into breeding values of parents}})$$

decomposition of u_3 into breeding values of parents

$$= \text{cov}(u_1, \frac{1}{2}u_1)$$

$$+ \frac{\text{cov}(u_1, \frac{1}{2}u_2)}{} \rightarrow -0$$

$$+ \frac{\text{cov}(u_1, \frac{1}{2}m_3)}{} \rightarrow -0$$

$$= \frac{1}{2} \cdot \text{cov}(u_1, u_1) = \frac{1}{2} \text{var}(u_1)$$

$$= \frac{1}{2} \sigma_{u_1}^2$$

$$G = \begin{bmatrix} \sigma_{u_1}^2 & 0 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 \\ 0 & \sigma_{u_2}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 \\ \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 \\ \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 \\ \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \frac{1}{2}\sigma_{u_1}^2 & \sigma_{u_1}^2 \end{bmatrix} \begin{bmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \cdot \sigma_{u_1}^2$$

$$\text{cov}(u_3, u_3) = \text{cov}\left(\left[\frac{1}{2}u_1 + \frac{1}{2}u_2 + m_3\right], \left[\frac{1}{2}u_1 + \frac{1}{2}u_2 + m_3\right]\right)$$

NRM II

$$\begin{bmatrix} \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 & \sigma_u^2 & \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 \\ \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 & \sigma_u^2 & \frac{1}{2}\sigma_u^2 \\ \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 & \sigma_u^2 \end{bmatrix} = \frac{1}{2} \quad \text{A}$$

$$\begin{aligned} \text{cov}(u_3, u_4) &= \text{cov}\left(\left[\frac{1}{2}u_1 + \frac{1}{2}u_2 + u_3\right], \left[\frac{1}{2}u_1 + \frac{1}{2}u_2 + u_4\right]\right) \\ &= \text{cov}\left(\frac{1}{2}u_1, \frac{1}{2}u_1\right) + \text{cov}\left(\frac{1}{2}u_1, \frac{1}{2}u_2\right) + \text{cov}\left(\frac{1}{2}u_1, u_3\right) \\ &\quad + \text{cov}\left(\frac{1}{2}u_2, \frac{1}{2}u_1\right) + \text{cov}\left(\frac{1}{2}u_2, \frac{1}{2}u_2\right) + \text{cov}\left(\frac{1}{2}u_2, u_3\right) \\ &\quad + \text{cov}\left(u_3, \frac{1}{2}u_1\right) + \text{cov}\left(u_3, \frac{1}{2}u_2\right) + \text{cov}\left(u_3, u_3\right) \\ &= 0 \\ &= \text{cov}\left(\frac{1}{2}u_1, \frac{1}{2}u_1\right) + \text{cov}\left(\frac{1}{2}u_2, \frac{1}{2}u_2\right) \\ &= \frac{1}{4} \text{cov}(u_1, u_1) + \frac{1}{4} \text{cov}(u_2, u_2) \\ &= \frac{1}{4} \text{var}(u_1) + \frac{1}{4} \text{var}(u_2) \\ &= \frac{1}{4} \sigma_u^2 + \frac{1}{4} \sigma_u^2 = \frac{1}{2} \sigma_u^2 \end{aligned}$$