HW10

2025-05-03

Q1

Let $(X^TX)^-$ be any generalized inverse of X^TX. A generalized inverse of a symmetric matrix is not necessarily symmetric. Thus, we cannot assume that

$$\left[(X^TX)^-\right]^T = [(X^TX)^T]$$

always holds.

Find a matrix X such that $\left[(X^TX)^-\right]^T \neq \left[(X^TX)^T\right]^-$.

However, it is also true that a symmetric generalized inverse can always be found for a symmetric matrix.

Answer

To find a matrix X such that the generalized inverse $(X^TX)^-$ is not symmetric, we can proceed with the following steps:

Let X be a 2×2 matrix:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Compute X^TX:

$$X^TX = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

A generalized inverse $(X^TX)^-$ must satisfy:

$$X^TX \cdot (X^TX)^- \cdot X^TX = X^TX$$

One such generalized inverse is:

$$(X^TX)^- = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

This matrix is not symmetric because:

$$\left[(X^TX)^-\right]^T = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = (X^TX)^-$$

Check the generalized inverse condition:

$$X^T X \cdot (X^T X)^- \cdot X^T X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = X^T X$$

The condition holds, and $(X^T X)^-$ is indeed non-symmetric.

The matrix

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

has a non-symmetric generalized inverse

$$(X^T X)^- = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

satisfying $\left[(X^T X)^-\right]^T \neq (X^T X)^-$.

This demonstrates that even for symmetric $X^T X$, a generalized inverse need not be symmetric.

Q2

See [10-1]

For each of the following special cases, derive the REML estimator of σ^2.

a)

Suppose $y_1, y_2, y_3 \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Answer

1) Find $n - \text{rank}(\mathbf{X}) = 3 - 1 = 2$ linearly independent vectors $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2]$ such that $\mathbf{a}_i'\mathbf{X} = \mathbf{0}'$. From the model, we have $\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, so one of the choices can be $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$.

2) Find the MLE of σ^2 using $\mathbf{w} \equiv \mathbf{A}'\mathbf{y} = \begin{bmatrix} y_1 - y_2 \\ y_1 - y_3 \end{bmatrix}$ as data.

$$\mathbf{w} = \mathbf{A}'\mathbf{y} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\boldsymbol{\epsilon} = \mathbf{0} + \mathbf{A}'\boldsymbol{\epsilon} = \mathbf{A}'\boldsymbol{\epsilon}$$

Thus, $\mathbf{w} \sim N(\mathbf{0}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})$, where

$$\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \sigma^2 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

And we have

$$\det(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}) = 3\sigma^4$$

$$\mathbf{w}'(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{w} = \frac{2}{3\sigma^2}\left[(y_1 - y_2)^2 - (y_1 - y_2)(y_1 - y_3) + (y_1 - y_3)^2\right] \equiv \frac{2}{3\sigma^2}\Delta$$

So, $\mathbf{w} \sim N\left(\mathbf{0}, \sigma^2 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right)$, and the log-likelihood function is

$$l(\sigma^2|\mathbf{w}) = -\frac{1}{2}\log(\det(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})) - \frac{1}{2}\mathbf{w}'(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{w} - \frac{1}{2}\log(2\pi)$$

The score equation is

$$\frac{\partial l}{\partial \sigma^2} = -\frac{1}{\sigma^2} + \frac{\Delta}{3\sigma^4} = 0 \implies \hat{\sigma}^2 = \frac{\Delta}{3}$$

Therefore, the REML estimator of σ^2 in this case is

$$\frac{\Delta}{3} = \frac{1}{3}\left[(y_1 - y_2)^2 - (y_1 - y_2)(y_1 - y_3) + (y_1 - y_3)^2\right].$$

b)

Suppose

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \sim \mathcal{N} \begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \end{pmatrix}, \begin{bmatrix} \sigma^2 & \sigma^2/2 & 0 & 0 \\ \sigma^2/2 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & \sigma^2/2 \\ 0 & 0 & \sigma^2/2 & \sigma^2 \end{bmatrix}$$

Answer

(b)

Follow the steps of slide 8 of set 20:

1) Find $n - \text{rank}(\mathbf{X}) = 4 - 2 = 2$ linearly independent vectors $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2]$ such that

$$\mathbf{a}_i'\mathbf{X} = \mathbf{0}'.$$

From the model we have

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

so one of the choices can be

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}.$$

2) Find the MLE of σ^2 using $\mathbf{w} \equiv \mathbf{A}'\mathbf{y} = \begin{bmatrix} y_1 - y_2 \\ y_3 - y_4 \end{bmatrix}$ as data.

$$\mathbf{w} = \mathbf{A}'\mathbf{y} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\boldsymbol{\epsilon} = \mathbf{0} + \mathbf{A}'\boldsymbol{\epsilon} = \mathbf{A}'\boldsymbol{\epsilon}$$

Thus $\mathbf{w} \sim N(\mathbf{0}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})$ where

$$\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma^2 & \sigma^2/2 & 0 & 0 \\ \sigma^2/2 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & \sigma^2/2 \\ 0 & 0 & \sigma^2/2 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

So $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and we can use the Gauss-Markov linear model result directly to find the MLE.

$$\hat{\sigma}^2 = \frac{\mathbf{w}'(\mathbf{I} - \mathbf{P})\mathbf{w}}{2} \quad \text{where} \quad \mathbf{P} = \mathbf{0} \quad \text{is the projection matrix for design matrix} \quad \mathbf{0}$$

$$= \frac{\mathbf{w}'\mathbf{w}}{2}$$

Therefore the REML estimator of σ^2 in this case is

$$\frac{\mathbf{w}'\mathbf{w}}{2} = \frac{1}{2}\left[(y_1 - y_2)^2 + (y_3 - y_4)^2\right].$$

Q3

See [10-2]

Suppose 100 maize genotypes were assigned to 304 plots in a field using an unbalanced completely randomized design in which some genotypes were assigned to only one plot while others were assigned to as many as six plots. Plots were planted with seed from their assigned genotypes, and yield in bushels per acre was recorded for each plot at the end of the growing season. The dataset is available in Canvas. Consider the model

$$y_{ij} = \mu + g_i + e_{ij},$$

where $\mu + g_i$ is the mean yield for the ith genotype, and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ for all i and j, with independence among all e_{ij} terms.

a)

Find the BLUE of $\mu + g_i$ for each $i = 1, \ldots, 100$.

Answer

If we use the parametrization $\boldsymbol{\beta} = (\mu_1, \cdots, \mu_{100})'$ where $\mu_i = \mu + g_i$, $i = 1, ..., 100$, the model matrix is:

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{n_1 \times 1} & & & & \\ & \mathbf{1}_{n_2 \times 1} & & & \\ & & \ddots & & \\ & & & \mathbf{1}_{n_{99} \times 1} & \\ & & & & \mathbf{1}_{n_{100} \times 1} \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n_1 & & & & \\ & n_2 & & & \\ & & \ddots & & \\ & & & n_{99} & \\ & & & & n_{100} \end{bmatrix} \text{ and } (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n_1} & & & & \\ & \frac{1}{n_2} & & & \\ & & \ddots & & \\ & & & \frac{1}{n_{99}} & \\ & & & & \frac{1}{n_{100}} \end{bmatrix}$$

Thus, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\bar{y}_1, ..., \bar{y}_{100})'$ and $\hat{\mu}_i = \widehat{\mu + g_i} = \bar{y}_i$ for $i = 1, ..., 100$.

R code:

```
dat <- read.table("https://dnett.github.io/S510/hw10GenotypeYield.txt",
                  header = TRUE, col.names = c("genotype", "yield"),
                  colClasses = c("factor", "numeric"))
dat$genotype <- factor(dat$genotype, levels = 1:100)
ols.f <- lm(yield ~ 0 + genotype, data = dat)
ols.f
```

```
##
## Call:
## lm(formula = yield ~ 0 + genotype, data = dat)
##
## Coefficients:
##   genotype1    genotype2    genotype3    genotype4    genotype5    genotype6
```

```
##       194.9       184.2       191.4       198.6       194.2       197.7
##    genotype7    genotype8    genotype9   genotype10   genotype11   genotype12
##       193.6       196.6       202.3       196.1       182.1       186.1
##   genotype13   genotype14   genotype15   genotype16   genotype17   genotype18
##       184.0       170.2       189.9       187.4       192.6       189.3
##   genotype19   genotype20   genotype21   genotype22   genotype23   genotype24
##       185.7       197.8       199.3       192.8       183.1       200.8
##   genotype25   genotype26   genotype27   genotype28   genotype29   genotype30
##       190.6       182.7       192.8       187.2       195.5       194.0
##   genotype31   genotype32   genotype33   genotype34   genotype35   genotype36
##       178.7       203.7       185.5       189.6       196.0       187.1
##   genotype37   genotype38   genotype39   genotype40   genotype41   genotype42
##       188.2       190.3       185.4       191.5       186.2       187.2
##   genotype43   genotype44   genotype45   genotype46   genotype47   genotype48
##       189.7       179.6       189.1       190.5       185.7       206.2
##   genotype49   genotype50   genotype51   genotype52   genotype53   genotype54
##       192.2       194.3       197.7       184.5       193.2       182.2
##   genotype55   genotype56   genotype57   genotype58   genotype59   genotype60
##       192.1       188.9       185.4       193.0       198.3       192.4
##   genotype61   genotype62   genotype63   genotype64   genotype65   genotype66
##       189.2       181.5       192.2       189.3       196.1       201.5
##   genotype67   genotype68   genotype69   genotype70   genotype71   genotype72
##       194.0       194.4       181.1       201.3       185.4       183.7
##   genotype73   genotype74   genotype75   genotype76   genotype77   genotype78
##       194.6       196.1       196.8       179.6       191.5       196.1
##   genotype79   genotype80   genotype81   genotype82   genotype83   genotype84
##       190.1       186.5       194.1       189.4       184.5       191.5
##   genotype85   genotype86   genotype87   genotype88   genotype89   genotype90
##       177.4       193.0       194.7       196.2       197.4       190.5
##   genotype91   genotype92   genotype93   genotype94   genotype95   genotype96
##       189.9       193.9       177.6       183.8       189.3       178.2
##   genotype97   genotype98   genotype99  genotype100
##       200.3       188.8       192.9       191.8
```

b)

For this and all subsequent parts of this problem, assume $g_1, \ldots, g_{100} \overset{iid}{\sim} \mathcal{N}(0, \sigma_g^2)$ and independent of all the e_{ij} terms. Find the REML estimates of σ_g^2 and σ_e^2.

Answer

Based on the output below, the REML estimates of σ_g^2 and σ_e^2 are $(2.6865)^2 = 7.2174$ and $(9.669)^2 = 93.4899$ respectively. The code and output are shown below:

```r
library(nlme)
set.seed(1234)
o=lme(yield~1,random=~1|genotype,data=dat)
o
```

```
## Linear mixed-effects model fit by REML
##   Data: dat
##   Log-restricted-likelihood: -1130.573
```

```
##   Fixed: yield ~ 1
## (Intercept)
##    190.6983
##
## Random effects:
##  Formula: ~1 | genotype
##         (Intercept) Residual
## StdDev:    2.686537 9.669021
##
## Number of Observations: 304
## Number of Groups: 100
```

c)

Find the BLUP of $\mu + g_i$ for each $i = 1, \ldots, 100$.

Answer

Note that $X = \mathbf{1}$, $\boldsymbol{\beta} = \mu$, and

$$Z = \begin{bmatrix} \mathbf{1}_{n_1 \times 1} \\ \mathbf{1}_{n_2 \times 1} \\ \vdots \\ \mathbf{1}_{n_{100} \times 1} \end{bmatrix}, \quad G = \sigma_g^2 I, \quad R = \sigma_e^2 I$$

The BLUP for $\mathbf{g} = (g_1, g_2, \ldots, g_{100})'$ is

$$\hat{\mathbf{g}} = GZ^\top \Sigma^{-1}(\mathbf{y} - X\hat{\beta}_\Sigma)$$

where $\Sigma = ZGZ^\top + R$. Then, the BLUP for $\mu + g_i$ is

$$\frac{n_i \sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2} \left(\bar{y}_{i\cdot} - \hat{\beta}_\Sigma \right) + \frac{\sigma_e^2}{\sigma_e^2 + n_i \sigma_g^2} \hat{\beta}_\Sigma$$

where

$$\hat{\beta}_\Sigma = \left(\mathbf{1}^\top \Sigma^{-1} \mathbf{1} \right)^{-1} \mathbf{1}^\top \Sigma^{-1} \mathbf{y} = \frac{\sum_{i=1}^{100} \frac{n_i \bar{y}_{i\cdot}}{\sigma_e^2 + n_i \sigma_g^2}}{\sum_{i=1}^{100} \frac{n_i}{\sigma_e^2 + n_i \sigma_g^2}}$$

R Code Output for Empirical BLUP

```r
b = fixef(o)
u = ranef(o)
blup = as.matrix(b + u)
blup
```

```
##     (Intercept)
## 1      191.4947
## 2      189.8359
```

```
## 3      190.8365
## 4      192.1842
## 5      191.7963
## 6      192.6427
## 7      191.4953
## 8      191.4877
## 9      193.4415
## 10     191.4208
## 11     188.6755
## 12     189.4232
## 13     189.4324
## 14     187.9567
## 15     190.5159
## 16     189.7853
## 17     190.9527
## 18     190.4291
## 19     189.2951
## 20     192.0275
## 21     193.4319
## 22     190.9861
## 23     190.1538
## 24     192.0494
## 25     190.6633
## 26     189.1880
## 27     191.2001
## 28     190.0405
## 29     191.6075
## 30     191.1466
## 31     188.4358
## 32     192.4440
## 33     189.7145
## 34     190.4333
## 35     191.4074
## 36     190.0154
## 37     190.3709
## 38     190.6517
## 39     189.9897
## 40     190.7989
## 41     190.3759
## 42     190.4476
## 43     190.4259
## 44     188.0738
## 45     190.3915
## 46     190.6610
## 47     189.7521
## 48     191.8093
## 49     190.8059
## 50     191.1867
## 51     192.0087
## 52     190.2541
## 53     191.4062
## 54     189.5617
## 55     190.9619
## 56     190.3601
```

```
## 57     189.0052
## 58     191.0129
## 59     192.1341
## 60     191.0246
## 61     190.4166
## 62     188.5280
## 63     190.9870
## 64     190.3684
## 65     191.7078
## 66     192.1430
## 67     191.1466
## 68     191.3881
## 69     188.8996
## 70     192.6920
## 71     189.4600
## 72     189.3760
## 73     191.6248
## 74     191.9669
## 75     192.4088
## 76     189.2206
## 77     190.9272
## 78     191.7078
## 79     190.5571
## 80     189.7019
## 81     191.3380
## 82     190.4604
## 83     190.2541
## 84     190.8491
## 85     188.1976
## 86     191.1374
## 87     190.9851
## 88     191.9846
## 89     191.9648
## 90     190.6610
## 91     190.5915
## 92     191.3067
## 93     188.9397
## 94     189.7824
## 95     190.4354
## 96     189.0334
## 97     192.9638
## 98     190.2445
## 99     190.9928
## 100    190.9641
```

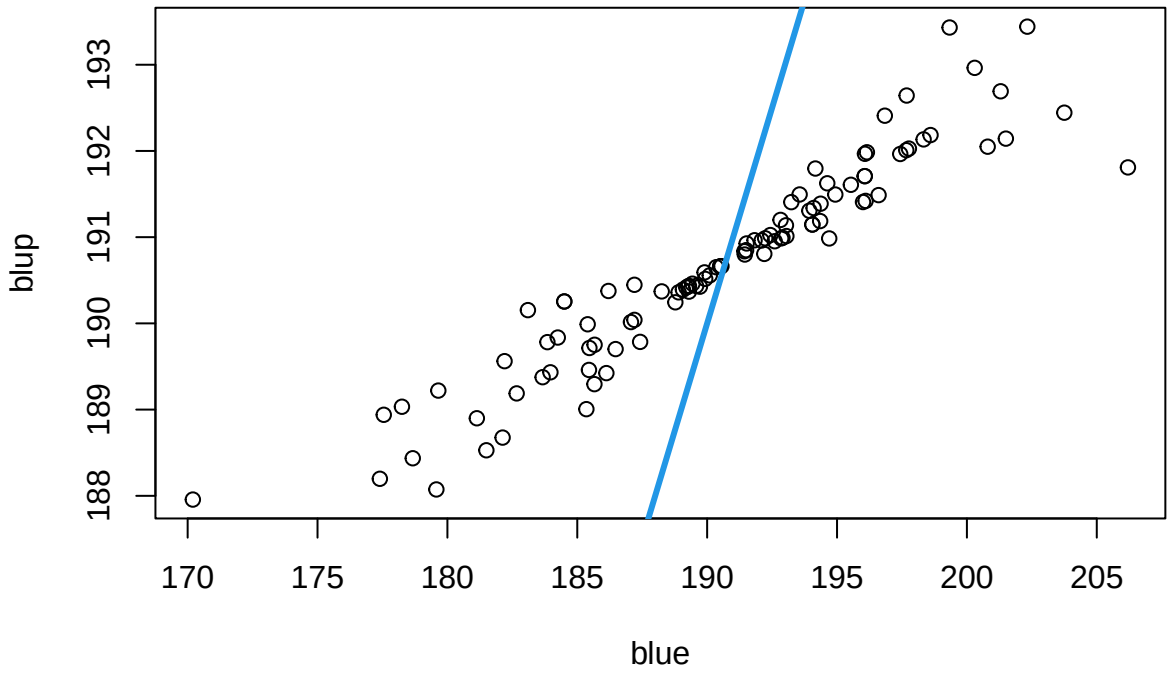
d)

Make a plot of the BLUPs (vertical axis) vs. the BLUEs from part (a) (horizontal axis) with one point for each genotype. Add the $y = x$ line to your plot. Explain why the plot looks the way it does.

Answer

The plot of the eBLUPSs (vertical axis) vs. the BLUEs from part (a) (horizontal axis) is produced by the R code that follows:

```
blue=as.vector(ols.f$coefficients)
plot(blue,blup)
abline(a=0,b=1,col=4,lwd=3)
```



e)

According to the BLUEs from part a), list the top five highest yielding genotypes.

Answer

According to the BLUEs from part (a), the top five highest yielding genotypes are as follows:

```
blue.ord = order(blue,decreasing = T)
top5 = blue.ord[1:5]
print(data.frame(Top5=top5,Blue=blue[top5]))
```

```
##   Top5    Blue
## 1   48 206.200
```

```
## 2   32 203.750
## 3    9 202.325
## 4   66 201.500
## 5   70 201.300
```

f)

According to the BLUPs, list the top five highest yielding genotypes.

Answer

According to the eBLUPs, the top five highest yielding genotypes are as follows:

```
blup.ord = order(blup,decreasing = T)
top5 = blup.ord[1:5]
print(data.frame(Top5=top5,Blup=blup[top5]))
```

```
##   Top5     Blup
## 1    9 193.4415
## 2   21 193.4319
## 3   97 192.9638
## 4   70 192.6920
## 5    6 192.6427
```

g)

Why is the top-yielding genotype according to the BLUEs from part a) not so highly rated according to the BLUPs?

Answer

(g)

The BLUE of $\mu + g_i$ from part (a) is simply the sample mean $\bar{y}_{i\cdot}$ for the i-th genotype.

The BLUP of $\mu+g_i$, on the other hand, is a convex combination of: - the sample mean $\bar{y}_{i\cdot}$, and - the weighted average of all sample means $\hat{\beta}_\Sigma$ from part (c).

The weights in this combination depend on the sample size n_i, as well as the variance components σ_e^2 (residual variance) and σ_g^2 (genotype variance).

Even if a BLUE from part (a) is large, the corresponding BLUP may be smaller, due to this weighting structure. This explains why the top-yielding genotype according to the BLUEs may not rank as highly under the BLUPs.

Example:
Genotype 48 has $n_{48} = 1$, so while the BLUE $\bar{y}_{48} = 206.2$ is the highest, it is based on a single observation and thus unreliable. As a result, the eBLUP for genotype 48 shrinks substantially toward the overall mean $\hat{\beta}_\Sigma$, giving it a lower rank among the eBLUPs.

Q4

See [11-2]

This is a repeated measures analysis. An experiment was designed to compare the effect of three drugs (A, B, and C) on the heart rate of women. Fifteen women were randomly assigned to the drugs using a completely randomized design with five women for each drug. The heart rate (in beats per minute) of each woman was measured at 0, 5, 10, and 15 minutes after the drug was administered. The data are provided in the file HeartRate.txt. Let y_{ijk} denote the heart rate at the kth time point for the jth woman treated with the ith drug. Suppose

$$y_{ijk} = \mu_{ik} + \epsilon_{ijk},$$

where μ_{ik} is an unknown constant for each combination of $i = 1, 2, 3$ and $k = 1, 2, 3, 4, 5$ and ϵ_{ijk} is a normally distributed error term with mean 0 for all $i = 1, 2, 3$, $j = 1, 2, 3, 4, 5$, and $k = 1, 2, 3, 4$. For all $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$, let

$$\epsilon_{ij} = (\epsilon_{ij1}, \epsilon_{ij2}, \epsilon_{ij3}, \epsilon_{ijd})^\top.$$

Suppose all the ϵ_{ij} vectors are mutually independent, and let $\mathbf{W}$ be the variance-covariance matrix of ϵ_{ij}, which is assumed to be the same for all $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$.

a)

Find the REML estimate of $\mathbf{W}$ under the assumption that $\mathbf{W}$ is a positive definite, compound symmetric matrix.

Answer

Under a compound symmetry assumption,

$$\boldsymbol{W} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix},$$

where the REML estimates for the heart rate data are $\hat{\sigma} = 6.12$ and $\hat{\rho} = 0.777$ ($\hat{\sigma}_s^2 = 29.13$, $\hat{\sigma}_e^2 = 8.36$).

b)

Find AIC and BIC for the case where $\mathbf{W}$ is a positive definite, compound symmetric matrix.

Answer

Using R,
AIC $= -2(-144.9602) + 2(14) = 317.92$
BIC $= -2(-144.9602) + (14)\log(48) = 344.12$

Using SAS,
AIC $= -2(-144.9602) + 2(2) = 293.9$
BIC $= -2(-144.9602) + (2)\log(15) = 295.3$

c)

Find the REML estimate of **W** under the assumption that **W** is a positive definite matrix with constant variance and an AR(1) correlation structure.

Answer

Under an AR(1) assumption,

$$\boldsymbol{W} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix},$$

where the REML estimates for the heart rate data are $\hat{\sigma} = 6.00$ and $\hat{\rho} = 0.828$.

d)

Find AIC and BIC for the case where **W** is a a positive definite matrix with constant variance and an AR(1) correlation structure.

Answer

Using R,
AIC $= -2(-142.9713) + 2(14) = 313.94$
BIC $= -2(-142.9713) + (14)\log(48) = 340.14$

Using SAS,
AIC $= -2(-142.9713) + 2(2) = 289.9$
BIC $= -2(-142.9713) + (2)\log(15) = 291.4$

e)

Find the REML estimate of **W** under the assumption that **W** is a positive definite, symmetric matrix.

Answer

Under a general symmetry assumption,

$$\boldsymbol{W} = \sigma^2 \begin{bmatrix} 1 & \rho_{12}\delta_2 & \rho_{13}\delta_3 & \rho_{14}\delta_4 \\ \rho_{12}\delta_2 & \delta_2^2 & \rho_{23}\delta_2\delta_3 & \rho_{24}\delta_2\delta_4 \\ \rho_{13}\delta_3 & \rho_{23}\delta_2\delta_3 & \delta_3^2 & \rho_{34}\delta_3\delta_4 \\ \rho_{14}\delta_4 & \rho_{24}\delta_2\delta_4 & \rho_{34}\delta_3\delta_4 & \delta_4^2 \end{bmatrix},$$

where the REML estimates for the heart rate data are:

$$\hat{\sigma} = 6.10, \quad \hat{\delta}_2 = 1.08, \quad \hat{\delta}_3 = 0.995, \quad \hat{\delta}_4 = 0.928,$$
$$\hat{\rho}_{12} = 0.850, \quad \hat{\rho}_{13} = 0.889, \quad \hat{\rho}_{14} = 0.625,$$
$$\hat{\rho}_{23} = 0.870, \quad \hat{\rho}_{24} = 0.631, \quad \hat{\rho}_{34} = 0.794.$$

f)

Find AIC and BIC for the case where $\mathbf{W}$ is a positive definite, symmetric matrix.

Answer

Using R,
$\text{AIC} = -2(-139.424) + 2(22) = 322.85$
$\text{BIC} = -2(-139.424) + (22)\log(48) = 364.01$

Using SAS,
$\text{AIC} = -2(-139.424) + 2(10) = 298.8$
$\text{BIC} = -2(-139.424) + (10)\log(15) = 305.9$

g)

Which of the three structures for $\mathbf{W}$ is preferred for this dataset?

Answer

The model with an AR(1) correlation structure has the smallest AIC and BIC of the three (regardless of whether you used R or SAS). Consequently, the AR(1) correlation structure is preferred for this dataset.

h)

Using the preferred structure for $\mathbf{W}$, compute a 95% confidence interval for the mean heart rate 10 minutes after treatment with drug A minus the mean heart rate 10 minutes after treatment with drug B.

Answer

There are several ways to find a 95% confidence interval for $\mu_{13} - \mu_{23}$ using the model with an AR(1) correlation structure.

- Using Cochran-Satterthwaite in SAS (df = 19.2): $(-3.54, 12.34)$
- Default SAS method (df = 36): $(-3.30, 12.10)$
- R's gls (df = 48): $(-3.23, 12.03)$

i)

Using the preferred structure for $\mathbf{W}$, compute a 95% confidence interval for the mean heart rate 10 minutes after treatment with drug A minus the mean heart rate 5 minutes after treatment with drug A.

Answer

An approximate 95% confidence interval for $\mu_{13} - \mu_{12}$:

- Cochran-Satterthwaite (df=35.9): $(-3.7946, 2.5946)$
- Default SAS (df=36): $(-3.7942, 2.5943)$
- R (df=48): $(-3.7667, 2.5668)$

Note: R Code

```
d=read.delim("https://dnett.github.io/S510/HeartRate.txt")
library(nlme)
```

```
attach(d)
woman <- as.factor(woman)
drug <- as.factor(drug)
time <- as.factor(time)
model.cs <- gls(y ~ drug * time,
correlation = corCompSymm(form = ~1 | woman),
method = "REML")
model.ar <- gls(y ~ drug * time,
correlation = corAR1(form = ~1 | woman),
method = "REML")
model.sy <- gls(y ~ drug * time,
correlation = corSymm(form = ~1 | woman),
weight = varIdent(form = ~1 | time),
method = "REML")
summary(model.cs)
```

```
## Generalized least squares fit by REML
##   Model: y ~ drug * time
##   Data: NULL
##        AIC      BIC    logLik
##   317.9204 344.1172 -144.9602
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | woman
##  Parameter estimate(s):
##       Rho
## 0.7769134
##
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept)   71.8  2.738308 26.220567  0.0000
## drugB          9.6  3.872553  2.478985  0.0167
## drugC          2.4  3.872553  0.619746  0.5384
## time5         10.8  1.829086  5.904588  0.0000
## time10        10.2  1.829086  5.576556  0.0000
## time15         1.8  1.829086  0.984098  0.3300
## drugB:time5   -7.6  2.586718 -2.938086  0.0051
## drugC:time5   -9.4  2.586718 -3.633948  0.0007
## drugB:time10 -14.0  2.586718 -5.412263  0.0000
## drugC:time10  -9.8  2.586718 -3.788584  0.0004
## drugB:time15  -4.0  2.586718 -1.546361  0.1286
## drugC:time15  -0.8  2.586718 -0.309272  0.7585
##
##  Correlation:
##              (Intr) drugB  drugC  time5  time10 time15 drgB:5 drgC:5 drB:10
## drugB        -0.707
## drugC        -0.707  0.500
```

```
## time5        -0.334  0.236  0.236
## time10       -0.334  0.236  0.236  0.500
## time15       -0.334  0.236  0.236  0.500  0.500
## drugB:time5   0.236 -0.334 -0.167 -0.707 -0.354 -0.354
## drugC:time5   0.236 -0.167 -0.334 -0.707 -0.354 -0.354  0.500
## drugB:time10  0.236 -0.334 -0.167 -0.354 -0.707 -0.354  0.500  0.250
## drugC:time10  0.236 -0.167 -0.334 -0.354 -0.707 -0.354  0.250  0.500  0.500
## drugB:time15  0.236 -0.334 -0.167 -0.354 -0.354 -0.707  0.500  0.250  0.500
## drugC:time15  0.236 -0.167 -0.334 -0.354 -0.354 -0.707  0.250  0.500  0.250
##              drC:10 drB:15
## drugB
## drugC
## time5
## time10
## time15
## drugB:time5
## drugC:time5
## drugB:time10
## drugC:time10
## drugB:time15  0.250
## drugC:time15  0.500  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.22111753 -0.58794288  0.04899524  0.55527938  2.35177151
##
## Residual standard error: 6.123044
## Degrees of freedom: 60 total; 48 residual
```

```
getVarCov(model.cs)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 37.492 29.128 29.128 29.128
## [2,] 29.128 37.492 29.128 29.128
## [3,] 29.128 29.128 37.492 29.128
## [4,] 29.128 29.128 29.128 37.492
##   Standard Deviations: 6.123 6.123 6.123 6.123
```

```
summary(model.ar)
```

```
## Generalized least squares fit by REML
##   Model: y ~ drug * time
##   Data: NULL
##        AIC      BIC    logLik
##   313.9425 340.1394 -142.9713
##
## Correlation Structure: AR(1)
##  Formula: ~1 | woman
##  Parameter estimate(s):
##       Phi
## 0.8277814
##
```

```
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept)   71.8  2.683679 26.754314  0.0000
## drugB          9.6  3.795296  2.529447  0.0148
## drugC          2.4  3.795296  0.632362  0.5302
## time5         10.8  1.575019  6.857062  0.0000
## time10        10.2  2.129354  4.790186  0.0000
## time15         1.8  2.496791  0.720925  0.4745
## drugB:time5   -7.6  2.227413 -3.412031  0.0013
## drugC:time5   -9.4  2.227413 -4.220143  0.0001
## drugB:time10 -14.0  3.011361 -4.649061  0.0000
## drugC:time10  -9.8  3.011361 -3.254343  0.0021
## drugB:time15  -4.0  3.530996 -1.132825  0.2629
## drugC:time15  -0.8  3.530996 -0.226565  0.8217
##
##  Correlation:
##              (Intr) drugB  drugC  time5  time10 time15 drgB:5 drgC:5 drB:10
## drugB        -0.707
## drugC        -0.707  0.500
## time5        -0.293  0.207  0.207
## time10       -0.397  0.281  0.281  0.676
## time15       -0.465  0.329  0.329  0.532  0.779
## drugB:time5   0.207 -0.293 -0.147 -0.707 -0.478 -0.376
## drugC:time5   0.207 -0.147 -0.293 -0.707 -0.478 -0.376  0.500
## drugB:time10  0.281 -0.397 -0.198 -0.478 -0.707 -0.551  0.676  0.338
## drugC:time10  0.281 -0.198 -0.397 -0.478 -0.707 -0.551  0.338  0.676  0.500
## drugB:time15  0.329 -0.465 -0.233 -0.376 -0.551 -0.707  0.532  0.266  0.779
## drugC:time15  0.329 -0.233 -0.465 -0.376 -0.551 -0.707  0.266  0.532  0.390
##              drC:10 drB:15
## drugB
## drugC
## time5
## time10
## time15
## drugB:time5
## drugC:time5
## drugB:time10
## drugC:time10
## drugB:time15  0.390
## drugC:time15  0.779  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.26633066 -0.59991106  0.04999259  0.56658267  2.39964423
##
## Residual standard error: 6.00089
## Degrees of freedom: 60 total; 48 residual
```

```
getVarCov(model.ar)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 36.011 29.809 24.675 20.426
## [2,] 29.809 36.011 29.809 24.675
```

```
## [3,] 24.675 29.809 36.011 29.809
## [4,] 20.426 24.675 29.809 36.011
##   Standard Deviations: 6.0009 6.0009 6.0009 6.0009
```

```
summary(model.sy)
```

```
## Generalized least squares fit by REML
##   Model: y ~ drug * time
##   Data: NULL
##        AIC      BIC   logLik
##   322.8481 364.0145 -139.424
##
## Correlation Structure: General
##  Formula: ~1 | woman
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.850
## 3 0.889 0.870
## 4 0.625 0.631 0.794
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | time
##  Parameter estimates:
##         0         5        10        15
## 1.0000000 1.0846078 0.9950670 0.9280264
##
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept)   71.8  2.728847 26.311483  0.0000
## drugB          9.6  3.859172  2.487580  0.0164
## drugC          2.4  3.859172  0.621895  0.5370
## time5         10.8  1.574804  6.857998  0.0000
## time10        10.2  1.283229  7.948698  0.0000
## time15         1.8  2.286190  0.787336  0.4350
## drugB:time5   -7.6  2.227109 -3.412496  0.0013
## drugC:time5   -9.4  2.227109 -4.220719  0.0001
## drugB:time10 -14.0  1.814760 -7.714519  0.0000
## drugC:time10  -9.8  1.814760 -5.400164  0.0000
## drugB:time15  -4.0  3.233161 -1.237179  0.2220
## drugC:time15  -0.8  3.233161 -0.247436  0.8056
##
##  Correlation:
##              (Intr) drugB  drugC  time5  time10 time15 drgB:5 drgC:5 drB:10
## drugB        -0.707
## drugC        -0.707  0.500
## time5        -0.136  0.096  0.096
## time10       -0.246  0.174  0.174  0.488
## time15       -0.502  0.355  0.355  0.278  0.684
## drugB:time5   0.096 -0.136 -0.068 -0.707 -0.345 -0.196
## drugC:time5   0.096 -0.068 -0.136 -0.707 -0.345 -0.196  0.500
## drugB:time10  0.174 -0.246 -0.123 -0.345 -0.707 -0.484  0.488  0.244
## drugC:time10  0.174 -0.123 -0.246 -0.345 -0.707 -0.484  0.244  0.488  0.500
## drugB:time15  0.355 -0.502 -0.251 -0.196 -0.484 -0.707  0.278  0.139  0.684
```

```
## drugC:time15  0.355 -0.251 -0.502 -0.196 -0.484 -0.707  0.139  0.278  0.342
##              drC:10 drB:15
## drugB
## drugC
## time5
## time10
## time15
## drugB:time5
## drugC:time5
## drugB:time10
## drugC:time10
## drugB:time15  0.342
## drugC:time15  0.684  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.05495379 -0.56510028  0.04660828  0.51373845  2.20692889
##
## Residual standard error: 6.101886
## Degrees of freedom: 60 total; 48 residual
```

```
getVarCov(model.sy)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 37.233 34.316 32.933 21.583
## [2,] 34.316 43.800 34.950 23.666
## [3,] 32.933 34.950 36.867 27.316
## [4,] 21.583 23.666 27.316 32.066
##   Standard Deviations: 6.1019 6.6182 6.0718 5.6627
```

```
ci.gls <- function(lmeout, C, df, a = 0.05) {
b = coef(lmeout)
V = vcov(lmeout)
Cb = C %*% b
se = sqrt(diag(C %*% V %*% t(C)))
tval = qt(1 - a / 2, df)
low = Cb - tval * se
up = Cb + tval * se
m = cbind( Cb, se, low, up)
dimnames(m)[[2]] = c("estimate", "se", paste(100 * (1 - a), "% Conf.", sep = ""), "limits")
return(m)
}

C2 <- matrix(c(0, 0, 0, 1, -1, 0, 0, 0, 0, 0, 0, 0), nrow = 1) # problem(h)
ci.gls(model.ar, C2, 19.2) # Cheated and took Cochran-Satterthwaite df value from SAS.
```

```
##      estimate       se 95% Conf.   limits
## [1,]      0.6 1.575019 -2.694229 3.894229
```

```
ci.gls(model.ar, C2, 36) # Default df method in SAS.
```

```
##      estimate       se 95% Conf.   limits
## [1,]      0.6 1.575019 -2.594286 3.794286
```

```
ci.gls(model.ar, C2, 48) # Default df method in R.
```

```
##      estimate       se 95% Conf.   limits
## [1,]      0.6 1.575019 -2.566787 3.766787
```

```
C3 <- matrix(c(0, 0, 0, -1, 1, 0, 0, 0, 0, 0, 0, 0), nrow = 1) # problem(i)
ci.gls(model.ar, C3, 35.9) # Cheated and took Cochran-Satterthwaite df value from SAS.
```

```
##      estimate       se 95% Conf.   limits
## [1,]     -0.6 1.575019 -3.794595 2.594595
```

```
ci.gls(model.ar, C3, 36) # Default df method in SAS.
```

```
##      estimate       se 95% Conf.   limits
## [1,]     -0.6 1.575019 -3.794286 2.594286
```

```
ci.gls(model.ar, C3, 48) # Default df method in R.
```

```
##      estimate       se 95% Conf.   limits
## [1,]     -0.6 1.575019 -3.766787 2.566787
```