Missing values

Facundo Muñoz

2015-11-25 breedR version: 0.10.19

Contents

The handling of missing values (i.e. `NA`) depends on *where* they are.

Missing response

It is perfectly valid to have missing vaules in the dependent variable. There is no need of removing those individuals from the dataset. Furthermore, including them will yield [predictions](https://github.com/famuvie/breedR/wiki/Overview#prediction for their phenotype, based on the predictive variables.

```
library(breedR)

N <- 1e3
x <- rep(1:4, each = N/4)
dat <- data.frame(y = x + rnorm(N),
                  x = factor(letters[x]))
dat$y[1] <- NA
head(dat)
```

```
##            y x
## 1         NA a
## 2 -0.1537691 a
## 3  4.3494888 a
## 4  1.1661482 a
## 5  2.7999953 a
## 6  1.2697673 a
```

```
res <- remlf90(y ~ x, data = dat)

## The predicted phenotype for y[1] is the estimated effect
## of the corresponding level of x
fitted(res)[1] == fixef(res)$x['a', 'value']
```

```
##    1
## TRUE
```

Missing value for a fixed effect

This is not allowed, as it would yield an underdetermined system of equations. `breedR` issues a warning if missing values are detected.

```r
N <- 1e3
x <- rep(1:4, each = N/4)
dat <- data.frame(y = x + rnorm(N),
                  x = factor(letters[x]))
dat$x[c(1, 3, 5)] <- NA
head(dat)
```

```
##            y    x
## 1  0.6419022 <NA>
## 2 -0.1239437    a
## 3  1.5444195 <NA>
## 4  0.9976375    a
## 5  0.9921047 <NA>
## 6  2.1273015    a
```

```r
res <- remlf90(y ~ x, data = dat)
```

```
## Error in progsf90(mf, effects, opt = union("sol se", progsf90.options), :
## Missing values in covariates are not allowed
## check individuals: 1, 3, 5
```

Idem for a regression variable.

```r
N <- 1e3
x <- runif(N)
dat <- data.frame(y = 1 + 2*x + rnorm(N),
                  x = x)
dat$x[c(1, 3, 5)] <- NA
head(dat)
```

```
##            y          x
## 1 -0.1313425         NA
## 2  3.3738401 0.94955025
## 3  4.4736444         NA
## 4 -0.1999409 0.01226402
## 5  0.5221936         NA
## 6  2.2613993 0.23531621
```

```r
res <- remlf90(y ~ x, data = dat)
```

```
## Error in progsf90(mf, effects, opt = union("sol se", progsf90.options), :
## Missing values in covariates are not allowed
## check individuals: 1, 3, 5
```

Missing value for a random effect

These **are** allowed. The incidence matrix will have a row of zeros for the corresponding individual.

```r
N <- 1e3
N.blk <- 20
blk.effects <- rnorm(N.blk, sd = 2)
blk.idx <- sample(seq_len(N.blk), N, replace = TRUE)
dat <- data.frame(y = 1 + blk.effects[blk.idx] + rnorm(N),
                  blk = factor(blk.idx))
dat$blk[1] <- NA
head(dat)
```

```
##            y  blk
## 1 -0.4926035 <NA>
## 2  1.2595595   10
## 3  4.0909867    1
## 4  2.7541333    7
## 5 -2.9540342   12
## 6  1.5301365   17
```

```r
res <- remlf90(y ~ 1, random = ~ blk, data = dat)

sum(model.matrix(res)$blk[1,])
```

```
## [1] 0
```

As a consequence, the predicted phenotype will be based on the remaining available effects. In this case, the global mean.

```r
fitted(res)[1] == fixef(res)$Intercept[1, 'value']
```

```
##    1
## TRUE
```

The spatial block effect is another way of writing the previous experiment. So it works in the same way.

```r
coord <- expand.grid(row = 1:20, col = 1:50)
res <- remlf90(y ~ 1,
               spatial = list(model = 'blocks',
                              coord = coord,
                              id    = 'blk'),
               data = dat)

c(sum(model.matrix(res)$spatial[1,]) == 0,
fitted(res)[1] == fixef(res)$Intercept[1, 'value'])
```

```
##         1
## TRUE TRUE
```

However, the empirical residuals of the individuals with missing values of the random effects will have an increased variance. We can show that by replicating the previous experiment and computing the variance of the residual for the first observation.

```
resid_sample <- replicate(1e3, sample_first_residual())
var(resid_sample)
```

```
## [1] 3.262426
```

This can be important when fitting several random effects. See below.

Missing values in genetic effects

For an additive genetic effect, the relationship between individuals is given in the pedigree. It is legitimate not knowing the relatives for some individual. This is what happens with founders, for example.

Use `NA` for unknown relatives. If both are unknown (e.g. founders), the genetic effect (Breeding Value) will be predicted based on its phenotype, the other effects, and the estimated heritability.

```
dat <- breedR.sample.phenotype(
  fixed = c(mu = 10, x = 2),
  genetic = list(model    = 'add_animal',
                 Nparents = c(10, 10),
                 sigma2_a = 2,
                 check.factorial = FALSE),
  N = 1e3)
head(dat)
```

```
##   self sire dam X.mu       X.x          BV      resid phenotype
## 1    1   NA  NA    1 0.9781896  1.48758459 -1.2836744 12.160289
## 2    2   NA  NA    1 0.4756239  3.00525741 -0.1231577 13.833348
## 3    3   NA  NA    1 0.5835971  0.01768209 -1.5475826  9.637294
## 4    4   NA  NA    1 0.7014859  0.12582218 -0.6011947 10.927599
## 5    5   NA  NA    1 0.4610591 -2.74823624 -0.4810666  7.692815
## 6    6   NA  NA    1 0.4729850  0.80748221  0.1783341 11.931786
```

```
res <- remlf90(phenotype ~ 1 + X.x,
               genetic = list(model = 'add_animal',
                              pedigree = dat[, 1:3],
                              id    = 'self'),
               data = dat)

str(ranef(res)$genetic)
```

```
##  atomic [1:1020] 0.929 2.701 -0.759 -0.452 -3.207 ...
##  - attr(*, "se")= num [1:1020] 0.461 0.474 0.467 0.468 0.471 ...
```

Important issue Having random effects with missing values in **combination** with genetic models, can yield spurious predictions of Breeding Values. This is due to the higher variability of the residual term, for the individuals with missing values in random effects.

Missing values in coordinates of spatial effects

Are allowed. Just like in any other random effect. For those cases, the spatial component will not participate in the prediction.

```
dat <- breedR.sample.phenotype(
  fixed = c(mu = 10, x = 2),
  spatial = list(model     = 'AR',
                 grid.size = c(10, 5),
                 rho       = c(.2, .8),
                 sigma2_s  = 1)
)
dat$Var1[1] <- NA
head(dat)
```

```
##   X.mu         X.x Var1 Var2     spatial      resid phenotype
## 1    1 0.258073758   NA    2 -1.20132917  0.4000320  9.714850
## 2    1 0.342574706    3    3 -0.60383572 -0.5985261  9.482788
## 3    1 0.075226751    7    1 -0.18285623 -0.5741285  9.393469
## 4    1 0.007293377    4    5 -1.40566332 -0.1105168  8.498407
## 5    1 0.020278299    5    5  0.08475658  1.9180202 12.043333
## 6    1 0.213822899    5    4  0.60052768 -0.1902395 10.837934
```

```
res <- remlf90(phenotype ~ 1 + X.x,
               spatial = list(model = 'AR',
                              coord = dat[, c('Var1', 'Var2')],
                              rho   = c(0.2, 0.8)),
               data = dat)

sum(model.matrix(res)$spatial[1,])
```

```
## [1] 0
```