Applied Statistical Methods - Solution 8

Peter von Rohr

2022-04-25

Problem 1: Repeated Measurements Data

Simulate a dataset with repeated measurements of `Body Weight` and `Breed`. The following dataset can be used as a basis:

```
## https://charlotte-ngs.github.io/asmss2022/data/asm_bw_flem.csv
```

The generated dataset should have the following properties

- For every observation, the ID of the animal, its `Body Weight` and its `Breed` should be contained in the dataset.
- Each animal of the given basis dataset should have 5 repeated observations of `Body Weight` and `Breed`.
- The phenotypic variance of `Body Weight` within the repeated observations of one animal should be 50% of the total phenotypic variance of `Body Weight` determined from the given basis dataset.

Your Tasks

- Analyse the generated dataset with an ANOVA
- Try to see whether you can re-cover the used input data in the results of the analysis

Solution

- Read the given basis dataset. First assign the variable with the datafile name

```
s_asm_ex08_p01_data_path <- "https://charlotte-ngs.github.io/asmss2022/data/asm_bw_flem.csv"
```

Read the data

```
tbl_ex08_p01 <- readr::read_csv(file = s_asm_ex08_p01_data_path)
```

```
## Rows: 10 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Breed
## dbl (5): Animal, Breast Circumference, Body Weight, BCS, HEI
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tbl_ex08_p01 <- dplyr::select(tbl_ex08_p01, Animal, `Body Weight`, Breed)
head(tbl_ex08_p01)
```

```
## # A tibble: 6 x 3
##   Animal `Body Weight` Breed
##    <dbl>         <dbl> <chr>
## 1      1           471 Angus
```

```
## 2      2           463 Angus
## 3      3           481 Simmental
## 4      4           470 Angus
## 5      5           496 Simmental
## 6      6           491 Simmental
```

- Loop over the records in the basis dataset and add the required number of records

```
set.seed(9875)
sd_bw <- sd(tbl_ex08_p01$`Body Weight`)
tbl_rep_obs_result <- NULL
for (idx in 1:nrow(tbl_ex08_p01)){
  tbl_rep_cur <- dplyr::bind_rows(tbl_ex08_p01[idx,],
                   tibble::tibble(Animal = c(rep(tbl_ex08_p01$Animal[idx], n_nr_rep - 1)),
                                  `Body Weight` = rnorm((n_nr_rep-1),
                                   mean = tbl_ex08_p01$`Body Weight`[idx],
                                   sd = n_sd_prop_bw * sd_bw),
                                   Breed = c(rep(tbl_ex08_p01$Breed[idx], n_nr_rep-1))))
  if (is.null(tbl_rep_obs_result)){
    tbl_rep_obs_result <- tbl_rep_cur
  } else {
    tbl_rep_obs_result <- dplyr::bind_rows(tbl_rep_obs_result, tbl_rep_cur)
  }
}
head(tbl_rep_obs_result)
```

```
## # A tibble: 6 x 3
##   Animal `Body Weight` Breed
##    <dbl>         <dbl> <chr>
## 1      1          471  Angus
## 2      1          450. Angus
## 3      1          493. Angus
## 4      1          474. Angus
## 5      1          471. Angus
## 6      2          463  Angus
```

The generated dataset is written to a file, such that it will be available for Problem 2

```
s_ex08_p01_rep_obs_data_dir <- file.path(here::here(), "docs", "data")
if (!dir.exists(s_ex08_p01_rep_obs_data_dir))
  dir.create(path = s_ex08_p01_rep_obs_data_dir, recursive = TRUE)
s_ex08_p01_rep_obs_data_path <- file.path(s_ex08_p01_rep_obs_data_dir,
                                          "asm_ex08_p01_rep_obs.csv")
if (!file.exists(s_ex08_p01_rep_obs_data_path))
  readr::write_csv(tbl_rep_obs_result, file = s_ex08_p01_rep_obs_data_path)
```

- Analyse the generated dataset with an ANOVA

```
tbl_rep_obs_result$Animal <- as.factor(tbl_rep_obs_result$Animal)
aov_ex08_p01 <- aov(`Body Weight` ~ Breed + Error(Animal), data = tbl_rep_obs_result)
(smry_aov_ex08_p01 <- summary(aov_ex08_p01))
```

```
##
## Error: Animal
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Breed      2  23641   11821   10.94 0.00702 **
## Residuals  7   7566    1081
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 40   5682   142.1
```

The results of the `aov()` analysis gives estimates for the residual error variance σ_e^2 and for the variance σ_a^2 within the repeated measurements of one animal. The estimate $\widehat{\sigma_e^2}$ is directly obtained from the section `Error: Within` in the column `Mean Sq`.

```
n_hat_sigmae2 <- smry_aov_ex08_p01$`Error: Within`[[1]]["Residuals","Mean Sq"]
```

The value for this estimate is $\widehat{\sigma_e^2} = 142.1$. The estimate for σ_a^2 is obtained by the formula

$$E(MSQ_a) = n_a * \sigma_a^2 + \sigma_e^2$$

For the expected value of the MSQ_a, we insert the `Mean Sq`-value of the `Residuals` row in the section `Error: Animal`. This leads to

$$\widehat{\sigma_a^2} = \frac{E(\widehat{MSQ_a}) - \widehat{\sigma_e^2}}{n_a}$$

The variable n_a is the number of observations for one animal. The numeric value for $\widehat{\sigma_a^2}$ is computed as

```
n_est_msqa <- smry_aov_ex08_p01$`Error: Animal`[[1]]["Residuals","Mean Sq"]
n_est_sigmaa2 <- (n_est_msqa - n_hat_sigmae2) / n_nr_rep
```

Hence, $\widehat{\sigma_a^2} = 187.8$

- Assess the results and compare them with the input used in the simulation

The variance of the residuals in the original basis dataset is obtained from

```
lm_bw_breed <- lm(`Body Weight` ~ Breed, data = tbl_ex08_p01)
smry_bw_breed <- summary(lm_bw_breed)
```

The estimate of the residual variance is 111.5 which is comparable to the value found by `aov()`. The used value for the variance within observations is given by

```
n_obs_var_bw <- var(tbl_ex08_p01$`Body Weight`)
n_obs_var_bw * n_sd_prop_bw^2
```

```
## [1] 154.5444
```

This variance is comparable to what was found by `aov()`.

Problem 2: Random Effects Model

Analyse the dataset generated in Problem 1 with a random effects model using the package `lme4`. If you had difficulties to solve Problem 1, then you can also use the following dataset.

```
## https://charlotte-ngs.github.io/asmss2022/data/asm_ex08_p01_rep_obs.csv
```

Solution

- Read generated dataset from Problem 1

```r
# read the data
tbl_ex08_p02 <- readr::read_csv(file = s_ex08_p02_data_path)
```

```
## Rows: 50 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Breed
## dbl (2): Animal, Body Weight
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# convert animal and breed to factors
tbl_ex08_p02$Animal <- as.factor(tbl_ex08_p02$Animal)
tbl_ex08_p02$Breed <- as.factor(tbl_ex08_p02$Breed)
tbl_ex08_p02
```

```
## # A tibble: 50 x 3
##    Animal `Body Weight` Breed
##    <fct>          <dbl> <fct>
##  1 1               471  Angus
##  2 1               450. Angus
##  3 1               493. Angus
##  4 1               474. Angus
##  5 1               471. Angus
##  6 2               463  Angus
##  7 2               434. Angus
##  8 2               435. Angus
##  9 2               469. Angus
## 10 2               447. Angus
## # ... with 40 more rows
```

- Analyse the data using `lme4::lmer()`

The mixed model analysis is done with

```r
lmer_ex08_p02 <- lme4::lmer(`Body Weight` ~ Breed + (1|Animal), data = tbl_ex08_p02)
summary(lmer_ex08_p02)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: `Body Weight` ~ Breed + (1 | Animal)
##    Data: tbl_ex08_p02
##
## REML criterion at convergence: 388.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.21272 -0.47641 -0.06473  0.67111  1.84776
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Animal   (Intercept) 187.8    13.70
##  Residual             142.1    11.92
## Number of obs: 50, groups:  Animal, 10
##
```

```
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)     467.951      8.489  55.127
## BreedLimousin    51.685     11.229   4.603
## BreedSimmental   21.119     12.005   1.759
##
## Correlation of Fixed Effects:
##             (Intr) BrdLms
## BreedLimosn -0.756
## BreedSmmntl -0.707  0.535
```

In the output of the `summary()` function, the formula of the model that produced the above results is shown. The REML criterion tells us that the parameter estimation process has converged to the solutions shown. The statistics on the residuals is comparable to what we have already seen in the output of the summary of the `lm()` function. The variance components were estimated with the REML method and are the same as the estimates found by `aov()`. But this is only the case when the dataset is balanced, i.e., for each animal the same number of observations are contained in the dataset.