STAT 260 - Nonparametric Statistics
Ying Sun
Statistics (STAT) Program
Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

HOMEWORK
II

Henrique Aparecido Laureano

Spring Semester 2018

Contents

Problem 1

Consider a model with two random effects of the form:

$$y_{ij} = \alpha + b_i + c_j + \epsilon_{ij},$$

where $i = 1, \dots, I$**,** $j = 1, \dots, J$**,** $b_i \sim \mathrm{N}(0, \sigma_b^2)$**,** $c_j \sim \mathrm{N}(0, \sigma_c^2)$**, and** $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ **and all these r.v.'s are mutually independent. If the model is fitted by least squares then**

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{IJ - I - J + 1}$$

is an unbiased estimator of σ^2**, where RSS is the residual sum of squares from the model fit.**

(a)

Show that, if the above model is correct, the averages $\bar{y}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}/J$ **are governed by the model:**

$$\bar{y}_{i\cdot} = a + e_i,$$

where e_i **are i.i.d.** $\mathrm{N}(0, \sigma_b^2 + \sigma^2/J)$ **and** a **is a random intercept term. Hence suggest how to estimate** σ_b^2**.**

Solution:

By averaging over each i, the random effect b_i is absorbed into the independent residual term

$$e_i = b_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij}.$$

Adding the random effect c_j to the intercept we have the random intercept a.

$$\begin{aligned} \bar{y}_{i\cdot} &= \alpha + c_j + e_i \\ &= a + e_i, \quad \text{with} \end{aligned}$$

- $e_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(0, \sigma_b^2 + \frac{\sigma^2}{J}\right)$,
- e_i's are mutually independent random variables.

From the first model we have

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{IJ - I - J + 1}.$$

From the averaged model we have

$$\hat{\sigma_b}^2 + \frac{\hat{\sigma}^2}{J} = \frac{\text{RSS}^\star}{J},$$

with RSS* being the residual sum of squares for this averaged model.

Hence, an unbiased estimator for σ_b^2 is

$$\hat{\sigma_b}^2 = \frac{\text{RSS}^\star}{J} - \frac{\text{RSS}}{J(IJ - I - J + 1)}.$$

□

(b)

Show that the averages $\bar{y}_{\cdot j} = \frac{1}{I}\sum_{i=1}^{J} y_{ij}$ are governed by the model:

$$\bar{y}_{\cdot j} = a^{'} + e_j^{'},$$

where the $e_j^{'}$ are i.i.d. $\text{N}(0, \sigma_c^2 + \sigma^2/I)$ **and $a^{'}$ is a random intercept parameter. Suggest how to estimate σ_c^2.**

Solution:

By averaging over each j, the random effect c_j is absorbed into the independent residual term

$$e_j^{'} = c_j + \frac{1}{I}\sum_{i=1}^{I} \epsilon_{ij}.$$

Adding the random effect b_i to the intercept we have the random intercept $a^{'}$.

$$\begin{aligned} \bar{y}_{\cdot j} &= \alpha + b_i + e_j^{'} \\ &= a^{'} + e_j^{'}, \qquad \text{with} \end{aligned}$$

- $e_j^{'} \overset{\text{i.i.d.}}{\sim} \text{N}\left(0, \sigma_c^2 + \frac{\sigma^2}{I}\right)$,
- $e_j^{'}$'s are mutually independent random variables.

From the first model we have

$$\hat{\sigma}^2 = \frac{\text{RSS}}{IJ - I - J + 1}.$$

From the averaged model we have

$$\hat{\sigma_c}^2 + \frac{\hat{\sigma}^2}{I} = \frac{\text{RSS}^\star}{I},$$

with RSS* being the residual sum of squares for this averaged model.

Hence, an unbiased estimator for σ_b^2 is

$$\hat{\sigma_c}^2 = \frac{\text{RSS}^\star}{I} - \frac{\text{RSS}}{I(IJ - I - J + 1)}.$$

$\square$

Problem 2

(a)

Show that if X and Z are independent random vectors, both of the same dimension, and with covariance matrices Σ_x and Σ_z, then the covariance matrix of $X+Z$ is $\Sigma_x+\Sigma_z$.

<u>Solution:</u>

In the diagonal of the covariance matrix of $X + Z$ we have the variance $\mathbb{V}(X + Z)$ and in the off-diagonal we have the covariances. If X and Z are independent $\text{Cov}(X + Z) = \Sigma_{x+z} = 0$, so by definition

$$\mathbb{V}(X + Z) = \mathbb{V}X + \mathbb{V}Z + 2\text{Cov}(X + Z) = \mathbb{V}X + \mathbb{V}Z + 2 \cdot 0 = \mathbb{V}X + \mathbb{V}Z,$$

that in matrix context is equivalent to write $\Sigma_x + \Sigma_z$.

Therefore, for the random vectors X and Z the covariance matrix of $X + Z$ is $\Sigma_x + \Sigma_z$.

$\square$

(b)

Consider a study examining patients' blood insulin levels 30 minutes after eating, y, in relation to sugar content, x, of the meal eaten. Suppose that each of 3 patients had their insulin levels measured for each of 3 sugar levels, and that an appropriate linear mixed model for the j-th measurement on the i-th patient is

$$y_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij},$$

where $b_i \sim \text{N}(0, \sigma^2)$, $\epsilon_{ij} \sim \text{N}(0, \sigma)$, and all the random effects and residuals are mutually independent.

i.

Write this model out in matrix vector form.

Solution:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, 3$$

with (for $j = 1, 2, 3$)

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix}, \quad \mathbf{X}_i = \mathbf{I}_3, \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha + \beta_1 \\ \alpha + \beta_2 \\ \alpha + \beta_3 \end{pmatrix}, \quad \mathbf{Z}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix}.$$

□

ii.

Find the covariance matrix for the response vector y.

Solution:

$$\begin{aligned}
\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_{i'}) &= \mathrm{Cov}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\epsilon}_i, \mathbf{X}_{i'}\boldsymbol{\beta} + \mathbf{Z}_{i'} b_{i'} + \boldsymbol{\epsilon}_{i'}) \\
&= 0, \quad i \neq i', \\
\mathbb{V}\mathbf{y}_i &= \mathbb{V}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\epsilon}_i) \\
&= \mathbb{V}(\mathbf{Z}_i b_i + \boldsymbol{\epsilon}_i) \\
&= \mathbf{Z}_i \mathbb{V} b_i \mathbf{Z}_i^\top + \mathbb{V}\boldsymbol{\epsilon}_i \\
&= \mathbf{Z}_i \sigma^2 \mathbf{Z}_i^\top + \sigma \mathbf{I}_3 \\
&= \begin{pmatrix} \sigma + \sigma^2 & \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma + \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 & \sigma + \sigma^2 \end{pmatrix}.
\end{aligned}$$

□

Problem 3

The data frame `Gun` (library `nlme`) is from a trial examining methods for firing naval guns. Two firing methods were compared, with each of a number of teams of 3 gunners; the gunners in each team were matched to have similar physique (Slight, Average or Heavy). The response variable `rounds` is rounds fired per minute, and

there are 3 explanatory factor variables, `Physique` (levels `Slight`, `Medium` and `Heavy`); `Method` (levels `M1` and `M2`) and `Team` with 9 levels. The main interest is in determining which method and/or physique results in the highest firing rate and in quantifying team-to-team variability in firing rate.

```
# <r code> =================================================================== #
library(nlme)                                                 # loading package
data(Gun)                                                     # loading dataset
# </r code> ================================================================== #
```

(a)

Identify which factors should be treated as random and which as fixed, in the analysis of these data.

Solution:

Fixed: `Method` & `Physique`
: We have interest in compare the levels of this variables. The interest is do inference about, take conclusions. Therefore, this factors should be treated as fixed effect.

Random: `Team`
: The interest about this variable is quantify and control variability. The levels are random, we don't have the interest in expand or generalize the conclusions about this factors to "all the population of possibly teams". Therefore, this factor should be treated as random effect.

□

(b)

Write out a suitable mixed model as a starting point for the analysis of these data.

Solution:

$$\text{rounds}_{ijk} = \mu + \text{method}_i + \text{physique}_j + a_k + b_{ik} + c_{jk} + \epsilon_{ijk}, \qquad i = 1, 2; \quad j = 1, 2, 3; \quad k = 1, \ldots, 9;$$

with

- $\text{a}_k \sim \text{N}(0, \sigma^2_{\text{intercept}})$; $\text{b}_{ik} \sim \text{N}(0, \sigma^2_{\text{method}})$; $\text{c}_{jk} \sim \text{N}(0, \sigma^2_{\text{physique}})$,
- $\epsilon_{ijk} \sim \text{N}(0, \sigma^2)$,
- all the random effects and residuals are mutually independent.

The dataset `Gun` have a "groupedData" class, the random effect for `Team` is already defined in the object structure. I'm specifying here this random effect structure.

□

(c)

Analyse the data using `lme` in order to answer the main questions of interest. Include any necessary follow-up multiple comparisons (as in the previous question) and report your conclusions.

Solution:

```
# <r code> =================================================================== #
                                                # fitting the model specified in (b)
# the dataset Gun have a "groupedData" class, the random effect for Team is alrea-
model <- lme(rounds ~ Method + Physique, Gun) # dy defined in the object structure
# </r code> ================================================================== #
```

Looking to the goodness of fit:

```
# <r code> =================================================================== #
library(latticeExtra)                                    # loading graphical library
res <- residuals(model, type = "pearson")                        # pearson residuals
fit <- fitted(model)                                                 # fitted values
print(                                         # graphical analysis of goodness of fit
  xyplot(res ~ fit, col = 1
         , xlab = "Fittted values", ylab = "Pearson residuals", main = "Model fit"
         , panel = function(...){
           panel.xyplot(...)
           panel.loess(fit, res, lwd = 3, col = "#0080ff")})
  , position = c(0, .5, .5, 1), more = TRUE)
print(
  xyplot(sqrt(abs(res)) ~ fit, col = 1, xlab = "Fitted values"
         , ylab = "Pearson residuals", main = "Mean/variance relation"
         , panel = function(...){
           panel.xyplot(...)
           panel.loess(fit, sqrt(abs(res)), lwd = 3, col = "#0080ff")})
  , position = c(.5, .5, 1, 1), more = TRUE)
print(
  xyplot(Gun$rounds ~ fit, col = 1, xlab = "Fitted values"
         , ylab = "Rounds fired per minute", main = "Observed x fitted values"
         , panel = function(...){
           panel.xyplot(...)
           panel.loess(fit, Gun$rounds, lwd = 3, col = "#0080ff")})
  , position = c(0, 0, .5, .5), more = TRUE)
print(
  qqmath(res, col = 1, xlab = "Theoretical quantile", ylab = "Sampling quantile"
         , main = "Normality") +
    layer(panel.qqmathline(res, lwd = 3, col = "#0080ff"))
  , position = c(.5, 0, 1, .5))
# </r code> ================================================================== #
```

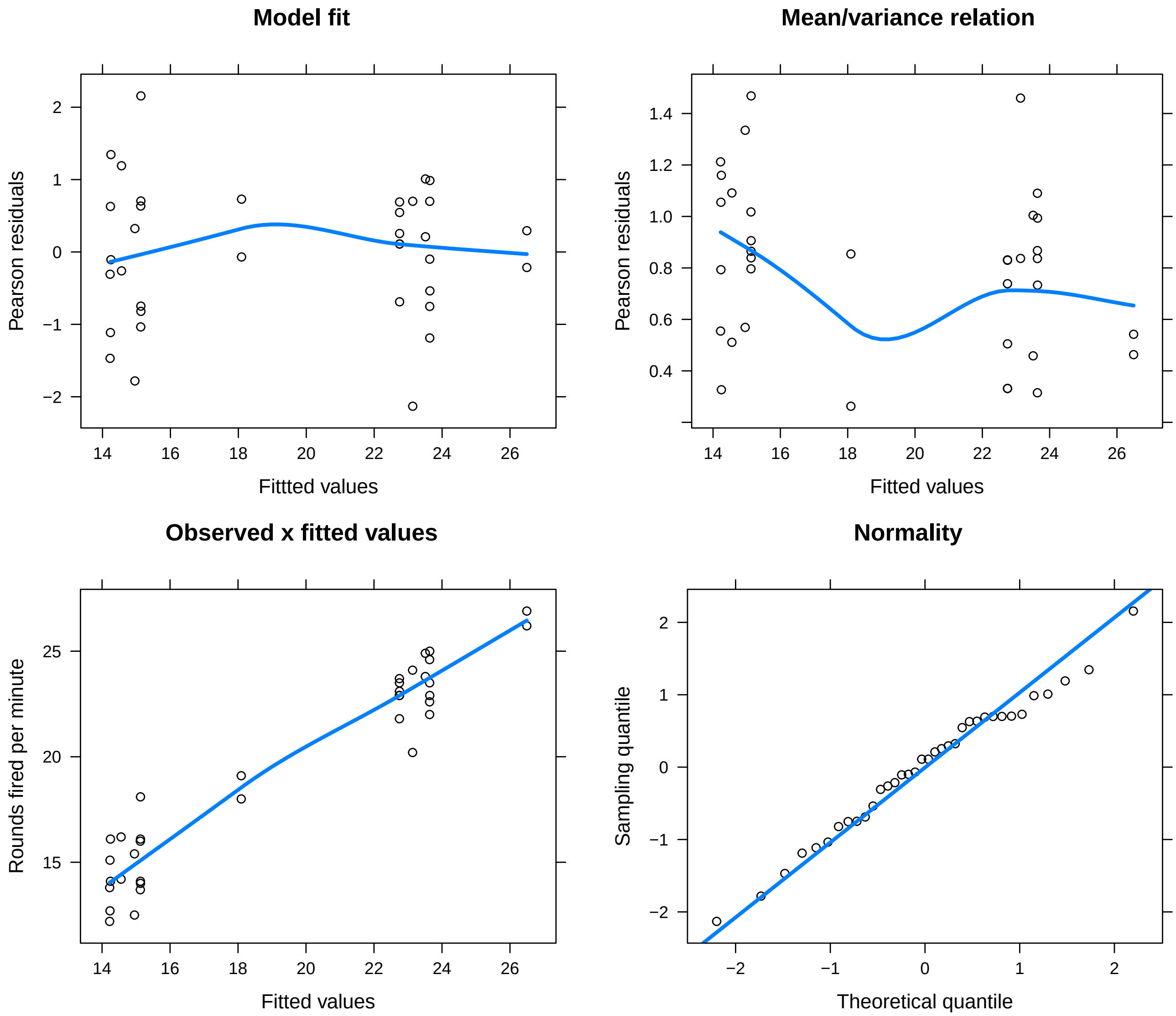

Figure 1: Graphical analysis of goodness of fit.

Given the small sample size, 36 observations, the behaviours observed in Figure 1 are very satisfatory, with no considerable deviation from the model assumptions (homocedasticity, "constant" mean/variance relation, normality of residuals).

After this verifications we are able to take conclusions about the model estimates.

Performing a variance analysis tables we see that `Physique` is not significant.

```
# <r code> ===================================================================== #
anova(model)                                                # variance analysis table
# </r code> ==================================================================== #
```

```
            numDF denDF  F-value p-value
(Intercept)     1    26 4827.168  <.0001
Method          1    26  343.511  <.0001
Physique        2     6    1.697  0.2606
```

Looking to the individual coefficients t-tests we can see better as we don't have a significant difference between the `Physique` levels.

```
# <r code> =================================================================== #
round(summary(model)$tTable, 3)                          # fixed effects estimates
# </r code> ================================================================== #
```

```
             Value Std.Error DF t-value p-value
(Intercept) 23.587     0.495 26  47.684   0.000
MethodM2    -8.511     0.460 26 -18.487   0.000
Physique.L  -1.149     0.904  6  -1.271   0.251
Physique.Q  -0.063     0.615  6  -0.103   0.921
```

In Figure 2 the boxplots for `Physique` levels are presented. The model results confirm what the figure show. The rounds fired per minute from one `Physique` level to the other are very similar.

```
# <r code> =================================================================== #
bwplot(rounds ~ Physique, Gun                           # boxplots for physique levels
       , xlab = "Physique", ylab = "Rounds fired per minute"
       , panel = function(...){
         panel.grid(v = 0, h = -1)
         panel.bwplot(...)})
# </r code> ================================================================== #
```

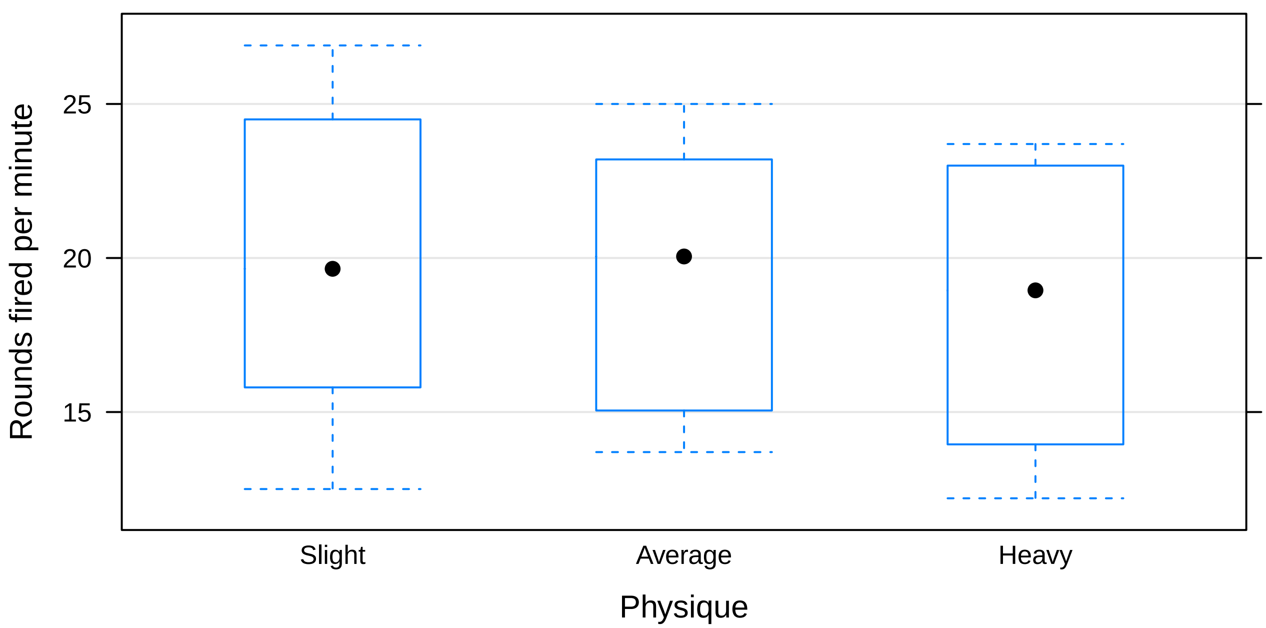

Figure 2: Boxplots for `Physique` levels.

The model present significant differences between the `Method`'s.
In Figure 3, looking to the data, we see that with the exception of T2S (level `Slight`) the medians don't differ so much. What differ is the variability.

```r
# <r code> =================================================================== #
bwplot(rounds ~ Method | Team, Gun, layout = c(9, 1)
       , xlab = "Method", ylab = "Rounds fired per minute"
       , strip = strip.custom(bg = "white")
       , panel = function(...){
         panel.grid(v = 0, h = -1)
         panel.bwplot(...)})
# <r code> =================================================================== #
```

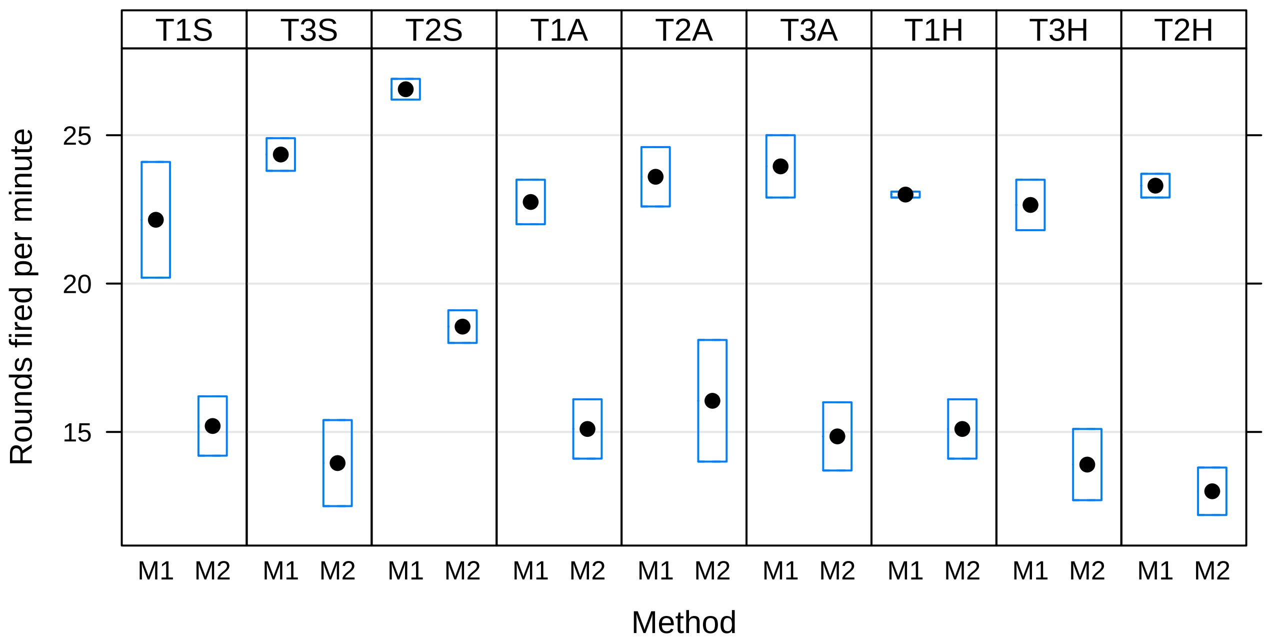

Figure 3: Boxplots for `Method` levels divided by the `Team`'s.

The random effects values are presented bellow. In the `(Intercept)` are represented the `MethodM2` and the `Physique.S`. The values differ from one level to another, which justifies the use of the random effect component.

```r
# <r code> =================================================================== #
ranef(model)                                           # extracting random effects
# <r code> =================================================================== #
```

```
    (Intercept)     MethodM2  Physique.L  Physique.Q
T1S -0.40267487 -0.064690941  0.87512523 -0.53207380
T3S -0.28073979 -0.045102590  0.61012596 -0.37095533
T2S  0.68956961  0.110782404 -1.49862760  0.91116198
T1A  0.02835393  0.004555287 -0.06162100  0.03746546
T2A  0.03280224  0.005270248 -0.07128835  0.04334332
T3A -0.01975017 -0.003173118  0.04292260 -0.02609691
T1H  0.05883425  0.009452670 -0.12786313  0.07774077
T3H -0.02301541 -0.003697796  0.05001885 -0.03041146
T2H -0.08337980 -0.013396164  0.18120743 -0.11017403
```

The biggest variability is observed in the level `Slight` and the smallest in the level `Average`.

About the fixed effect:

```
# <r code> ================================================================= #
round(summary(model)$tTable[ , 1:2], 3)                # fixed effects estimates
# </r code> ================================================================ #
```

```
             Value Std.Error
(Intercept) 23.587     0.495
MethodM2    -8.511     0.460
Physique.L  -1.149     0.904
Physique.Q  -0.063     0.615
```

For the `Method` 1 and `Slight Physique` the estimate rounds fired per minute is 23.587. If you change for the `Method` 2 the value decrease to 15.076 (23.587 - 8.511). For `Method` 1 and `Average Physique` the estimate rounds fired per minute is 22.438 (23.587 - 1.149). For `Heavy Physique` the estimate is 23.524 (23.587 - 0.063). The same reasoning is applied to achieve the other estimates. The `Method` 1 and `Slight Physique` results in the highest firing rate.

■
