HMMA 307:
Advanced Linear Modeling

Chapter 5 : Random Effects Models

ISKOUNEN SELENA AMAHJOUR WALID RUDOLF RÖMISCH

https://github.com/selenaiskounen/CA-HMMA307

Université de Montpellier

Overview

Random Effect Models

- One-Way ANOVA
- More than one Factor
- Nesting

Motivation

We consider cases where we get random samples from large populations. Since the goal is to make statements about properties of whole populations and not about observed individuals, it is rather natural to assume random samples. Now we elaborate an example with machines. We assume that we assess the quality of produced samples from some machines.

One-Way ANOVA: Model

Our model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Here we have the following variables and assumptions:

- Y_{ij} as the quality of the j-th sample on the i-th machine
- μ as the global mean
- α_i i.i.d. $\sim \mathcal{N}(0, \sigma_\alpha^2)$ as the effect of the i-th machine
- ε_{ij} i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ as the error term for the j-th sample and the i-th machine

Specialty

The model has strong similarities with the fixed models. But the $\alpha_i' s$ are random variables and not fixed unkown parameters. With that change the properties of the model will be strongly influenced.

- We get the new parameter σ_α^2
- Our model is also called variance components models since we have now two different variances σ_α^2 and σ^2

Properties

We consider some properties for our model:

- The expected value of Y_{ij} is $\mathbb{E}[Y_{ij}] = \mu$, since $\mathbb{E}[Y_{ij}] = \mathbb{E}[\mu + \alpha_i + \varepsilon_{ij}] = \mathbb{E}[\mu] + \mathbb{E}[\alpha_i] + \mathbb{E}[\varepsilon_{ij}] = \mu$
- The variance is $\mathbb{V}(Y_{ij}) = \sigma_\alpha^2 + \sigma^2$ since

$$
\begin{aligned}
\mathbb{V}(Y_{ij}) =& \mathbb{E}[Y_{ij}^2] - \mathbb{E}[Y_{ij}]^2 \\
=& \mathbb{E}[\mu^2] + \mathbb{E}[\mu\alpha_i] + \mathbb{E}[\mu\varepsilon_{ij}] + \mathbb{E}[\alpha_i\mu] + \mathbb{E}[\alpha_i\alpha_i] \\
& + \mathbb{E}[\alpha_i\varepsilon_{ij}] + \mathbb{E}[\varepsilon_{ij}\mu] + \mathbb{E}[\varepsilon_{ij}\alpha_i] + \mathbb{E}[\varepsilon_{ij}\varepsilon_{ij}] - \mu^2 \\
=& \mu^2 + \mu \underbrace{\mathbb{E}[\alpha_i]}_{=0} + \mu \underbrace{\mathbb{E}[\varepsilon_{ij}]}_{=0} + \underbrace{\mathbb{E}[\alpha_i]}_{=0} \mu + \mathbb{E}[\alpha_i^2] \\
& + \underbrace{\mathbb{E}[\alpha_i]}_{=0} \underbrace{\mathbb{E}[\varepsilon_{ij}]}_{=0} + \underbrace{\mathbb{E}[\varepsilon_{ij}]}_{=0} \mu + \underbrace{\mathbb{E}[\varepsilon_{ij}]}_{=0} \underbrace{\mathbb{E}[\alpha_i]}_{=0} + \mathbb{E}[\varepsilon_{ij}^2] - \mu^2 \\
& = \mathbb{E}[\alpha_i^2] + \mathbb{E}[\varepsilon_{ij}^2] = \sigma_\alpha^2 + \sigma^2
\end{aligned}
$$

For the correlation structure we get:

$$D_{it} = \begin{cases} 0 & i \neq k \\ \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma^2) & i = k, j \neq l \\ 1 & i = k, j = l \end{cases}$$

Here we can see that observations from different machines are uncorrelated (here independent) while observations from the same machine are correlated.
$\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma^2)$ is also called the intraclass correlation (ICC). For $\sigma_\alpha^2 >> \sigma^2$, the ICC gets "large" what means that observations from the same "group" are very similar to each other.
So values "sharing" the same $"\alpha_i"$ are correlated, while in the fixed effects model all values would be independent (because there the α_i are parameters, that is fixed, unknown quantities).

Parameter estimation for the variance components σ_α^2 and σ^2 is typically being done with a technique called restricted maximum likelihood (REML).

It is also possible to use the "classical" maximum-likelihood estimator but REML estimates are less biased. The parameter μ is estimated with maximum-likelihood assuming that the variances are known.

Now we can extend our model to the two-way ANOVA situation. As an example we have a manufacturer who is developing a new spectrophometer for medical labs. A critical issue is consistency of measurements from day to day among different machines. The response is the triglyceride level [mg/dl] of a sample.

More than One Factor: Model

We use the model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij},$$

for the random effects we have the usual assumptions

- Y_{ijk} is the triglyceride level of the k-th sample on day i and machine j,
- α_i i.i.d. $\sim \mathcal{N}(0, \sigma_\alpha^2)$ is the random effect of day,
- β_j i.i.d. $\sim \mathcal{N}(0, \sigma_\beta^2)$ is the random effect of machine,
- $(\alpha\beta)_{ij}$ i.i.d. $\sim \mathcal{N}(0, \sigma_{\alpha\beta}^2)$ random interaction term between day and machine.

Motivation

Now we introduce a new concept. We consider the strength of a chemical paste product which was measured for a total of sixty samples coming from ten randomly selected delivery batches each containing three randomly selected casks. Now we can find a new way of combining factors. Cask one in batch 1 has nothing to do with cask one in batch two and so on. The "1" of cask has a different meaning for every batch. Hence, cask and batch are not crossed. We say cask is nested in batch.
The battch effect does not seem to be very pronounced. On the other side, casks within the same batch can be substantially different. Let us now model this with appropriate random effects.

Nesting: Model

We can use the model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{k(ij)}.$$

- Y_{ijk} is the strength of the k-th sample of cask j in batch i,
- α_i i.i.d. $\sim \mathcal{N}(0, \sigma_\alpha^2)$ is the (random) effect of batch,
- $\beta_{j(i)}$ i.i.d. $\sim \mathcal{N}(0, \sigma_\beta^2)$ is the (random) effect of cask within batch,
- $\varepsilon_{k(ij)}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ is the "usual" error term.

Here we can see the special notation for $\beta_{j(i)}$ and $\varepsilon_{k(ij)}$ which emphasizes that cask is nested in batch.