HMMA307

REstricted Maximum Likelihood

Article: Nikolay Oskolkov *Scribes: Ophélie Coiffier*

1 Introduction

This project speaks about the REstricted Maximum Likelihood (REML).
When we calculate a variance estimator with the Maximum Likelihood, we must check if the estimator isn't biased. Actually, in many cases, it is biased. The obtained value with the Maximum Likelihood method overestimates (or underestimates) the true value. That's why we need to calculate the variance estimator with REML method.
In the first part we will illustrate the issue and we will give an answer to the question :

How the REML approach affects the linear mixed model ?

Then, we will mathematically explain and solve the problem.

2 Illustration of our problem

In the first part, we illustrate the variance problem with an example.

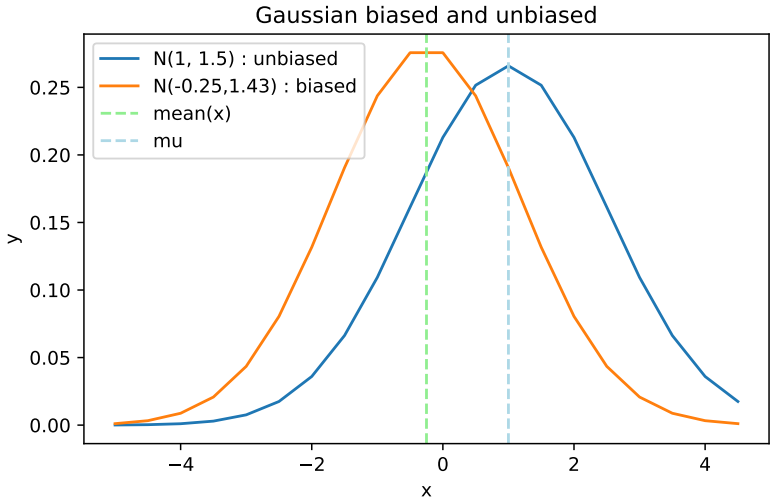


FIGURE 1 – Difference between biased Normal distribution and Unbiased Normal distribution.

The Figure 1 shows us the difference between an unbiased Normal distribution and a biased Normal distribution. We see that the mean and the variance are different. On the Figure 1, we use the next data to know the mean of the second Gaussian. It's the estimation of the mean (we will demonstrate it in the next part). Now, we use real data.

Ind column is the group of the individual, *Resp* column is the treatment response and *Treat* column is an

Ind	**Resp**	**Treat**
1	10	0
1	25	1
2	3	0
2	6	1

indicator (if the individual gets the treatment, he has 1 else he has 0).
The model used in linear regression is the impact of the treatment on the response.

$$Y_{Resp} = \mu + \beta X_{Treat} + \varepsilon$$

Where ε is a noise following a centered, reduced Normal distribution ($\mathcal{N}(0,1)$).
The model used in linear mixed effects is :

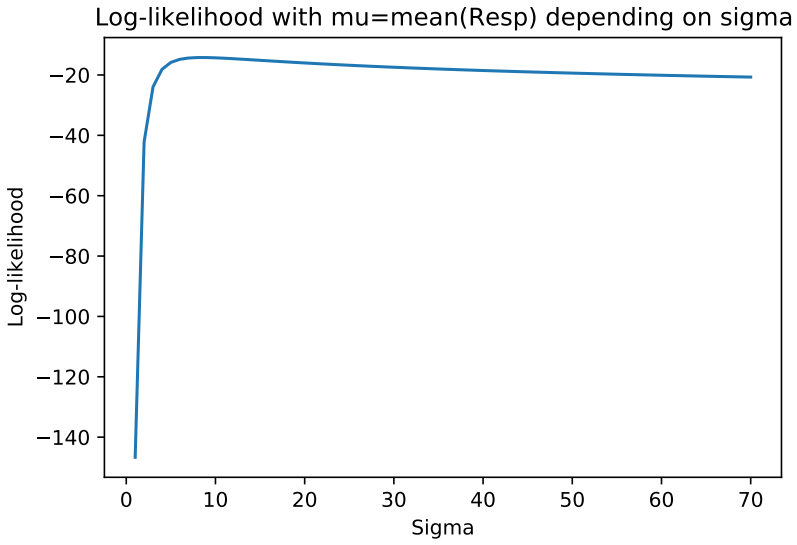
$$Y_{Resp} = \mu + \beta X_{Treat} + \alpha X_{Ind} + \varepsilon$$

When we compare the log-likelihood with both methods, we don't obtain the same result. With the linear regression, we have -14.23 (we can observe this in the Figure 2, at the point $(8, -14.23)$) while we find -7.89 when we use the linear mixed effects regression (REML).

Remark. *The code is available in the repository called REML but we notice lines we use to show the log-likelihood comparison*

```
linear_reg = sm.OLS(df.Resp, df.Treat)
linear_reg_fit = linear_reg.fit()
linear_reg_fit.summary()
```

```
mixed_random = smf.mixedlm("Resp~Treat", df, groups = df['Ind'])
mixed_fit = mixed_random.fit()
mixed_fit.summary()
```



OLS Regression Results

Dep. Variable:	Resp	R-squared (uncentered):	0.624
Model:	OLS	Adj. R-squared (uncentered):	0.499
Method:	Least Squares	F-statistic:	4.979
Date:	Mon, 26 Oct 2020	Prob (F-statistic):	0.112
Time:	19:57:12	Log-Likelihood:	-14.239
No. Observations:	4	AIC:	30.48
Df Residuals:	3	BIC:	29.87
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Treat	15.5000	6.946	2.231	0.112	-6.606	37.606

Omnibus:	nan	Durbin-Watson:	0.687
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.587
Skew:	-0.782	Prob(JB):	0.746
Kurtosis:	1.965	Cond. No.	1.00

FIGURE 2 – The log-likelihood using linear regression depending on σ and the result of the linear regression with Python function.

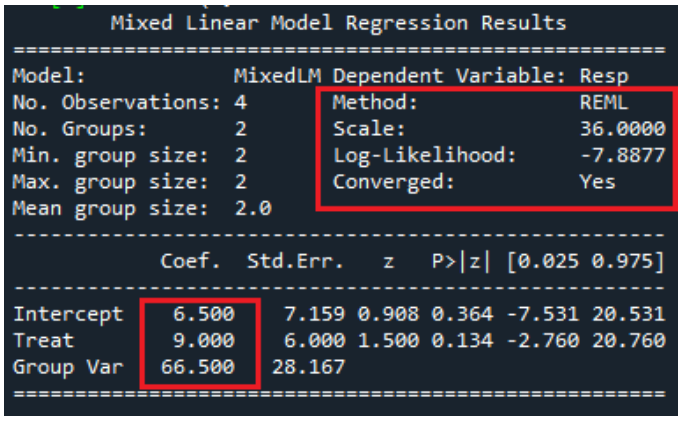


FIGURE 3 – The log-likelihood using REML regression.

Moreover, we notice the different value of the coefficient : $\sigma^2 = \sqrt{scale} = \sqrt{36.00} = 6.00$
$\sigma_s^2 = \sqrt{GroupVar} = \sqrt{66.5} = 8.15$, $\beta_1 = 6.0$ and $\beta_2 = (6.5 + 9.0) = 15.5 = coef.Treat(OLSregression)$
Now, we admit the next results and we demonstrate it in the other parts of this report.
Thanks to that values, we will compare our results with Python function results.
We write out :

$$Y = \begin{pmatrix} 3 & 10 \\ 6 & 25 \end{pmatrix}$$

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

$$\Sigma_y = \begin{pmatrix} \sigma^2 + \sigma_s^2 & \sigma_s^2 & 0 & 0 \\ \sigma_s^2 & \sigma_s^2 + \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 + \sigma_s^2 & \sigma_s^2 \\ 0 & 0 & \sigma_s^2 & \sigma_s^2 + \sigma^2 \end{pmatrix}$$

$$|\Sigma_y| = 4\sigma_s^4\sigma^4 + 4\sigma_s^2\sigma^6 + \sigma^8$$

Where, Σ_y is the variance-covariance matrix, $|\Sigma_y|$ its determinant.
So, we can maximize the integrated log-likelihood (REML purpose, detailed in next part).
The code is available in the Github field (REML.py).
Actually, we define a function calculating the integrated log-likelihood with two parameters that aren't fixed : σ^2 and σ_s^2. We know β thanks to the linear regression and Y is represented by the *Resp* values. The implemented function is :

$$log(\int L(\beta, \Sigma_y)d\beta) = -\frac{1}{2}log(2\pi) - \frac{1}{2}log(|\Sigma_y|) - \frac{(Y - X\beta)^\intercal \Sigma_y^{-1}(Y - X\beta)}{2} - \frac{1}{2}log(|X^\intercal \Sigma_y^{-1} X|)$$

(This equality will be explained in the next part). Then, we search the maximum of this function. Our results are the value of σ^2 and σ_s^2 which maximize the function.
We obtain : $log(\int L(\beta, \Sigma_y)d\beta) = -6.05$, $\sigma_s^2 = 8.15$ and $\sigma^2 = 6.00$.
There are the same results that the values calculating with linear regression.

3 The biased variance problem

When we calculate the maximum likelihood, we transform the likelihood to log-likelihood, then we derive and we equate to zero. And, obviously, we calculate the second derivative of the log-likelihood to check the sign. We need a negative sign to have a maximum.

Example. *First, we calculate the maximum likelihood in* 1 *dimension.*
We take

$$y = (y_1, \cdots, y_N) \sim \mathcal{N}(\mu, \sigma^2)$$

where μ *is the mean and* σ^2 *is the variance.*
The likelihood is

$$L(y, \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

The log-likelihood is

$$l(y, \mu, \sigma^2) = -\frac{N}{2} ln(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{2\sigma^2}$$

Now, we derive this function and equate to zero to find the maximum

$$\begin{cases} \frac{\partial}{\partial \mu} \mid_{\hat{\mu}} l(y, \mu, \sigma^2) = 0 \\ \frac{\partial}{\partial \sigma^2}) \mid_{\widehat{\sigma^2}} l(y, \mu, \sigma^2) = 0 \end{cases}$$

We assure that we have a negative second derivative.
Finally, we find

$$(\hat{\mu}, \widehat{\sigma^2}) = (\frac{1}{N} \sum_{i=1}^{N} y_i, \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\mu})^2)$$

Usually, we stop there, we have our answer but if we calculate the expected value of the variance estimator, we should obtain the variance estimator (if it's an unbiased estimator).

Example. *Return to the previous example. We need to know if the variance estimator is unbiased.*
We write out $\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \hat{\mu}$

$$\begin{aligned} \mathbb{E}[\widehat{\sigma^2}] = \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\mu})^2] &= \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} (y_i - \mu + \mu - \hat{\mu})^2] \\ &= \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} ((y_i - \mu)^2 + (\hat{\mu} - \mu)^2 - 2(y_i - \mu)(\hat{\mu} - \mu))] \end{aligned}$$

But, we have $\hat{\mu} - \mu = \frac{1}{N} \sum_{i=1}^{N} y_i - \mu = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)$
So, $\sum_{i=1}^{N} (y_i - \mu) = N(\hat{\mu} - \mu)$
Let's return to our expected value

$$\mathbb{E}[\widehat{\sigma^2}] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[(y_i - \mu)^2] - \frac{2}{N} \mathbb{E}[N(\hat{\mu} - \mu)^2] + \mathbb{E}[(\hat{\mu} - \mu)^2] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[(y_i - \mu)^2] - \mathbb{E}[(\hat{\mu} - \mu)^2]$$

We need to find the variance of $(y_i - \mu)$ *to carry on our calculation.*

$$Var(y_i - \mu) = \mathbb{E}[(y_i - \mu)^2] - (\mathbb{E}[(y_i - \mu)])^2 = \mathbb{E}[(y_i - \mu)^2] = \sigma^2$$

Finally, we have these equations

$$\mathbb{E}[\widehat{\sigma^2}] = \sigma^2 - \mathbb{E}[(\hat{\mu}-\mu)^2] = \sigma^2 - \mathrm{Var}(\hat{\mu}-\mu) = \sigma^2 - \frac{1}{N^2}\mathrm{Var}(\sum_{i=1}^{N} y_i) = \sigma^2\frac{N-1}{N} \neq \sigma^2$$

The variance estimator is biased (underestimated the true variance because $\frac{N-1}{N} < 1$*). To remove the bias, we can change the variance estimator :* $\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{\mu})^2$.

We also put these calculations in a higher dimension. In the real life, e.g our illustration, the dimension isn't 1 but k where $k > 1$.
The model writes out $Y = X\beta + \varepsilon$. where ε follows a Normal distribution $\mathcal{N}(0, \sigma^2 I_k)$ and Y follows a Normal distribution $\mathcal{N}(X\beta, \sigma^2 I_k)$.
The log-likelihood becomes

$$l(\beta, \sigma_2) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^\intercal(Y - X\beta)$$

We don't realize all calculations but it the same principle than with 1 dimension. So, we obtain these estimators

$$\hat{\beta} = (X^\intercal X)^{-1}X^\intercal Y$$

and

$$\widehat{\sigma^2} = \frac{1}{N}(Y - X\hat{\beta})^\intercal(Y - X\hat{\beta})$$

We calculate the expected value of the variance estimator.

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma^2}] = \frac{1}{N}\mathbb{E}[(Y - X\hat{\beta})^\intercal(Y - X\hat{\beta}] &= \frac{1}{N}\mathbb{E}[Y^\intercal(I_k - A)(I_k - A)Y] \\
&= \frac{1}{N}\mathbb{E}[Y^\intercal(I_k - A)Y] \\
&= \frac{1}{N}\mathbb{E}[Y^\intercal Y - Y^\intercal AY] \\
&= \frac{1}{N}(\mathbb{E}[Y^\intercal Y] - \mathbb{E}[Y^\intercal AY]) \\
&= \frac{1}{N}(N\sigma^2 + (X\beta)^\intercal(X\beta) - (k\sigma^2 + (X\beta)^\intercal(X\beta)) \\
&= \frac{N-k}{N}\sigma^2
\end{aligned}$$

where $k = rg(A)$ is the number of X's column.
We notice $A = X(X^\intercal X)^{-1}X^\intercal$ and consequently, $AY = X\hat{\beta} = \hat{Y}$

Finally, we have a bias, but if we choose $\widehat{\sigma^2} = \frac{1}{N-k}(Y - X\hat{\beta})^\intercal(Y - X\hat{\beta})$, we don't have a bias.

Thanks to these calculations, we see that the maximum likelihood is a good method when $k << N$. But the biased results are obtained when $N \approx k$.

That's why, we use the REML method.

4 Solve the bias problem

The main issue is that we use an unknown estimator for the mean. The principle that we will use is : if the log-likelihood has any information about the mean, we can optimize it and find a unbiased variance estimator.

- First step : integrate the likelihood in relation to μ and calculate the log of this integration. The parameter μ will be removed from the equation.
- Second step : use Taylor development in the log-likelihood to simplify the formula.
- Third step : separate the formula in 2 parts (the log-likelihood use in the maximum likelihood and the bias, the REML approach).
- Fourth step : finish previous calculations to find the unbiased estimator.

Example. *We continue with the previous example (the Normal distribution and $\beta \in \mathbb{R}^{2\times 2}$). The likelihood is*

$$L(\beta, \sigma_s^2, \sigma^2) = \frac{1}{\sqrt{2\pi|\Sigma_y|}} e^{-\frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2}}$$

where σ_s^2 represents standard deviation of random effects, σ^2 represents standard deviance of residual effects and $|\Sigma_y|$ is the variance-covariance matrix.
First step :

$$log(\int L(\beta, \Sigma_y) d\beta) = -\frac{1}{2} log(2\pi) - \frac{1}{2} log(|\Sigma_y|) + log(\int e^{-\frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2}} d\beta)$$

We write out $f(\beta) = -\frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2}$ and we use the Taylor development formula :
$f(\beta) \approx f(\hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^2 f''(\hat{\beta})$
Therefore, we obtain :

$$f(\beta) \approx -\frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2} - \frac{1}{2}\frac{(\beta-\hat{\beta})^\intercal X^\intercal \Sigma_y^{-1} X(\beta-\hat{\beta})}{2}$$

So, the new log-likelihood is :

$$\begin{aligned} log(\int L(\beta, \Sigma_y) d\beta) &= -\frac{1}{2} log(2\pi) - \frac{1}{2} log(|\Sigma_y|) + log(\int e^{-\frac{1}{2}\frac{(\beta-\hat{\beta})^\intercal X^\intercal \Sigma_y^{-1} X(\beta-\hat{\beta})}{2}} d\beta) - \frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2} \\ &= -\frac{1}{2} log(2\pi) - \frac{1}{2} log(|\Sigma_y|) - \frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2} - \frac{1}{2} log(|X^\intercal \Sigma_y^{-1} X|) \end{aligned}$$

We recognize the solution of the Maximum Likelihood

$$-\frac{1}{2} log(2\pi) - \frac{1}{2} log(|\Sigma_y|) - \frac{(Y-X\beta)^\intercal \Sigma_y^{-1}(Y-X\beta)}{2}$$

And we have the REML approach (the bias) :

$$-\frac{1}{2} log(|X^\intercal \Sigma_y^{-1} X|$$

To conclude calculations, we derivate the new log-likelihood in relation to σ^2 and equate to 0. The solution is the unbiased estimator of the variance. There are the same calculations that we realise to find the maximum likelihood.

In the first part of the document, we have seen an applied example. It shows us how the fee affects the mixed linear model.

5 Deal with linear models in depth

We can read the article and the lesson to learn more about the REML and the linear regression model : [Osk], [Sal19]
To visualize the Python code, we can go to the Github : [Coi20]

Références

[Coi20] Ophélie Coiffier. REML to Python code. https ://github.com/opheliecoiffier/REML, 2020. 8

[Osk] Nikolay Oskolkov. Maximum likelihood (ml) vs. reml. https ://towardsdatascience.com/maximum-likelihood-ml-vs-reml-78cf79bef2cf. 8

[Sal19] Joseph Salmon. HMMA 307 - modèles linéaires avancées, 2019. 8