Maximum de vraisemblance vs. Maximum de vraisemblance restreint

Mégane Diéval

Dans cet article, nous reprenons l'excellent travail mené par Nikolay Oskolkov à propos de l'estimateur du maximum de vraisemblance restreint (REML). Nous expliquerons pourquoi il est utilisé et quelle est la différence avec l'estimateur du maximum de vraisemblance (ML).

1 Introduction du problème

Supposons qu'un ensemble de N observations $X = (x_1, x_2, ..., x_N)$ ait une distribution de type gaussienne, la variable aléatoire X suit alors une loi normale de paramètres μ et σ^2, i.e. $X \sim N(\mu, \sigma^2)$. Il s'agit alors d'estimer ces paramètres. On montrera dans cet article que l'estimateur du maximum de vraisemblance donne des résultats biaisés pour $\hat{\sigma^2}$.

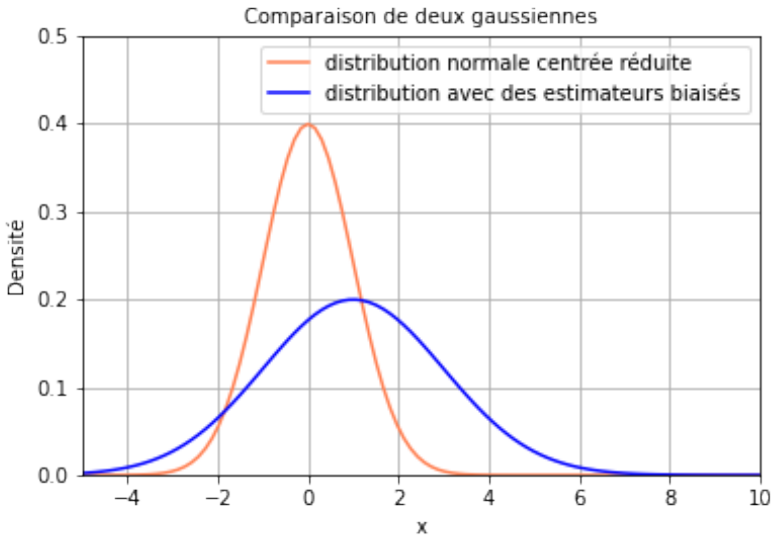

Figure 1: La courbe bleue estime (mal) la orange par des paramètres biaisés.

Exemple. Considérons maintenant un jeu de données très simple qui conserve les propriétés nécessaires d'un modèle linéaire mixte. On suppose que nous avons quatre points, deux proviennent de l'individu 1 ($Ind = 1$), et les deux autres de l'individu 2 ($Ind = 2$). La variable d'intêret est *Resp*. Pour chacun des individus, les deux données sont associées aux modalités "traités" ($Treat = 0$) et "non traitées" ($Treat = 1$) :

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

Table 1: Jeu de données décrit précédemment

Remarque. L'ensemble du code utilisé pour cet article se trouve dans le répertoire Github du projet.

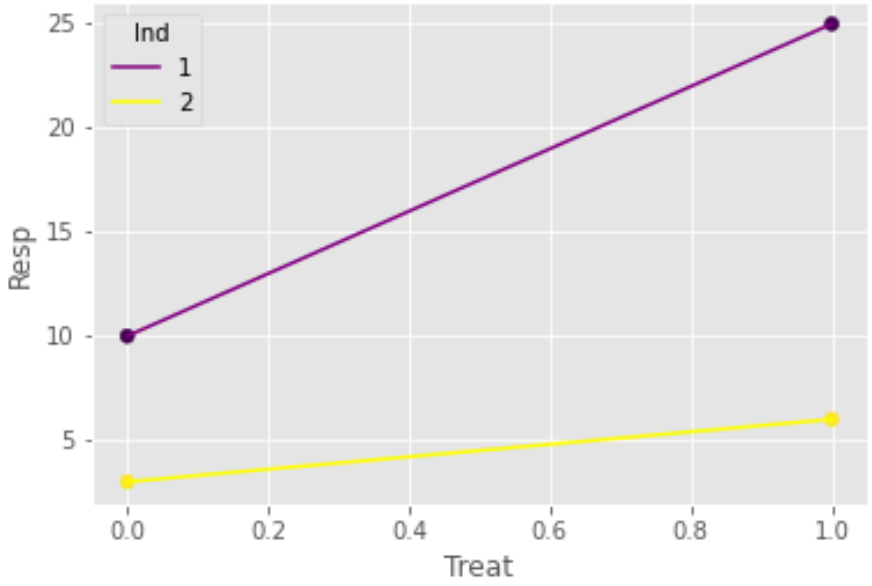

Figure 2: Représentation graphique du jeu de données

Si on étudie l'effet des variables $Treat$ et Ind sur la variable $Resp$, on peut constuire un modèle linéaire mixte (LMM):

$$Y_{Resp} = \beta X_{Treat} + \alpha K_{Ind} + \epsilon \quad (0)$$

Le traitement est modélisé comme un effet fixe tandis que les effets individuels comme un effet aléatoire. On suppose que l'effet fixe n'a pas d'erreur associée mais que l'effet aléatoire en a et suit une loi normale $N(0, \sigma_s^2 I)$. Enfin l'erreur résiduelle ϵ suit également une loi normale $N(0, \sigma^2 I)$.
On peut également réaliser une régression linéaire des moindres carrés (OLS) pour étudier l'effet de la variable $Treat$ uniquement.

Out[6]:

OLS Regression Results

Dep. Variable:	Resp	**R-squared:**	0.283
Model:	OLS	**Adj. R-squared:**	-0.075
Method:	Least Squares	**F-statistic:**	0.7902
Date:	Wed, 28 Oct 2020	**Prob (F-statistic):**	0.468
Time:	19:20:08	**Log-Likelihood:**	-13.549
No. Observations:	4	**AIC:**	31.10
Df Residuals:	2	**BIC:**	29.87
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5000	7.159	0.908	0.460	-24.302	37.302
Treat	9.0000	10.124	0.889	0.468	-34.561	52.561

Omnibus:	nan	**Durbin-Watson:**	1.176
Prob(Omnibus):	nan	**Jarque-Bera (JB):**	0.337
Skew:	0.000	**Prob(JB):**	0.845
Kurtosis:	1.579	**Cond. No.**	2.62

Figure 3: Résultats obtenus par le modèle linéaire pour l'estimateur ML

```
       Mixed Linear Model Regression Results
============================================================
Model:            MixedLM Dependent Variable: Resp
No. Observations: 4       Method:             ML
No. Groups:       2       Scale:              18.0001
Min. group size:  2       Log-Likelihood:     -13.0029
Max. group size:  2       Converged:          Yes
Mean group size:  2.0
------------------------------------------------------------
             Coef.  Std.Err.   z   P>|z| [0.025 0.975]
------------------------------------------------------------
Intercept     6.500    5.062 1.284 0.199 -3.422 16.422
Treat         9.000    4.243 2.121 0.034  0.685 17.315
Group Var    33.250   14.083
============================================================
```

(a) Résultats pour ML

Out[7]:

Model:	MixedLM	Dependent Variable:	Resp
No. Observations:	4	Method:	REML
No. Groups:	2	Scale:	36.0000
Min. group size:	2	Log-Likelihood:	-7.8877
Max. group size:	2	Converged:	Yes
Mean group size:	2.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	6.500	7.159	0.908	0.364	-7.531	20.531
Treat	9.000	6.000	1.500	0.134	-2.760	20.760
Group Var	66.500	28.167				

(b) Résultats pour REML

Figure 4: Résultats obtenus par le modèle linéaire mixte pour les deux estimateurs

On a donc constitué trois modèles :

1. Un modèle linéaire
2. Un modèle linéaire mixte avec estimateur du maximum de vraisemblance
3. Un modèle linéaire mixte avec estimateur du maximum de vraisemblance restreint

On constate déjà que que la valeur de la log-vraisemblance diffère entre les deux premiers modèles et le dernier avec une valeur de -13.5 contre -7.9. Les valeurs des coefficients de

l'intercepte et de la variable $Treat$ sont les mêmes mais les intervalles de confiances diffèrent selon les modèles. Enfin en fonction de l'estimateur utilisé pour les modèles linéaires mixtes le coefficient associé à la variable de groupe est différent.
Notons que les 4 points ne sont pas indépendants et que l'utilisation d'une régression des moindres carrés n'est pas pertinente ni adaptée aux données. Le modèle linéaire mixte prend en compte la dépendance des données pour chaque individu.

Finalement, la question que l'on peut se poser ici concerne la différence entre les deux estimateurs ML et REML.

2 Les limites de l'estimateur du maximum de vraisemblance

Dans cette partie nous allons soulever le problème essentiel de l'estimateur du maximum de vraisemblance de la variance. Ce dernier se calcule à partir de l'estimation de la moyenne qui peut avoir une erreur, ce qui entraîne un biais conséquent.

Proposition. L'estimateur du maximum de vraisemblance (ML) de σ^2 est biaisé.

Considérons le cas simple à une dimension pour démontrer que l'estimateur du maximum de vraisemblance de la variance est biaisé. Soit $y = (y_1, y_2, ..., y_N)$ suivant une loi normale centrée réduite, avec N le nombre d'observations. La vraisemblance de ce modèle est :

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

avec $\hat{\mu}$ l'estimateur de la moyenne μ, et $\hat{\sigma}^2$ l'estimateur de la variance. On choisit de travailler avec la log-vraisemblance par souci de simplicité des calculs de dérivation :

$$l(\hat{\mu}, \hat{\sigma}^2) = \log\left(L(\hat{\mu}, \hat{\sigma}^2)\right) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\hat{\sigma}^2) - \frac{\sum_{i=1}^{N}(y_i - \hat{\mu})^2}{2\hat{\sigma}^2}$$

Pour établir les équations de vraisemblance, on calcule à partir de la log-vraisemblance les dérivées du premier ordre par rapport à $\hat{\mu}$ et $\hat{\sigma^2}$:

$$\frac{\partial l(\hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\mu}} = \frac{\sum_{i=1}^{N}(y_i - \hat{\mu})}{2\hat{\sigma}^2} = 0 \quad (1)$$

$$(1) \quad \Leftrightarrow \sum_{i=1}^{N} y_i - \hat{\mu}N = 0 \;\Leftrightarrow \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\frac{\partial l(\hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = -\frac{N}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{N}(y_i - \hat{\mu})^2 = 0 \quad (2)$$

$$(2) \Leftrightarrow \frac{\partial l(\hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = -\frac{N}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{N}(y_i - \hat{\mu})^2 = 0$$

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

Soit μ la vraie valeur de la moyenne. La valeur attendue de σ^2 est donc:

$$\sigma^2 \equiv \mathrm{Var(y)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}(\mathrm{y_i} - \mu)^2$$

Toutefois, si on restructure l'expression de $\hat{\sigma^2}$, on a:

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

D'autre part on exprime le terme $\sum_{i=1}^{N}(y_i - \mu)$ en fontion de $\hat{\mu}$:

$$\hat{\mu} - \mu = \frac{1}{N}\sum_{i=1}^{N} y_i - \mu = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N}\mu = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu) \quad (3)$$

$$(3) \Rightarrow \sum_{i=1}^{N}(y_i - \mu) = N(\hat{\mu} - \mu)$$

Enfin on remplace l'expression de $\sum_{i=1}^{N}(y_i - \mu)$ obtenue dans la formule de l'estimateur du maximum de vraisemblance de la variance:

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

On calcule à partir de cette expression le biais de $\hat{\sigma^2}$, sachant qu'on connaît l'expression théorique de la variance σ^2 attendue:

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

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

$E[\hat{\sigma}^2] = \sigma^2 - \frac{1}{N^2}N\sigma^2 = \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2$

On constate que l'espérance de l'estimateur ML de la variance n'est pas égale à la variance σ^2. L'estimateur ML sous-estime la vraie variance.
En revanche, la différence entre la variance réelle et la variance estimée devient plus petite pour de grands échantillons.
Cependant, nous avons considéré ici le cas unidimensionnel le plus simple. Lorsque Y n'est pas un vecteur mais une matrice, c'est-à-dire pour des données de grande dimension, il peut être montré que l'expression précédente prend la forme :

$$E[\hat{\sigma}^2] = \frac{N-k}{N}\sigma^2$$

où k est le nombre de dimensions. Par conséquent, le problème de la sous-estimation de la vraie variance par ML devient particulièrement aigu, et l'estimateur de variance ML devient de plus en plus biaisé lorsque le nombre de dimensions k se rapproche du nombre d'observations N. Ici, nous voyons clairement que dans un espace de grande dimension, le principe du maximum de vraisemblance ML ne fonctionne bien que dans la limite $k << N$, tandis que des résultats biaisés peuvent être trouvés lorsque $k \approx N$.

3 L'estimateur REML, une solution au problème de biais

L'incertitude sur l'estimateur de la variance vient du fait que celui-ci est calculé à partir de l'estimateur de la moyenne. Notre stratégie ici va être d'exprimer la log-vraisemblance sans aucune information sur la moyenne. Un moyen de se passer des informations sur la moyenne de la fonction log-vraisemblance est de calculer une probabilité marginale, c'est-à-dire d'intégrer la log-vraisemblance sur la moyenne. Ici, nous allons intégrer la log-vraisemblance par rapport à β et obtenir une estimation sans biais pour les composantes de la variance.

On se place ici dans un espace de dimension $k = 2$, dans le contexte du modèle linéaire mixte (0) exposé dans la première partie dans l'exemple où, on le rappelle, σ^2 est la variance de l'erreur résiduelle et σ_s^2 celle de l'effet aléatoire. On a alors $Y \sim N(X\beta, \Sigma_y)$ où Σ_y vaut:

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

La vraisemblance du modèle, relative à une distribution gaussienne multivariée est:

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

On intègre cette vraisemblance par rapport à β pour calculer la "log-vraisemblance":

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

On utilise l'approche du point de selle. La fonction exponentielle sous l'intégrale décroît très rapidement, il suffit donc de calculer l'intégrale pour le minimum de la fonction $f(\beta)$ où:

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

qui donnera une contribution maximale à l'exposant de l'exponentielle et donc à la log-vraisemblance.

On obtient par un développement de Taylor: $f(\beta) \approx f(\hat{\beta}) + (1/2)(\beta - \hat{\beta})^2 f''(\hat{\beta})$. Ce développement se fait au point $\hat{\beta}$, l'estimateur ML de β. On suppose que $\hat{\beta}$ est assez proche de la vraie valeur de β de sorte que la vraisemblance soit maximale et que l'on puisse réaliser ce développement. On obtient:

$$f(\beta) = -\frac{(Y - X\beta)^T \Sigma_y^{-1} (Y - X\beta)}{2} \approx -\frac{\left(Y - X\hat{\beta}\right)^T \Sigma_y^{-1} \left(Y - X\hat{\beta}\right)}{2} - \frac{\left(\beta - \hat{\beta}\right)^T X^T \Sigma_y^{-1} X \left(\beta - \hat{\beta}\right)}{2}$$

Finalement, en réinjectant cette expression dans celle de la log-vraisemblance, on a:

$$l = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_y|) - \frac{\left(Y - X\hat{\beta}\right)^T \Sigma_y^{-1} \left(Y - X\hat{\beta}\right)}{2} + \log\left[\int \mathrm{e}^{-\frac{\left(\beta - \hat{\beta}\right)^{\mathrm{T}} \mathrm{X}^{\mathrm{T}} \Sigma_{\mathrm{y}}^{-1} \mathrm{X} \left(\beta - \hat{\beta}\right)}{2}} \mathrm{d}\beta\right]$$

$$l = \log\left[\int \mathrm{L}(\beta, \Sigma_{\mathrm{y}})\mathrm{d}\beta\right] = -\frac{1}{2}\log(|\Sigma_y|) - \frac{1}{2}\left(Y - X\hat{\beta}\right)^T \Sigma_y^{-1} \left(Y - X\hat{\beta}\right) - \frac{1}{2}\log\left(|X^T\Sigma_y^{-1}X|\right)$$

On reconnaît dans les premiers termes l'expression de la log-vraisemblance du modèle linéaire mixte pour $k > 1$. Le dernier terme $-\frac{1}{2}\log\left(|X^T\Sigma_y^{-1}X|\right)$ est issu de l'approximation REML et semble constituer le biais dans l'estimation ML classique.

Exemple. Revenons au petit jeu de données présenté précédemment en partie 1. Le modèle linéaire mixte sous forme matricielle qui lui est associé est le suivant:

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{12} \\ \epsilon_{22} \end{bmatrix} \tag{1}$$

avec $(y_{11}, y_{21}, y_{12}, y_{22}) = (3, 6, 10, 25)$.

On va estimer la matrice de variances-covariances de ce modèle à partir des formules établies précédemment. On se sert des valeurs de $(\beta_1, \beta_2) = (6.5, 15.5)$ communes aux trois modèles construits dans la partie 1. On a également besoin des expressions suivantes pour le calcul de la vraisemblance:

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

$(Y - X\beta)^T \Sigma_y^{-1} (Y - X\beta) = \frac{1}{\sigma^2(\sigma^2+2\sigma_s^2)} [(y_{11} - \beta_1)^2(\sigma^2 + \sigma_s^2) - 2(y_{11} - \beta_1)(y_{21} - \beta_2)\sigma_s^2 +$
$(y_{21} - \beta_2)^2(\sigma^2 + \sigma_s^2) + (y_{12} - \beta_1)^2(\sigma^2 + \sigma_s^2) - 2(y_{12} - \beta_1)(y_{22} - \beta_2)\sigma_s^2 + (y_{22} - \beta_2)^2(\sigma^2 + \sigma_s^2)]$

$|X^T\sigma_y^{-1}X| = \frac{4}{\sigma^2(\sigma^2+2\sigma_s^2)}$

On trouve en maximisant l'expression de la log-vraisemblance approchée (voir Illustrations.py): $\sigma^2 = 6$ et $\sigma_s^2 = 8.15$ qui donnent à la log-vraisemblance la valeur maximale de -6.049856. Dans la partie 1 (voir figure 4.*b*), on trouvait les mêmes valeurs pour les estimateurs REML de la variance puisque $\sigma^2 = \sqrt{Scale} = \sqrt{36} = 6$ et $\sigma_s^2 = \sqrt{Groupe\ var} = \sqrt{33.25} = 8.15$.

References

[1] Nikolay Oskolkov, Maximum Likelihood (ML) vs. Restricted Maximum Likelihood (REML), https://towardsdatascience.com/maximum-likelihood-ml-vs-reml-78cf79bef2cf, 2020

[2] Mégane Diéval, Maximum de vraisemblance vs. Maximum de vraisemblance restreint https://github.com/MegDie/ML_VS_REML, 2020

[3] Joseph Salmon. HMMA307 - Modèles linéaires avancés, 2020