HMM307: Modèles linéaires avancés

Maximum de vraisemblance vs. Maximum de vraisemblance restreint

Mégane Diéval

https://github.com/MegDie/ML_VS_REML

Université de Montpellier

Sommaire

Sommaire

La problématique des estimateurs biaisés

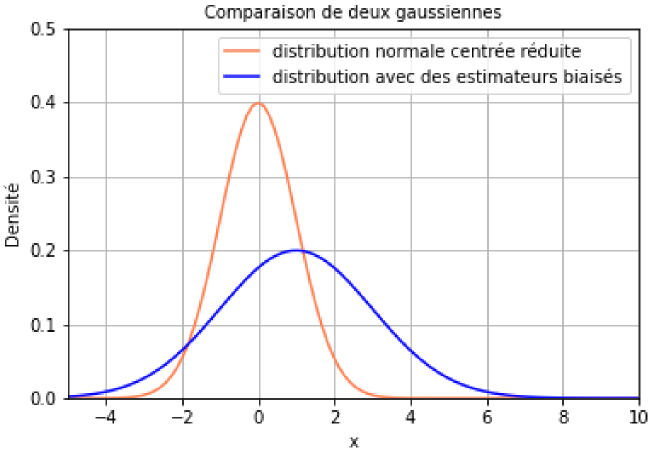

Figure: La courbe bleue est une estimation de la courbe orange par des paramètres biaisés

Rappels sur les modèles linéaires mixtes

Le modèle linéaire mixte

$$Y = X\beta + K\alpha + \epsilon$$

$$avec\, Y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times q},\ K \in \mathbb{R}^{n\times q},\ \beta \in \mathbb{R}^q,\ \alpha \in \mathbb{R}^q,\ \epsilon \in \mathbb{R}^n$$

$$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 Id_n),\ \alpha_j \sim \mathcal{N}(0, \sigma_j^2 Id_{q_j}),\ Y \sim \mathcal{N}(X\beta, Var(Y))$$

$$Var(Y) = \Sigma_y = Var(\epsilon) + Var(K\alpha)$$

- X est l'effet fixe du modèle et ne comporte pas d'erreur
- K est l'effet aléatoire du modèle
- ϵ est le bruit du modèle

Les estimateurs du maximum de vraisemblance

Estimateurs ML de β et de la variance

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

$$\hat{\Sigma}_y = \frac{1}{n} \; (Y - X\hat{\beta})^t \; (Y - X\hat{\beta})$$

L'estimation de la variance **dépend d'un estimateur qui peut comporter une erreur**.

$\mathbb{E}[\hat{\Sigma}_y] = \frac{n-q}{n} \; \Sigma_y$

- Sous-estimation de la vraie variance
- Erreur d'autant plus grande que $q \approx n$

Sommaire

L’estimateur REML, une solution

Stratégie : Exprimer la log-vraisemblance en se passant de l’information sur la moyenne.

L'estimateur REML, une solution

Stratégie : Exprimer la log-vraisemblance en se passant de l'information sur la moyenne.

- 1) Intégrer par rapport à β pour se passer de l'information estimée:

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

L'estimateur REML, une solution

Stratégie : Exprimer la log-vraisemblance en se passant de l'information sur la moyenne.

- 1) Intégrer par rapport à β pour se passer de l'information estimée:

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

- 2) Utiliser cette expression pour calculer la log-vraisemblance:

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

L'estimateur REML, une solution

Stratégie : Exprimer la log-vraisemblance en se passant de l'information sur la moyenne.

- 3) Réaliser un développement de Taylor sur l'exposant de l'exponentielle:

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

$$f(\beta) \approx f(\hat{\beta}) + (1/2)(\beta - \hat{\beta})^2 f''(\hat{\beta})$$

L'estimateur REML, une solution

Stratégie : Exprimer la log-vraisemblance en se passant de l'information sur la moyenne.

- 3) Réaliser un développement de Taylor sur l'exposant de l'exponentielle:

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

$$f(\beta) \approx f(\hat{\beta}) + (1/2)(\beta - \hat{\beta})^2 f''(\hat{\beta})$$

- 4) Identifier le biais après avoir calculé la log-vraisemblance de cette façon:

$$-\frac{1}{2}\log\left(|\Sigma_y|\right) - \frac{1}{2}\left(Y - X\hat{\beta}\right)^T \Sigma_y^{-1}\left(Y - X\hat{\beta}\right) - \frac{1}{2}\log\left(|X^T\Sigma_y^{-1}X|\right)$$

Sommaire

Présentation du jeu de données

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

Table: jeu de données

Treat : Indicatrice du traitement
Resp : Variable d'intérêt
Ind : Indicatrice d'individu

4 données qui conservent les propriétés du modèle linéaire mixte

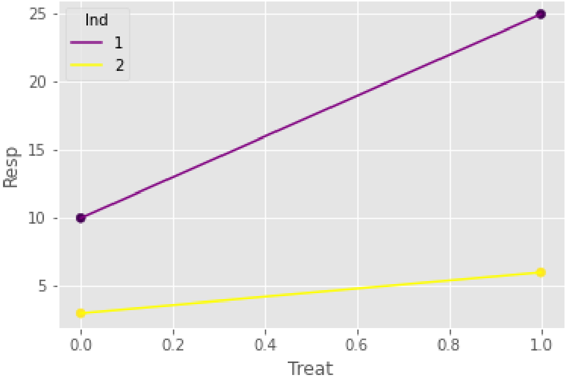

Figure: Représentation graphique du jeu de données

Modèle : $Y_{Resp} = X_{Treat}\ \beta + K_{Ind}\ \alpha + \epsilon$

Calcul des estimateurs selon les deux méthodes

- Calcul des estimateurs ML : $\hat{\beta}_1 = 6.5$ et $\hat{\beta}_2 = 15.5$

Commandes Python:

mm_ml = smf.mixedlm("Resp ˜ Treat", df, groups = df['Ind'])
result_ml = mm_ml.fit(reml=False)

- Calcul des estimateurs REML : $\hat{\beta}_1 = 6.5$ et $\hat{\beta}_2 = 15.5$

Commandes Python:

result_reml = mm_ml.fit()

Méthode	**Log-vraisemblance**	$\hat{\sigma_\epsilon}2$	$\hat{\sigma_j^2}$
REML	-7.89	6.00	8.15
ML	-13.0	4.24	5.77

Conclusion

- L'estimateur REML résout les problèmes de biais de l'estimateur ML.
- La log-vraisemblance est supérieure avec REML mais les coefficients β sont les mêmes.

Bibliographie

- Nikolay Oskolkov. Maximum Likelihood (ML) vs. Restricted Maximum Likelihood (REML), 2020
- Mégane Diéval. Maximum de vraisemblance vs. Maximum de vraisemblance restreint, 2020
- Joseph Salmon. HMMA307 - Modèles linéaires avancés, 2020