EFISIENSI MODEL CAMPURAN LINEAR DISTRIBUSI T DENGAN PROSES AUTOREGRESIF PADA DATA LONGITUDINAL

Cucu Sumarni*)
*) Departemen Statistika FMIPA Institut Pertanian Bogor
cucu_s31@yahoo.com

ABSTRAK

Data longitudinal adalah data dengan ciri khas pengukurannya dilakukan secara berulang terhadap objek amatan yang sama (repeated measurement), sehingga dalam subjek amatan antar waktu terdapat autokorelasi. Salah satu tujuan pengumpulan data longitudinal adalah untuk efisiensi sampel. Dalam penerapannya, data longitudinal sering dimodelkan dengan model campuran linear, dimana model ini sangat ketat dengan asumsi normalitas, yaitu komponen error dan komponen acaknya diasumsikan berdistribusi Normal. Padahal dalam kenyataannya, asumsi normalitas ini sulit dipenuhi, apalagi jika sampelnya kecil. Distribusi t biasanya cukup efektif dalam menggambarkan sebaran data dengan sampel yang kecil. Sementara proses Autoregresif(AR) sering digunakan untuk kasus time series yang dapat menangani masalah data objek amatan yang mengandung autokorelasi antar waktu. Sehingga alternatif model campuran linear yang dapat menangani kedua masalah tersebut adalah model campuran linear dimana bagian error-nya diasumsikan berdistribusi-t dan mengandung proses Autoregresif, AR(p). Di samping itu, model campuran linear dengan metode estimasi Restricted Maximum Likelihood(REML) juga merupakan salah satu pendekatan yang robust terhadap asumsi normalitas. Oleh karena itu, dalam makalah ini akan dikaji mengenai efisiensi kedua model campuran linear tersebut. Berdasarkan kajian simulasi data dan aplikasi data real, untuk kasus data yang mengandung korelasi serial yang cukup tinggi antar amatan, model campuran linear distribusi-t dengan proses autoregresif AR(1) lebih efisien dibanding model campuran linear metode REML. Namun dalam penghitungan estimasi parameternya secara komputasional seperti menggunakan software R, model campuran linear dengan metode REML lebih praktis.

Kata-kata kunci: autoregresif, data longitudinal, distribusi-t, efisiensi model, model campuran linear

PENDAHULUAN

Model campuran linear banyak diaplikasikan untuk analisis data kontinu yang pengukurannya dilakukan secara berulang terhadap setiap subjek amatan yang dikenal dengan data longitudinal. Model campuran linear adalah suatu pemodelan dimana variabel respon Y selain dipengaruhi oleh variabel penjelas (X) atau dikenal dengan efek fix (fixed effect), juga dipengaruhi oleh faktor keacakan dalam pemilihan sampel, waktu, area, dll yang dikenal dengan efek acak (random effect).

Pada umumnya, model campuran linear yang diterapkan kebanyakan peneliti mengasumsikan bahwa variabel acak dan error-nya berdistribusi normal. Namun pada kenyataannya, sangat mungkin terjadi pelanggaran terhadap asumsi ini ketika data diantara subjek amatan sangat bervariasi, sampel kecil atau adanya data yang outlier. Pelanggaran asumsi ini dapat mengakibatkan estimasi koefisien regresi dan komponen varians menjadi tidak akurat.

Dalam tiga dekade terakhir, distribusi t telah direkomendasikan sebagai generalisasi dari distribusi normal untuk ke-robust-an model regresi linear seperti Zelner pada tahun 1976 dan Lang et al. pada tahun 1989 dalam Lin [1] dan model campuran linear (Lin [1] dan Pinheiro, et al [2]. Selain itu, Lin dan Pinheiro [1-2] juga mengungkapkan bahwa pengukuran yang berulang dari waktu ke waktu dapat menyebabkan observasi dalam setiap individu/subjek terdapat autokorelasi. Hal ini dikenal dengan istilah serial correlation. Sehingga, Lin [1] mengusulkan model campuran linear yang mengakomodir masalah outlier dan serial correlation, yaitu model campuran linear distribusi-t yang mengandung struktur

dependensi AR(p) atau kita sebut t-linear mixed model (TLMM).

Sementara itu, Jiang [3] mengungkapkan bahwa metode REML yang merupakan salah satu metode estimasi quasi-likelihood dapat digunakan untuk mengestimasi model campuran linear baik yang mengasumsikan normalitas pada random effect dan error maupun tidak. Sebut saja model ini sebagai LMM-REML. Oleh karena itu, tujuan studi ini adalah untuk mengevaluasi model manakah yang lebih efisien antara model TLMM-AR(p) dengan model LMM-REML.

Batasan Masalah

Studi ini hanya dibatasi pada kasus data longitudinal dengan korelasi serial yang tinggi, dilihat dari estimasi parameter fixed effect dan model TLMM-AR(p) yang digunakan adalah TLMM-AR(1).

METODE DAN BAHAN

Model Longitudinal

Model longitudinal merupakan salah satu bentuk model campuran linear yang menganalisis data yang pengamatannya dilakukan berulang-ulang pada setiap subjek (repeated measurement) (Verbeke dan Molenberghs [4]). Selain itu, data longitudinal dapat dianalisis dengan model campuran linear, karena pemilihan subjek pengamatannya biasanya dilakukan secara random dari suatu populasi (McCulloch dan Searle [5]). Sehingga dalam model longitudinal, terdapat intercept random effect dan slope random effect dari variabel waktu. Contoh model longitudinal (Zhang, et al [6]):

$$y_{ij} = b_{0i} + \beta_1 age_i + \beta_2 sex_i + b_{1i} t_{ij} + e_{ij} \quad (1)$$

Dimana

y_{ij} = data kolesterol individu ke-i pada waktu ke-j

age_i = usia pada individu ke-i

sex_i = jenis kelamin pada individu ke-i

t_{ij} = variabel waktu = (j-5)/10

e_{ij} = error diasumsikan independen berdistribusi N(0,τ^2)

$b_i = (b_{0i}, b_{1i})'$ adalah random effects diasumsikan independen berdistribusi $N(0,\sigma^2)$

$\beta = (\beta_1, \beta_2)'$ adalah fixed effects

Jika $x_{ij} = (age_i, sex_i)'$ dan $z_{ij} = (1, t_{ij})$ maka persamaan (1) menjadi:

$$y_{ij} = \beta x_{ij} + b_i z_{ij} + e_{ij}$$

Berdasarkan Verbeke dan Molenberghs [4] secara umum bentuk model longitudinal adalah

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,\ i=1,\ldots,N \quad (2)$$

Dimana y_i adalah vektor observasi dari individu ke-i berdimensi n_i; X_i dan Z_i adalah matriks yang diketahui masing-masing berdimensi (n_i x q1) dan (n_i x q2); $\boldsymbol{\beta}$ adalah vektor berdimensi q1 yang terdiri dari fixed effects; $\mathbf{b}_i$ adalah vektor berdimensi q2 yang terdiri dari random effects; ϵ_i adalah vektor komponen error berdimensi n_i; dengan asumsi:

$$\begin{cases} b_i \sim N(0, G) \\ \epsilon_i \sim N(0, R) \\ b_i \text{ dan } \epsilon_i \text{ indepeden} \end{cases}$$

Bentuk persamaan seperti ini adalah sama dengan bentuk model campuran linear.

McCulloch dan Searle [5], mengungkapkan bahwa pengukuran data secara berulang dilakukan untuk tujuan:

- Meningkatkan sensitivitas di antara subjek pengamatan
- Mempelajari perubahan karena waktu
- Efisiensi sampel

Contoh data longitudinal: data kolesterol seseorang yang diamati tiap hari selama 1

minggu (Zhang, et al [6]). Untuk melihat efektifitas terapi hormon decapeptyl, dilakukan percobaan pada tikus yang diamati produksi testosteronnya setiap 10 hari hingga 5 kali pengamatan (Verbeke dan Molenberghs [4].

Model Campuran Linear

Berdasarkan Jiang [3], bentuk umum model campuran linear dapat ditulis:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z\alpha} + \boldsymbol{\epsilon} \quad (3)$$

dimana $\mathbf{y}$ adalah vektor observasi berdimensi n_i, $\mathbf{X}$ adalah matriks covariate yang diketahui, $\boldsymbol{\beta}$ adalah vektor koefisien regresi yang disebut fixed effect, $\mathbf{Z}$ adalah matriks yang diketahui, $\boldsymbol{\alpha}$ adalah vektor random effect, dan $\boldsymbol{\epsilon}$ adalah vektor error.

Asumsi dasar yang harus dipenuhi dalam model campuran linear adalah pertama, bahwa random effect dan error mempunyai mean nol dan varians tertentu, misalkan *var(α)*=**G** dan var($\boldsymbol{\epsilon}$)=**R**. Asumsi kedua yang harus dipenuhi adalah α dan $\boldsymbol{\epsilon}$ tidak berkorelasi.

Jika random effect dan error diasumsikan berdistribusi Normal, maka model (3) disebut model campuran linear Gaussian, asumsinya dapat ditulis:

- $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{G})$
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{R})$

Sehingga distribusi variabel responnya,

$$\mathbf{y} \sim N(\mathbf{X\beta}, \boldsymbol{V})$$

dimana $\boldsymbol{V} = Var(\mathbf{y}) = \boldsymbol{ZGZ}' + \boldsymbol{R}$.

Jika random effect dan error tidak diasumsikan berdistribusi Normal, maka model (3) disebut model campuran linear non-Gaussian.

Estimasi Model Campuran Linear

Berdasarkan Jiang [3], estimasi model campuran linear terdiri dari estimasi fixed effect dan prediksi random effect, yang diberikan sebagai berikut:

Estimasi fixed effect Best Linear Unbiased Estimator (BLUE):

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{BLUE}} = \tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{y} \quad (4)$$

Prediksi random effect yang Best Linear Unbiased (BLUP) adalah:

$$\tilde{\boldsymbol{\alpha}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \quad (5)$$

Pada kenyataannya, komponen varians V tidak diketahui. Oleh karena itu untuk estimasi fixed effect dan random effect, komponen variansnya diganti dengan estimasinya $\hat{\boldsymbol{V}}$ dan hasil estimasinya masing-masing disebut empirical BLUE (EBLUE) dan empirical BLUP (EBLUP).

Estimasi komponen varians **V**

$$\hat{\boldsymbol{V}} = \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}' + \hat{\boldsymbol{R}} \quad (6)$$

dimana $\hat{\boldsymbol{G}}$ dan $\hat{\boldsymbol{R}}$ adalah estimator MLE (Maximum Likelihood Estimation) atau REML (Restricted Maximum Likelihood).

Jiang [3] mengungkapkan bahwa estimator REML untuk komponen varians adalah estimator yang konsisten tanpa mengasumsikan normalitas dari variabel acak dan error. Sehingga, REML dapat diterapkan baik pada model campuran linear Gaussian maupun non-Gassian. Teknis komputasi REML dapat dilihat di Jiang [3].

Model Campuran Linear Distribusi-t dengan Autoregresif (TLMM-AR)

Berdasarkan Lin [1], bentuk model campuran linear TLMM adalah sama dengan model (2), yaitu:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \quad (7)$$

Dengan asumsi:

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{\epsilon}_i \end{bmatrix} \sim t_{n_i+m_2}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_i \end{bmatrix}, v\right) \quad (8)$$

Dimana $\boldsymbol{y}_i$ adalah vektor observasi dari individu ke-i berdimensi n_i *(i=1,...,N)*; $\boldsymbol{X}_i$ dan $\boldsymbol{Z}_i$ adalah matriks kovariat masing-masing berdimensi (n_i x m_1) dan (n_i x m_2); $\boldsymbol{\beta}$ adalah vektor fixed effects berdimensi m_1; $\mathbf{b}_i$ adalah vektor random effects berdimensi m_2; $\boldsymbol{\epsilon}_i$ adalah vektor komponen error berdimensi n_i; $\boldsymbol{\Gamma}$ adalah matriks positif

definit tidak terstruktur berdimensi m_2 x m_2; **Ci** adalah matriks dependensi terstruktur AR(p) berdimensi m_1 x m_1. $\boldsymbol{C_i = C_i(\phi) = [\rho_{|r-s|}(\phi)]},\ r,s = 1,\dots,n_i$; ρ_k adalah fungsi autoregresif dengan parameter $\boldsymbol{\phi} = (\phi_1,\dots,\phi_p)$ dan memenuhi persamaan:

$$\rho_k = \phi_1\rho_{k-1} + \cdots + \phi_p\rho_{k-p},$$

dimana $k = 0,\dots,n_i - 1, \rho_0 = 1$

v adalah derajat bebas dari distribusi t-multivariat.

Untuk mengestimasi parameter baik fixed effect maupun random effect pada model campuran linear distribusi-t (TLMM), Lin [1] mengusulkan untuk menggunakan algorithma Expectation Conditional Maximization Either (ECME), yang merupakan campuran antara algorithma Expectation and Maximization (EM) dengan metode Fisher scoring. Teknis komputasi algorithma ECME dapat dilihat di Lin [1].

Bahan

Data yang digunakan pada studi ini adalah data hasil simulasi dan aplikasi data real.

Simulasi Data

Untuk melihat efisiensi model TLMM dibanding model LMM-REML, kita kaji berdasarkan simulasi data. Untuk data simulasi, kita bangkitkan 50 himpunan data dengan menggunakan software R berdasarkan model campuran linear di bawah ini:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + e_{ij} \qquad (9)$$

Dimana kita tetapkan $j = 1,\dots,5$, $t_{ij} \sim N(0,1)$, $\beta_0 = 3$, $\beta_1 = 2$, e_{ij} mengikuti proses AR(1) yaitu $e_{ij} = \phi e_{ij-1} + \zeta_{ij}$, $\phi = 0{,}9$, ζ_{ij} adalah whitenoise $\zeta_{ij} \sim N(0,1)$, $b_i \sim t(df = n)$, dan kita ambil n=10, dan n=50, sehingga *i=1,...,N*, dimana $N = n \times j$.

Untuk setiap himpunan data yang dibangkitkan dari model (9), kita estimasi sebanyak 3 kali sehingga didapat:

Model-1: model LMM-REML

Model-2: model TLMM dengan e_{ij} whitenoise = AR(0) (TLMM-WN)

Model-3: model TLMM dengan e_{ij}=AR(1), (TLMM-AR1)

Data Real

Data yang akan digunakan adalah data panel untuk 29 provinsi di Indonesia tahun 2007-2009, dengan variabel Y adalah produktivitas padi (kuintal/ha) dan variabel X adalah tenaga kerja per hektar. Provinsi Kepulauan Riau, DKI Jakarta, Sulawesi Barat dan Papua Barat tidak disertakan karena permasalahan data, di mana terdapat beberapa pengamatan yang bernilai 0 (nol) sehingga tidak dapat digunakan fungsi logaritma. Data diperoleh dari Badan Pusat Statistik (BPS).

HASIL DAN DISKUSI

Berdasarkan simulasi data sebanyak 50 data set, diperoleh hasil estimasi sebanyak 50 set yang dirangkum pada Tabel 1. Estimasi model-1, LMM-REML menggunakan software R, package nlme dengan fungsi lme, sedangkan model TLMM menggunakan package lmm dengan fungsi fastml, yang merupakan hybrid ECME dengan Fisher Scoring.

Estimasi Model LMM-REML dengan menggunakaan package nlme, relative lebih mudah diimplementasikan. Namun, package lmm ternyata tidak mudah untuk diimplementasikan, terutama untuk model TLMM dengan proses autoregresif AR(p). Artinya, model LMM-REML lebih praktis dalam penghitungan estimasinya dibanding model TLMM-AR(p).

Berdasarkan Tabel 1, kita lihat bahwa baik model LMM-REML maupun TLMM dapat dikatakan menghasilkan estimasi yang asymptotically unbiased, yaitu biasnya menuju nilai nol untuk n menuju tak hingga ($n \to \infty$) (Shao [7]). Selain itu, estimasi parameter model LMM-REML dan TLMM juga menghasilkan estimator yang konsisten, karena dengan bertambahnya jumlah n sampel, bias dan standard error-nya semakin kecil atau dapat dilihat dari

nilai mean square error (MSE) yang semakin kecil/semakin menuju nilai nol untuk n menuju tak hingga (Mood, Graybill dan Boes [8]).

Tabel 1. Mean, Standard Deviasi (SD), Rata-rata Standard Error (AVE SE) dan Mean Square Error (MSE) estimasi parameter fixed effect dari 50 data set bangkitan

Para-meter	Mean	SD	AVE SE	MSE
(1)	(2)	(3)	(4)	(5)
n=10				
Model_1 REML				
$\beta_0(3)$	3.3444	1.2447	0.6358	0.5524
$\beta_1(2)$	1.9878	0.1731	0.1574	0.0258
Model_2 TLMM-WN				
$\beta_0(3)$	3.3443	1.2447	0.6027	0.5085
$\beta_1(2)$	1.9876	0.1732	0.1554	0.0251
Model_3 TLMM-AR1				
$\beta_0(3)$	3.3536	1.2462	0.5947	0.5040
$\beta_1(2)$	1.9820	0.1167	0.1136	0.0137
$\phi(0,9)$	0.8097			
n=50				
Model_1 REML				
$\beta_0(3)$	2.9593	0.5143	0.3199	0.1054
$\beta_1(2)$	1.9902	0.0045	0.0706	0.0051
Model_2 TLMM-WN				
$\beta_0(3)$	2.9593	0.7172	0.3167	0.1033
$\beta_1(2)$	1.9902	0.0671	0.0704	0.0051
Model_3 TLMM-AR1				
$\beta_0(3)$	2.9611	0.7127	0.3124	0.1005
$\beta_1(2)$	1.9863	0.0490	0.0522	0.0029
$\phi(0,9)$	0.8813			

Efisiensi model

Suatu estimasi model dikatakan lebih efisien dari model yang lain, jika menghasilkan standard error koefisien regresi yang lebih kecil (Rao dan Toutenberg [9])

Jika ada barisan estimator bagi β, misalkan $\{\hat{\beta}_n\}$ yang asymptotically unbiased dikatakan asymptotically efficient jika dan hanya jika $Var_n(\beta) = [\mathfrak{I}_n(\beta)]^{-1}$ (Shao [7])

Dimana $\mathfrak{I}^{-1}$ adalah inverse matriks Informasi Fisher yang tiada lain adalah matriks varians kovarian β (dalam output di R ditulis Cov.Beta untuk package lmm dan akarnya disebut standard error/SE untuk package nlme).

Hasil estimasi dari model REML hampir sama dengan hasil estimasi model TLMM-WN (lihat Tabel 1). Artinya model REML terbukti robust terhadap asumsi normalitas. Sedangkan model TLMM-AR1 terlihat menghasilkan standard error yang paling kecil diantara kedua model lainnya. Untuk memastikan apakah model TLMM-AR1 lebih efisien dibanding kedua model lainnya, kita perhatikan Tabel 2.

Tabel 2. Efisiensi Model REML, Model TLMM-WN dan TLMM-AR(1) berdasarkan 50 set data simulasi

Parameter	SD^2 ($Var_n\beta$)	AVE Cov β	$\lvert Var_n(\beta) - AVE\ Cov\ \beta \rvert$
(1)	(2)	(3)	(4)
n=10			
Model 1 REML			
$\beta_0(3)$	1.5493	0.4042	1.1451
$\beta_1(2)$	0.0300	0.0248	0.0052
Model 2 TLMM-WN			
$\beta_0(3)$	1.5493	0.3899	1.1594
$\beta_1(2)$	0.0300	0.0250	0.0050
Model 3 TLMM-AR1			
$\beta_0(3)$	1.5530	0.3790	1.1741
$\beta_1(2)$	0.0136	0.0134	0.0003
n=50			
Model 1 REML			
$\beta_0(3)$	0.5143	0.1037	0.4106
$\beta_1(2)$	0.0045	0.0050	0.0005
Model 2 TLMM-WN			
$\beta_0(3)$	0.5143	0.1016	0.4127
$\beta_1(2)$	0.0045	0.0050	0.0005
Model 3 TLMM-AR1			
$\beta_0(3)$	0.5080	0.0990	0.4089
$\beta_1(2)$	0.0024	0.0027	0.0003

Berdasarkan Tabel 2, model yang efisien adalah model yang menghasilkan nilai selisih antara varians dari barisan estimasi

β sebanyak 50 data set dengan rata-rata varians error estimasi β yang semakin kecil/mendekati nilai nol (dapat dilihat dari kolom 4). Ternyata model TLMM-AR(1) menghasilkan estimasi yang lebih efisien dibanding model REML dan TLMM-WN terutama estimasi parameter β_1 pada saat n=10, dan ketika n=50 baik estimasi β_0 maupun β_1 menghasilkan estimasi yang lebih efisien dibanding kedua model lainnya.

Aplikasi Data Real

Menurut teori Cobb-Douglas, produksi dipengaruhi oleh modal dan tenaga kerja (Debertin [10]). Dalam hal ini, modal dapat didekati dengan luas lahan. Sementara itu, data yang tersedia di BPS adalah luas panen. Oleh karena itu, luas lahan dapat didekati dengan luas panen dibagi 2 per tahunnya, karena biasanya panen di sebagian besar provinsi di Indonesia terjadi dua kali dalam setahun. Sehingga model Cobb-Douglas-nya dapat ditulis:

$$Y^* = f[LH, X^*] \quad (10)$$

dimana : Y* = produksi padi, LH = luas lahan, X* = tenaga kerja di pertanian(padi)

Sementara itu, data produktivitas (Y) padi dapat dihitung dari data produksi padi dibagi luas lahan (Y*/LH), sehingga model (10) dapat dimodifikasi menjadi:

$$Y = f[1, X] \quad (11)$$

dimana : Y = produktivitas padi (Y*/LH), dan X = tenaga kerja per hektar (X*/LH)

Dalam fungsi double log, sebagaimana yang digunakan Cobb-Douglas, persamaan (11) dapat dituliskan:

$$Y = \beta_0 X^{\beta_1} \quad (12)$$

Kemudian fungsi (12) dilinierkan menjadi :

$$\ln Y = \ln \beta_0 + \beta_1 \ln X \quad (13)$$

Data produksi padi dikumpulkan setiap subround (caturwulan/empat bulanan) melalui Survei Ubinan. Untuk mengakomodir adanya pengaruh acak (random effect) yang diakibatkan oleh pengambilan sampel ubinan, maka, salah satu model yang mungkin untuk data longitudinal produktivitas padi, dapat ditulis sebagai berikut:

$$\ln Y_{ij} = \ln \beta_0 + \beta_1 \ln X_{ij} + \alpha_i + \epsilon_{ij} \quad (14)$$

Y_{ij} = produktivitas padi provinsi ke-i, pada waktu ke-j

X_{ij} = tenaga kerja pada provinsi ke-i dan waktu ke-j

j = $1,\ldots,5$ (dimana $j = 1$ untuk tahun 2007, $j = 2$ untuk tahun 2008, $j = 3$ untuk tahun 2009).

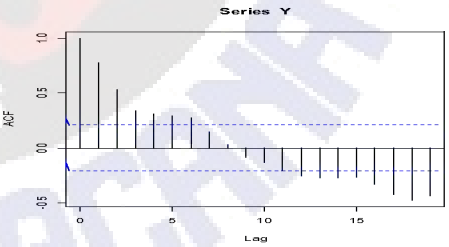
β_0, β_1, adalah fixed effect

α_i adalah random effect

ϵ_{ij} adalah error

α dan ϵ diasumsikan tidak berkorelasi.

Pengumpulan data produksi padi dilaksanakan pada waktu petani panen pada unit sampel yang sama pada setiap subround-nya. Sehingga data produksi padi atau produktivitas padi kemungkinan besar terdapat korelasi serial dalam subjek. Dapat kita lihat dari Gambar 1 plot ACF data respon Y yang berpola sinusoidal.



Gambar 1. Plot ACF untuk data produktivitas padi (Y) tahun 2007-2009

Berdasarkan Gambar 2, PACF data respon Y terlihat menonjol pada lag 1. Hal ini menunjukkan adanya proses Autoregresif AR(1).

Untuk melihat model mana yang lebih baik dalam aplikasi data real antara model

REML dan TLMM, Model (14) kita estimasi sebanyak dua kali:

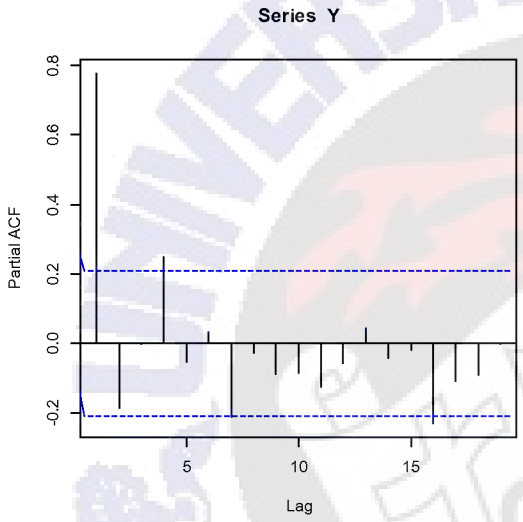
Model-1a: model LMM-REML

Model-2a: model TLMM-AR(1)

Kita gunakan Akaike Information Criteria (AIC) untuk memilih model terbaik. Model yang memberikan nilai AIC minimum adalah model terbaik(Lin,2008).

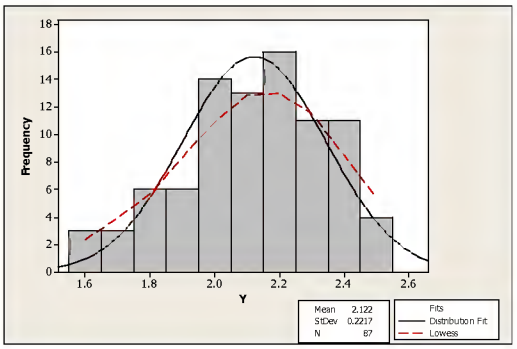
$$AIC = -2[log\ likelihood - m] \qquad (15)$$

Dimana m adalah jumlah parameter



Gambar 2. Plot Partial ACF untuk data Produktivitas padi (Y) tahun 2007-2009

Berdasarkan Gambar 3, kita lihat pola sebaran data respon Y, produktivitas padi menyerupai sebaran t-student dengan ekor yang gendut. Karena model campuran linear LMM-REML adalah model yang robust terhadap asumsi normalitas dan model TLMM-AR(1) adalah model yang mengasumsikan bahwa Y berdistribusi t, maka kedua model tersebut dapat digunakan untuk mengestimasi data longitudinal produktivitas padi. Hasil estimasi kedua model tersebut dapat kita lihat pada Tabel 3.



Gambar 3. Pola distribusi data respon Y

Berdasarkan Tabel 3, kita lihat bahwa model TLMM-AR(1) menghasilkan nilai AIC yang lebih kecil dibanding model REML dan mengahasilkan standard error koefisien regresi yang juga lebih kecil. Estimasi korelasi serial untuk data produktivitas padi dari tahun 2007-2009 sebesar 0,77 menunjukkan bahwa terdapat korelasi serial yang cukup tinggi. Dengan demikian model TLMM-AR(1) dapat dikatakan lebih efisien dibanding model REML untuk kasus korelasi serial yang cukup tinggi.

Tabel 3. Hasil Estimasi Data Produktivitas Padi Berdasarkan Model REML dan Model TLMM-AR(1)

	REML		TLMM-AR(1)	
	estimasi	SE	estimasi	SE
(1)	(2)	(3)	(4)	(5)
β_0	2.1096	0.0782	2.2095	0.0713
β_1	0.0060	0.0308	-0.0424	0.0288
ϕ			0.7721	0.0659
AIC	-188.316		-307.598	

KESIMPULAN

Model campuran linear dengan asumsi distribusi t dan mengandung proses autoregresif AR(p) (TLMM-AR(p)) lebih efisien diterapkan pada model longitudinal yang mengandung korelasi serial yang cukup tinggi dalam subjek antar waktu dibanding model campuran linear dengan metode REML.

Akan tetapi, penghitungannya secara komputasional dengan menggunakan software R, model campuran linear metode REML lebih praktis dibanding model TLMM-AR(p).

Untuk penelitian selanjutnya, dapat dicek apakah model TLMM-AR(p) masih lebih efisien atau tidak dibanding model REML untuk kasus dimana korelasi serial yang kecil.

UCAPAN TERIMAKASIH

Kami ucapkan terimakasih yang sebesar-besarnya kepada Prof. Dr. Khairil Anwar Notodiputro dan Dr, Kusman Sadik, M.Sc sebagai dosen yang senantiasa mengarahkan kami, Septiandiantoro yang telah membantu dalam pembuatan simulasi data dengan software R dan keluarga kami tercinta yang telah memberikan dukungan yang penuh.

DAFTAR PUSTAKA

[1] Lin, T., "Longitudinal Data Analysis Using t-Linear Mixed Models with Autoregressive Dependence Structures," Journal of Data Science **6**, 333-355, 2008.

[2] Pinheiro, J.C., Liu, C. dan Wu, Y.N., "Efficient Algorithms for Robust Estimation in Linear Mixed-effects Models using The Multivariate t distribution," Journal of Computational and Graphical Statistics **10**, 249-276, 2001.

[3] Jiang, J., Linear and Generalized Linear Mixed Models and Their Applications, New York: Springer, 2007.

[4] Verbeke, G. dan Molenberghs, G., Linear Mixed Models for Longitudinal Data, New York: Springer, 2000.

[5] McCulloch, C.E dan Searle, S.R., Generalized, Linear, and Mixed Models, Canada: John Willey&Son, 2001.

[6] Zhang, D. dan Davidian, M., "Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data," Biometrics **57,** 795-802, 2001.

[7] Shao, J., Mathematical Statistics Second Edition, New York: Springer, 2003.

[8] Mood, A.M, Graybill, A.F dan Boes, C.D., Introduction to The Theory of Statistics – third edition, Singapura: McGraw-Hill, 1974

[9] Rao, C.R. dan Toutenberg, H., Linear Models: Least Square and Alternatives Second Edition, New York: Springer, 1999.

[10] Debertin, D.L, Agricultural Production Economics. New York: Macmilan Publishing Company, 1986.

DISKUSI

Pertanyaan : Simulasi untuk X mengapa dipilih normal baku? Bagaimana jika dipilih selain normal baku?

Jawab : karena menggunakan model Gaussian sehingga dipilih normal baku, tetapi bisa digunakan selain normal baku