Prosiding Seminar Nasional MIPA 2016
"Peran Penelitian Ilmu Dasar dalam Menunjang Pembangunan Berkelanjutan"
Jatinangor, 27-28 Oktober 2016
ISBN 978-602-72216-1-1

Analisis Data Ketersediaan Layanan Air Ledeng di Daerah Perkotaan Indonesia (Studi Kasus Pada Indonesia Family Life Survey)

Bornok Fagabe*, Septiadi Padmadisastra, Yudhie Andriyana

Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Padjadjaran
*E-mail: bfagabe@gmail.com

Abstrak

Data longitudinal adalah data yang diperoleh dari hasil pengukuran berulang (*repeated measures*) terhadap beberapa individu (*cross-sectional*). Di dalam pemodelan data longitudinal untuk respon biner seperti data ketersediaan air ledeng dapat menggunakan model regresi logistik. Penaksiran parameter dari model regresi logistic sangat sulit dilakukan karena asumsi independensi antar pengamatan tidak terpenuhi, oleh sebab itu untuk menanggulangi autokorelasi antar pengamatan tersebut dapat menggunakan *Generalized Estimating Equations* dengan beberapa struktur korelasi. Dari hasil simulasi, diperoleh bahwa pendekatan GEE dengan bentuk korelasi *AR(1)* menghasilkan penaksir parameter yang efisien dibandingkan dengan bentuk korelasi lainnya.

Kata Kunci :RegresiLogistik, *Generalized Estimating Equations*, Data Longitudinal

1. PENDAHULUAN

Pada saat ini untuk mendapatkan air yang layak dikonsumsi harus mengeluarkan biaya. Awalnya masyarakat Indonesia menggunakan air dari sumber mata air tanah sebagai air yang layak untuk dikonsumsi, tetapi menurut Badan Pengelolaan Lingkungan Hidup Daerah (BLPHD) Jakarta menunjukkan bahwa 41% sumur gali yang digunakan oleh rumah tangga berjarak kurang dari 10 meter dari *septic tank*, selain itu sumber air tanah jumlahnya juga sangat terbatas, berbeda halnya dengan sumber air permukaan. Hal ini menjadi alasan PDAM untuk mengolah air permukaan menjadi air ledeng. Bila dilihat dari segi kualitasnya air ledeng merupakan air yang layak dikonsumsi. Menurut Hutton dan Haller untuk meningkatkan kesehatan pada negara-negara berkembang sebaiknya menggunakan air ledeng. Hal tersebut juga didukung oleh laporan Riskesdas 2007 yang menyatakan bahwa terjadi kasus diare pada anak-anak akibat konsumsi air dari sumur terbuka sebesar 34%. Angka ini lebih tinggi dibandingkan dengan anak-anak yang menggunakan air ledeng (UNICEF, 2012). Fakta ini menunjukkan bahwa air ledeng yang diolah PDAM bias menjadi air berkualitasbaik.

Ketersediaan layanan air ledeng (*Piped Water Connection*) di rumah dipengaruhi oleh Median pengeluaran rumah tangga perbulan (*Median Household Monthly Expenditure*), luas rumah *(Floor Area)*, kondisi rumah (*House Condition*). Untuk melihat ketersediaan data mengenai variabel-variabel tersebut dapat menggunakan data pada *Indonesia Family Life Survey* (IFLS) terkhususnya pada daerah perkotaan. Pada survei ini yang menjadi unit sampling adalah rumah tangga, dari masing unit sampling diamati sebanyak 4 kali,yaitu pada tahun 1993, 1997, 2000, dan 2007. Survei ini dilakukan pada 13 provinsi (provinsi di Jawa, Bali, NTB, Sulawesi Selatan, Kalimantan Selatan, Sumatera Selatan, Lampung, Sumatra Barat, dan Sumatra Utara). Untuk variabel ketersediaan layanan air ledeng merupakan variabel yang bersifat biner.

Dengan banyaknya metode pemodelan yang tersedia maka diperlukan pemahaman bagaimana membuat model untuk respon data biner dengan pengukuran berulang seperti data *Piped Water Connection* .

2. METODOLOGI PENELITIAN

2.1 Data Longitudinal

Data longitudinal adalah data yang diperoleh dari hasil pengukuran berulang (*repeated measures*) terhadap beberapa individu (*cross-sectional*).

Sekumpulan hasil pengukuran berulang dari setiap subjek dapat dibuat menjadi sebauh vektor$\mathbf{Y_i} = (y_{ij}, \dots, y_{in_i})^{\mathrm{T}}$ dengan rata-rata $\mathbf{E(Y_i)} = \boldsymbol{\mu_i}$, matriks varians kovarians $\mathbf{Var(Y_i)} = \mathbf{V_i}$ berukurann_ixn_i, $\mathbf{X_i}$ adalah martriks variabel penjelas yang berukuran n_ix p, atau dapat ditulis sebagai $\mathbf{X_i} = (\mathbf{x_{i1}}, .., \mathbf{x_{in_i}})^{\mathbf{T}}$. Dengan$\mathbf{x_{ij}}$ merupakan sebuah vektor variabel penjelas yang berukuran p, dengan p adalah banyak variabel penjelas untuk pengamatan $j = 1, .., n_i$ pada subjek $i = 1, .., K$, sedangkan $n_i$adalah banyaknya pengukuran berulang yang dilakukan. Nilai untuk masing rata-rata dan varians dari y_{ij} adalah $\mathrm{E}(y_{ij}) = \mu_{ij}$ dan $\mathrm{Var}(\mathrm{Y_{ij}}) = \mathrm{v_{ij}}$.

2.2 Generalized Linear Models

Awal perkembangan pemodelan dalam statistik diperkenalkan oleh Gauss, Boole, Cayleydan Sylvester yaitu *General Linear Models*, dimana untuk data longitudinal model tersebut dapat dituliskan sebagai berikut :

$$Y_{ij} = \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \epsilon_{ij} \qquad 2.1$$

Pada Persamaan 2.1 terdapat asumsi yang harus dipenuhi yaitu ϵ_{ij} merupakan variabel acak yang berdistribusi normal dan saling independen $N(0, \sigma^2)$. Pada kenyataannya sangat sulit untuk memenuhi asumsi normalitas pada residual maka untuk menanggulangi hal tersebut *Liang dan Zeger (1986)* melakukan generalisasi terhadap asumsi normalitas. Pada model ini diasumsikan Y_{ij} merupakan anggota keluarga eksponensial yang memiliki distribusi peluang sebagai berikut:

$$f(y_{ij}) = \exp\left[\{y_{ij}\theta_{ij}\text{-a}(\theta_{ij}) + b(y_{ij})\}\phi\right]$$

Dengan mengkaitkan $\mathrm{E}(y_{ij}) = \mu_{ij}$ terhadap predict or linier $\eta_{it} = \mathbf{x_{ij}\beta}$ oleh fungsi $\mu_{ij} = h(\eta_{ij}) = h(\mathbf{x_{ij}\beta}\ )$,
sehingga invers dari fungsi tersebut adalah $\eta_{ij} = g(\mu_{ij})$, dengan $\boldsymbol{\beta}$ adalah vektor parameter berukuran p+1 yang tidak diketahui, $g(.)$ adalah *link function* atau fungsi invers dari h(.).

2.3 Regresi Logistik

Suatu variable acak yang bersifat biner merupakan suatu variabel yang berdistribusi Bernouli. Jika data berdistribusi Bernouli maka model regresi yang cocok untuk pendekatan *Generalized Linear Models* adalah regresi logistic dengan *link function logit*. Model regresi logistic untuk variable penjelas sebanya kp, dengan parameter π_{ij} tidak diketahui dapat dituliskan sebagai berikut:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \log\left(\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right)$$
$$= \beta_0 + \mathbf{x_{ij}\beta}$$

π_{ij}adalah peluang suksesnya terjadi suatu kejadian. Parameter β_0 dan $\boldsymbol{\beta}$ akan ditaksir dengan *Generalized Estimating Equations*.

2.4 Generalized Estimating Equations

GEE adalah metode quasi likelihood dengan asumsi bahwa data yang digunakan berasal dari keluarga eksponensial. Dimana untuk menaksir parameter β_0 dan $\boldsymbol{\beta}$ dapat menggunakan *general estimating equations* berikutini:

$$\textstyle\sum_{i=1}^{K} \mathbf{D_i'V_i^{-1}S_i} = \mathbf{0},$$

keterangan:

$$\mathbf{D_i'} = \frac{\mathbf{d\{\mu_i\}}}{\mathbf{d\beta}} = \mathbf{A_i\Delta_i S_i}.$$

$$\mathbf{V_i} = \mathbf{A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}}/\phi.$$

$\mathbf{V_i}$yang akan sama dengan $\mathbf{cov(Y_i)}$ jika $\mathbf{R(\alpha)}$ yang merupakan matriks korelasi dari subjek $\mathbf{Y_i}$. $\mathbf{R(\alpha)}$ merupakan sebuah matrik simetris yang berukuran $n_i x n_i$ yang disebut dengan *working correlation matrix*. *Working Correlation Matrix* terdiri dari beberapa bentuk yaitu *Independent structure, Exchangeable structure, Autoregressive Structure, Unstructure Stucture, Fixed Correlation* (Swan,2006).

Persamaan GEE dapat diselesaikan dengan persamaan *Fisher's Scoring Method*, adapun persamaan iterasinya adalah

$$\mathbf{\widehat{\beta}_{j+1}} = \mathbf{\widehat{\beta}_j} + (\mathbf{E(H)}^{(j)})^{-1}\mathbf{q}^{(j)},$$

dimana:

$$\mathbf{q} = \textstyle\sum_{i=1}^{K} \mathbf{D_i^T}(\widehat{\boldsymbol{\beta}_j})\, \widetilde{\mathbf{V}}_i^{-1}(\widehat{\boldsymbol{\beta}_j})\mathbf{S_i}(\widehat{\boldsymbol{\beta}_j}),$$

$$\mathbf{E(H)} = \textstyle\sum_{i=1}^{K} \mathbf{D_i^T}(\widehat{\boldsymbol{\beta}_j})\, \widetilde{\mathbf{V}}_i^{-1}(\widehat{\boldsymbol{\beta}_j})\mathbf{D_j}(\widehat{\boldsymbol{\beta}_j}),$$

$$\widetilde{\boldsymbol{V_i}}^{-1}(\boldsymbol{\beta}) = \boldsymbol{V_i}[\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}].$$

Untuk ukuran sampel yang besar maka $K^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}_G} - \boldsymbol{\beta})$mengikuti distribusi *multivariate Gaussian* dengan rata-rata adalah nol dan martrik kovariansnya $\mathbf{V_G}$ adalah:

$$\left(\sum_{i=1}^{K} \mathbf{D_i^T V_i^T D_i}\right)^{-1} \left\{\sum_{i=1}^{K} \mathbf{D_i^T V_i^{-1} cov(Y_i) V_i^{-1} D_i}\right\} \left(\sum_{i=1}^{K} \mathbf{D_i^T V_i^T D_i}\right)^{-1}$$

2.5 EfisiensiPenaksirParameter

Liang dan Zeger (1986) menyatakan bahwa efisiensi estimator dari $\boldsymbol{\beta}$ bergantung pada nilai α yang terdapat pada matriks $\mathbf{R(\alpha)}$.

3. Hasil dan Pembahasan

Pada Penelitian ini digunakan sampel sebayak 1284 rumah tangga. Untuk mempermudah proses perhitungan pada penelitian ini menggunakan sofware R 3.2.3 dengan packages *geepack*.

3.1 Model Regresi Logistik

Model regresi logistik yang dapat menjelaskan keterkaitan variable *Median Household Monthly* (X_1), *Floor Area* (X_2) dan *House Condition* (X_3) terhadap ketersediaan layanan air ledeng (*Piped Water Connection)* di rumah adalah:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta_1} + \boldsymbol{\beta_1}\mathbf{X}_{ij1} + \beta_2\,\mathbf{X}_{ij2} + \beta_3\mathbf{X}_{ij3}$$

3.2 Penentuan *Working Correlation Matrix*

Adapun matriks korelasi yang digunakan dalam penelitian ini adalah: *Independent Structure, Exchangeable Structure, Autoregressive Structure, Unstructure Structure,* adapun bentuk masing-masing matriks korelasi teresebut adalah :

a. Independent Stucture

$$\mathbf{R_i}(\hat{\boldsymbol{\alpha}}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

b. Exchangeable Structure

$$\mathbf{R_i}(\hat{\boldsymbol{\alpha}}) = \begin{bmatrix} 1 & 0.492 & 0.492 & 0.492 \\ 0.492 & 1 & 0.492 & 0.492 \\ 0.492 & 0.492 & 1 & 0.492 \\ 0.492 & 0.492 & 0.492 & 1 \end{bmatrix}$$

c. Autoregressive Structure

$$\mathbf{R_i}(\hat{\boldsymbol{\alpha}}) = \begin{bmatrix} 1 & 0.62 & 0.384 & 0.238 \\ 0.62 & 1 & 0.62 & 0.384 \\ 0.384 & 0.62 & 1 & 0.62 \\ 0.238 & 0.384 & 0.62 & 1 \end{bmatrix}$$

d. Unstucture Stucture

$$\mathbf{R_i}(\hat{\boldsymbol{\alpha}}) = \begin{bmatrix} 1 & 0.478 & 0.420 & 0.273 \\ 0.478 & 1 & 0.630 & 0.445 \\ 0.420 & 0.630 & 1 & 0.663 \\ 0.273 & 0.445 & 0.663 & 1 \end{bmatrix}$$

3.3 Penaksiaran Parameter

Parameter-parameter yang terdapat dalam persamaan model regresi logistic akan ditaksir *Generalized Estimating Equations* (GEE) dengan menggunakan beberapa struktur korelasi yang telah disebut pada bagian sebelumnya, sedangkan untuk iterasinya menggunakan *Fisher Scoring Method.* Berikut ini merupakan hasil taksiran parameter dengan GEE mengggunakan beberapa struktur korelasi yang mungkin :

1. Model dengan *Independent Structure*

(Model I)

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = -1.4532 + 0.4686\ \ln x_{ij1} + 0.1500\ln x_{ij2} + 0.2292 x_{ij3}$$

2. Model dengan *Exchangeable Structure* (Model II)

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = -2.2870 + 0.2269\ \ln x_{ij1} + 0.1851\ \ln x_{ij2} + 0.3733 x_{ij3}$$

3. Model denganAR-1 Structure (Model III)

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = -1.9704 + 0.2333\ \ln x_{ij1} + 0.0762\ \ln x_{ij2} + 0.2879 x_{ij3}$$

4.Model dengan Unstructure Stucture (Model IV)

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = -2.1406 + 0.2247\ \ln x_{ij1} + 0.0615\ \ln x_{ij2} + 0.3186\, x_{ij3}$$

3.4 Varians Penaksir Parameter

Akar dari varians $\hat{\boldsymbol{\beta}}$, merupakan *standard error* dari $\hat{\boldsymbol{\beta}}$. *Standard error* digunakan untuk melihat efisiensi dari penaksiran parameter tersebut.

Tabel 1.KoefisienRegresidan*Standar Error* Koefisien regresi

Variabel Independen	*Working Correlation Matrix*							
	AR(1)		Independent		Exchangeable		Unstructure	
	Koefisien	Robust (SE)	Koefisien	Robust (SE)	Koefisien	Robust (SE)	Koefisien	Robust (SE)
Intercept	-1.9704	0.1558	-1.4532	0.2305	-2.2870	0.1771	-2.1406	0.1655
x1	0.2333	0.0365	0.4686	0.0506	0.2269	0.0401	0.2247	0.0385
x2	0.0762	0.0582	0.1500	0.0832	0.1851	0.0631	0.0615	0.0617
x3	0.2879	0.0314	0.2292	0.0456	0.3733	0.0355	0.3186	0.0333

BerdasarkanTabel 1 dapat lihat bahwa *standar error* $\hat{\boldsymbol{\beta}}$ yang paling kecil adalah *standard error* dari matriks korelasi *Exchangeable Structure,* dan nilai-nilai

standar error $\widehat{\boldsymbol{\beta}}$ yang paling besar adalah dari matriks korelasi *Independent Structure.*

Oleh sebab itu didapatkan model yang dapat menjelaskan ketersediaan layanan air ledeng dan keterkaitannya dengan *Median Household Monthly Expenditure, House Condition, dan Floor Area* adalah Model III.

4. Kesimpulandan Saran

Dari hasil dan pembahasan sebelumnya diperoleh kesimpulan sebagaiberikut ini.

a. Penentuan bentuk hubungan antar pengamatan dalam subjek $\mathbf{R_i}(\widehat{\boldsymbol{\alpha}})$ perlu ditentukan, karena penggunaan bentuk korelasi yang berbeda-beda mengakibatkan *standard error* penaksir parameter yang berbeda-beda
b. Penentuan struktur korelasi yang benar menentukan efisiensi dari taksiran parameter.
c. Dari model regresi logistik yang terpilih maka dapat dinterpretasikan bahwa :
   i. Setiap bertambah satusatuan *Median Household Monthly* akan menyebabkan *exp (0.02333)*=1,26276 kali tersedianya layanan air ledeng dirumah.
   ii. Setiap bertambah satu satuan *Floor Area* menyebabkan exp(0.0762) =1,079178kali tersedianya layanan air ledeng dirumah.
   iii. Apabila*House Condition* semakin baik maka akan menyebabkan exp(0.2879) = 1,333624 kali tersedianya layanan air ledeng dirumah.

DAFTAR PUSTAKA

Beacham, Lauren. 2012. *Using Generalized Estimating Equations To Analyze Repetead Measure Binary Data From Young Adolescent Crowd Study. Thesis.*Louisiana State University and Agricultural and Mechanical College: United States.

Chadidjah, Anna dan Indra Elfiyan .2009. *Model Regresi Data Panel untuk Menaksir Realisasi Total Investasi Asing dan Dalam Negeri*. Fakultas MIPA UNY:Yogyakarta.

Diggle, Peter J.1994.*Analysis of Longitudinal Data*.Oxford University Press.New York.

Myers, Raymond H., dkk.1937. *Generalized Linear Models: with applications in engieering and the sciences. John Wiley and Sons*, NewYork.

Liang, K-Y &Zeger, S. 1986. *Longitudinal data analysis using Generalized Linear Model*.Biometrika.

Nugraha, Jaka, dkk. 2006.*Model Regresi Logistik untuk Respons Biner Multivariate dengan Generalized Estimating Equation*. Paper Universitas Isalam. Yogyakarta.

Swan,Taryn.2006. *Generalized Estimating Equation when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling*. Disertasi. University of Southern Queensland: Toowoomba.