ISSN: 2528-4630

PEMODELAN REGRESI MULTILEVEL ORDINAL PADA DATA PENDIDIKAN DI JAWA BARAT

Bertho Tantular
Departemen Statistika FMIPA Universitas Padjadjaran
bertho@unpad.ac.id

ABSTRAK. Dalam *generalized linear models*, respon ordinal dianalisis menggunakan *proportional oddsmodels*. Penggunaan *proportional oddsmodels* akan menghasilkan sebanyak kategori di kurang satu (*C*-1) model logit yang mana peluang yang digunakan adalah peluang kumulatif. Penaksiran parameter untuk model *proportional odds* menggunakan metode *maximum likelihood* dengan menggunakan algoritma *Fisher's scoring* pada proses iterasinya. Data mengenai tingkat pendidikan seseorang dipengaruhi oleh faktor individu dan faktor lingkungan, sehingga datanya merupakan data hierarki. Data hierarki dengan respon ordinal tidak dapat dimodelkan dengan model regresi ordinal biasa karena efek dari lingkungan tidak dapat diperoleh. Selain itu juga akan menghasilkan penaksir yang tidak efisien. Pemodelan respon ordinal pada data hierarki disebut dengan *multilevel proportional odds models*. Dalam model ini, baik level individu maupun lingkungan diakomodasi dalam pembentukan modelnya. Model *multilevel proporstional odds* tidak dapat ditaksir menggunakan metode *maximum likelihood* biasa tetapi harus menggunakan pendekatan *penalized quasi likelihood*. Penaksir ini dapat menghasilkan taksiran bagi level individu maupun level lingkungan.

Kata Kunci: *multilevel proportional odd models; penalized quasi likelihood; data pendidikan*

1. PENDAHULUAN

Penelitian mengenai tingkat pendidikan seseorang seringkali terkonsentrasi pada masalah bagaimana menelusuri hubungan antara individu dengan lingkungannya misalnya: sekolah, kecamatan atau kabupaten. Konsep umum suatu individu berkorelasi dengan lingkungannya adalah bahwa suatu individu dipengaruhi lingkungan sosial (sekolah atau kecamatan) tempat mereka berada, dan sifat-sifat dari lingkungan sosial tersebut terbentuk oleh individu-individu yang membuat lingkungan tersebut. Secara umum individu dan lingkungan sosial merupakan suatu sistem hierarki atau dapat dikatakan sebagai suatu struktur tersarang (*nested*). Penelitian data yang berstruktur hierarki peubah-peubah dapat didefinisikan pada tingkat individu dan pada tingkat lingkungan. (Hox [8])

Beberapa peneliti telah membuat beberapa pendekatan untuk menganalisis data berstruktur hierarki. Ringdal [13] menyebutkan bahwa pada awalnya analisis digunakan tanpa memperhatikan informasi mengenai keanggotaan individu dalam lingkungan meskipun data yang diperoleh berisi informasi tersebut. Hal ini mengakibatkan ketidakpuasan pada hasil analisisnya karena tidak bisa didapatkan simpulan yang lebih khusus untuk masing-masing lingkungan. Selain itu secara teori, mengabaikan informasi ini dapat menimbulkan masalah dalam inferensinya.

Jones dan Steenbergen [9] menyebutkan bahwa masalah yang muncul akibat mengabaikan informasi lingkungan adalah munculnya heteroskedastisitas dalam galat. Pendekatan lain untuk menganalisis data berstruktur hierarki adalah dengan cara

membuat model-model yang terpisah untuk setiap taraf pada tingkat lingkungan. Pendekatan ini menimbulkan masalah yaitu banyak informasi mengenai lingkungan menjadi tidak tercakup. Masalah lain yang muncul dari pendekatan ini adalah bahwa interaksi antar faktor dari tingkat yang berbeda tidak bisa didapatkan.

Pendekatan lain yang lebih baik adalah dengan menggunakan model regresi peubah boneka. Model regresi dengan peubah boneka dapat digunakan untuk mengatasi masalah heterogenitas. Akan tetapi model regresi dengan peubah boneka tetap tidak dapat mengatasi apabila ada hubungan antara peubah pada tingkat yang berbeda.

Model multilevel mulai diperkenalkan oleh Goldstein [5]. Model ini disebutkan dapat mengatasi semua masalah yang muncul dari data dengan struktur hierarki. Dalam model multilevel, struktur hierarki didefinisikan sebagai level. Tingkat yang paling rendah yaitu individu disebut level 1 dan tingkat yang lebih tinggi yaitu lingkungan disebut level 2. Model multilevel selain dapat menentukan keragaman antar lingkungan juga dapat menunjukkan korelasi antar dua individu yang pada model lain diasumsikan tidak ada. Selain itu model multilevel juga dapat mengukur interaksi yang mungkin terjadi antara peubah pada tingkat yang berbeda.

Dalam model regresi apabila responnya bersifat kategori maka biasanya digunakan model regresi logistik yang dalam pemodelannya harus menggunakan fungsi penghubung (*link function*). Apabila responnya berdistribusi binomial dengan parameter proporsi (π_{ij}) maka fungsi penghubung yang digunakan adalah *logit* ($\log\{\pi/(1-\pi)\}$) sehingga modelnya disebut dengan model logistik. Penaksiran parameter dari model logistik menggunakan metode maksimum likelihood, tidak dapat diperoleh penaksir yang eksplisit sehingga harus melalui proses iterasi. Metode iterasi yang digunakan biasanya metode Newton-Rhapson atau metode Scoring.

Pada data mengenai tingkat pendidikan, respon yang digunakan memiliki skala ukur ordinal sehingga untuk memodelkan data tersebut harus menggunakan model logistik ordinal yang dikenal dengan *cumulative logit model* atau *proportional odds models*. (Agresti [1]). Oleh karena data mengenai tingkat pendidikan merupakan data berstruktur hierarki maka model yang digunakan adalah *cumulative logit mixed model* atau *multilevel proportional odds models*.

2. METODE PENELITIAN

Secara umum rumusan matematis untuk model logistik multilevel dengan fungsi penghubung *logit* dapat dituliskan dalam bentuk matriks sebagai berikut

$$\mathbf{Logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad \ldots(1)$$

Dengan **X** adalah matriks variabel bebas *fixed* dari level pertama, $\boldsymbol{\beta}$ adalah vektor koefisien regresi dan **u** adalah vektor galat untuk level 2.

Persamaan 1 dapat ditulis juga sebagai

$$\text{Logit}(\pi_{ij}) = \mathbf{x}_j\boldsymbol{\beta} + u_j \qquad \ldots(2)$$
$$y_{ij} \sim \text{Bernouli}(\pi_{ij})$$

Dengan x_j adalah vektor variabel bebas, $\boldsymbol{\beta}$ adalah vektor koefisien regresi dan u_j adalah vektor galat untuk level 2. dalam hal ini u_j berdistribusi normal dengan rata-rata nol dan varians σ^2_2.(Sneijder adn Bosker [14])

Secara umum dari Model (1) dapat ditentukan bahwa galat total untuk model tersebut adalah $\xi_{ij} = \varepsilon_{ij} + u_j$. Sehingga secara umum varians untuk ξ_{ij} dengan fungsi penghubung *logit* adalah:

$$\text{var}(\xi_{\text{ij}}) = \sigma_2^2 + \frac{\pi^2}{3} \qquad \ldots(3)$$

Dan korelasi *intaclass* untuk galat total adalah

$$\rho \equiv \mathrm{Cor}(\xi_{ij},\xi_{i'j}) = \frac{\sigma_2^2}{\sigma_2^2+\pi^2/3} \quad \ldots(4)$$

Hesketh [7]

Secara umum rumusan matematis dalam bentuk matriks untuk model *random-coefficient* dua level dengan fungsi penghubung *logit* dapat dituliskan dalam bentuk matriks sebagai berikut

$$\mathbf{Logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} \quad \ldots(5)$$

Dengan **X** adalah matriks variabel bebas *fixed* dari level pertama, $\boldsymbol{\beta}$ adalah vektor koefisien regresi dan **u** adalah vektor galat untuk level 2 dan **Z** adalah matriks *random-effect* pada level 2.

Untuk menaksir parameter-parameter yang terlibat dalam model multilevel untuk respon biner bisa digunakan *Marginal Quasi-Likelihood* (MQL) yang diusulkan oleh Goldstein [5]. Menurut Goldstein [5] penaksiran koefisien dengan menggunakan MQL akan menyebabkan *underestimate* terutama untuk sampel kecil. Begitu pula menurut Rodriguez dan Goldman [13] penaksiran yang diturunkan menggunakan MQL untuk respon biner akan menyebabkan bias pada saat kuantitas klasternya cukup besar. Selain menggunakan MQL parameter-parameter tersebut juga bisa ditaksir dengan menggunakan *Penalized Quasi-Likelihood* (PQL) yang diusulkan oleh Hedeker [6].

Model multilevel non linier atau lebih khusus lagi model dengan respon biner adalah model regresi yang tidak bisa dijelaskan sebagai kombinasi linier dari koefisiennya sehingga dengan demikian vektor **y** tidak lagi berdistribusi normal. Secara umum fungsi densitas peluang dari **y** dapat dituliskan sebagai berikut:

$$f(y) = \int\int f(y,\beta_1,\beta_2,\ldots,\beta_k)\, d\beta_1 d\beta_2 \ldots d\beta_k \quad \ldots(6)$$

untuk β_i berdistribusi kontinyu. Secara umum persamaan 6 tidak bisa diselesaikan kecuali melalui prosedur iterasi. Untuk menyelesaikan persamaan fungsi kemungkinan

$$L = \prod_{i=1}^{n} f(y_i)$$

harus menggunakan teknik integrasi numerik. Dalam model multilevel prosedur yang digunakan adalah prosedur *Gaussian Quadrature* untuk menghitung integral secara numerik. Metode Kemungkinan Maksimum memerlukan nilai awal yang baik untuk parameter-parameternya.

Parameter untuk model non linier ditaksir dengan menggunakan Kuadrat Terkecil Biasa yang kemudian dijadikan sebagai nilai awal. Selanjutnya prosedur mencocokkan model yang digunakan adalah *Maximum Aposterior* untuk parameter yang tidak diketahui sehingga dapat dibentuk fungsi densitas peluang untuk parameter sebagai berikut

$$f(\beta|y) = \frac{f(\beta)f(y|\beta)}{f(y)} \quad \ldots(7)$$

sehingga fungsi kemungkinannya adalah

$$\ln f(\beta|y) = \ln f(\beta) + \ln f(y|\beta) - \ln f(y) \quad \ldots(8)$$

Dengan nilai awal tadi (sebut saja β_{0i} untuk i = 1, 2, …, k) fungsi kemungkinan persamaan (6) dapat diturunkan dan disamakan dengan nol

$$\frac{\partial}{\partial \hat{\beta}_{0i}} \ln f\left(\hat{\beta}_{0i}|y\right) = 0 \text{untuk } i = 0, 1,\ldots, k \quad \ldots(9)$$

sehingga bisa diperoleh penaksir β_i. Penaksir β_i ini digunakan kembali untuk memperoleh penaksir β_i yang baru. Proses diulang sehingga didapatkan hasil yang konvergen.

Rodriguez dan Goldman [13] dalam tulisannya menyatakan bahwa pendekatan prosedur penaksiran termotivasi dengan mempertimbangkan model multilevel logit yang dianggap linier. Persamaan 2 dapat diubah menjadi

$$Y = \pi + \varepsilon, \text{ dengan } \pi = f(\mathrm{X}\beta + u) \quad \ldots(10)$$

dengan **Y** adalah vektor respon, *f* adalah transformasi invers *logit* (atau anti *logit*), dan **ε** adalah galat heteroskedastik dengan rata-rata 0 dan varians adalah matriks diagonal dengan elemen $\pi(1-\pi)$.

Marginal Quasi-Likelihood (MQL) order pertama untuk π menggunakan pendekatan order pertama deret Taylor dari $f(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})$ dengan $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ dan $\mathbf{u}=0$, dan $\boldsymbol{\beta}_0$ adalah penaksir dari efek tetap. Pendekatan ini memiliki struktur model linier multilevel yang dapat dicocokkan dengan menggunakan algoritma standar (Goldstein [5]), yang diarahkan untuk memperbaiki penaksir beta, yang kemudian digunakan sebagai nilai awal yang baru. Prosedurnya diiterasi hingga didapat nilai yang konvergen.

Menurut *Rodriguez* dan Goldman [13] penggunaan MQL akan memberikan hasil yang baik apabila efek acaknya kecil dalam arti variansnya mendekati nol. Tetapi hasilnya kurang baik apabila efek acaknya sedang atau besar.

Prosedur alternatif yang akan memberikan hasil yang lebih baik adalah dengan menggunakan nilai awal yang tidak nol untuk efek acak dalam deret Taylor. Caranya adalah dengan memperluas $f(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})$ dengan $\mathbf{u} = \mathbf{u}_0$ yaitu penaksir Bayes empiris atau prediktor dari efek acak, yang didefinisikan sebagai rata-rata dari $f(\mathbf{u}|\mathbf{y})$ yang ditaksir pada nilai parameter yang ditentukan. Untuk sebuah nilai $\mathbf{u}_0$ tertentu penyelesaiannya juga dengan mendekati model linier multilevel yang dapat ditaksir dengan menggunakan algoritma standar. Penaksir-penaksir yang telah diperbaiki baik untuk efek tetap maupun acak kemudian digunakan untuk mendapatkan model linier pendekatan yang baru, dan prosedur diiterasi hingga didapat hasil yang konvergen.

Secara umum untuk respon dengan kategori ordinal dapat dianalisis menggunakan model logistik multilevel dengan mengubah logit pada Persamaan 1 dengan *cumulative logit*. Misalkan respon memiliki *C* kategori dan peluang masing-masing kategori secara berurutan adalah $p_1, p_2, \ldots, p_C$ maka fungsi penghubung logit untuk respon ordinal adalah sebagai berikut.

$$logit(\pi_{ijc}) = log\left(\frac{\pi_{ijc}}{1-\pi_{ijc}}\right) = \beta_{0c} + \beta_1 X_1 + \ldots + \beta_p X_p + u_{ijc} \qquad ...(11)$$

dengan

$$\pi_{ijc} = P(Y_{ij} \leq c) = \sum_{k=1}^{c} p_{ijk}$$

Sehingga akan terbentuk sebanyak *C*-1 persamaan model. Model-model yang terbentuk akan memilki nilai intersep yang berbeda-beda dengan slope yang sama. Untuk menaksir model logistik ordinal multilevel seperti pada persamaan 11 juga menggunakan PQL seperti telah dijelaskan sebelumnya. (Hedeker [6]).

3. HASIL PENELITIAN DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang dipublikasikan oleh RAND *Labor and Population* tahun 2004 yaitu data survei rumah tangga dan komunitas gelombang ketiga (IFLS3) dan data hasil PODES Jawa Barat Tahun 2006. Data IFLS diukur pada tingkat rumah tangga sedangkan data PODES 2006 merupakan hasil sensus yang diukur pada tingkat desa yang kemudian disesuaikan pada tingkat kecamatan. Peubah-peubah yang terlibat dalam penelitian ini adalah sebagai berikut: Peubah respon (Y) adalah Tingkat Pendidikan Anak sedangkan peubah penjelas pada level 1 adalah Jenis Kelamin Anak (X_1), Pendidikan Ibu (X_2) dan Pendidikan Ayah (X_3). Untuk peubah penjelas pada level 2 adalahstatus daerah (Z_1),banyak SMA di kecamatan (Z_2), danpersentase petani di kecamatan (Z_3).

Model yang digunakan dalam penelitian ini adalah model intersep acak yaitu model dengan nilai *intersep* merupakan komponen acak tetapi *slope* tetap.Pada model ini peubah

penjelas pada level 2 disertakan kedalam model. Dalam proses pengolahan, perhitungan serta analisis data, digunakan paket "ordinal" dalam software R versi 3.0.1. (Bliese [3])

Hasil dari pengolahan data disajikan dalam bentuk tabel dan akan dijelaskan pada uraian berikut ini. Hasil yang pertama adalah pengujian struktur hierarki dari data diperlihatkan pada tabel berikut ini

Tabel 1
Pengujian Struktur Hierarki

Model	Jum Parameter	AIC	logLik	LR.stat	df	*p-value*
Tanpa struktur hierarki	4	2446.7	-1219.4			
Dengan struktur hierarki	5	2228.7	-1109.4	220	1	< 2.2e-16

Berdasarkan hasil tersebut terlihat bahwa hasil pengujiannya signifikan yang berarti bahwa terdapat struktur hierarki pada data. Oleh karena itu model multilevel akan digunakan untuk menganalisis data yang digunakan.

Model multilevel yang digunakan adalah model intersep acak. Hasil pengujan kecocokan modelintersep acak untuk data pendidikan adalah sebagai berikut:

Tabel 2
Pengujian Kecocokan Model

Model	Jum Parameter	AIC	logLik	LR.stat	df	*p-value*
Null Model	5	2228.7	-1109.4			
Full Model	17	1995.1	-978.6	261.61	12	< 2.2e-16

Berdasarkan hasil tersebut terlihat bahwa hasil pengujiannya signifikan yang berarti bahwa model yang diusulkan cocok dengan data. Dengan demikian model multilevel yang digunakan adalah model intersep acak.

Selanjutnya penaksiran parameter dan pengujian hipotesis secara parsial dapat dilihat pada tabel berikut ini

Tabel 3
Taksiran Parameter Slope dan Pengujian Parsial

Variabel	Penaksir	*Std.Error*	Z	*p-value*
JK (Perempuan)	-0.327325	0.136488	-2.398	0.016476
Pend.ibu (SD)	1.054753	0.184196	5.726	1.03e-08
Pend.ibu (SLTP)	1.235133	0.361812	3.414	0.000641
Pend.ibu (SLTA)	3.184304	0.489255	6.508	7.59e-11
Pend.ibu (PT)	0.744840	0.639589	1.165	0.244197
Pend.ayah (SD)	0.419018	0.190405	2.201	0.027759
Pend.ayah (SLTP)	1.567451	0.291900	5.370	7.88e-08
Pend.ayah (SLTA)	1.595201	0.322790	4.942	7.74e-07
Pend.ayah (PT)	0.722488	0.426850	1.693	0.090531
Status (Urban)	1.151834	0.358342	3.214	0.001307
Jml.SMA	0.068036	0.020289	3.353	0.000798
Persen.petani	0.005901	0.007122	0.829	0.407361

Berdasarkan Tabel 3 penaksir parameter Level 1 untuk variabel Pendidikan Ibu untuk kategori PT dan variabel Pendidikan Ayah untuk kategori PT tidak signifikan. Sedangkan penaksir parameter level 2 untuk variabel persentase petani tidak signifikan. Untuk penaksir parameter intersep dapat dilihat pada tabel berikut ini

Tabel 4
Taksiran Parameter Intersep dan Pengujian Parsial

Model	Penaksir	Std.Error	z-value
Tidak Tamat SD|SD	-1.3284	0.5059	-2.626
SD|SLTP	1.2223	0.5064	2.414
SLTP|SLTA	2.7791	0.5160	5.386
SLTA|PT	5.0837	0.5373	9.461

Berdasarkan Tabel 4 terlihat bahwa efek intersep semakin tinggi tingkat pendidikan memberikan efek yang semakin besar. Hasil taksiran parameter acak untuk data tersebut adalah sebagai berikut

Tabel 5
Taksiran Parameter dan Pengujian Parsial

Groups	Varians	Simp Baku
Kecamatan	0.2577	0.5077

Dari Tabel 5 terlihat bahwa efek acak dari data tersebut sebesar 0.2577 yang berarti variasi tingkat pendidikan seseorang di antara kecamatan yang ada di Jawa Barat sebesar 0.2577.

4. SIMPULAN

Dari uraian pada bagian sebelumnya dapat disimpulkan bahwa untuk data hierarki dengan respon kategori berskala ukur ordinal dapat dianalisis menggunakan model regresi logistik multilevel dengan fungsi penghubung *cumulative logit*. Model yang digunakan adalah *multievel proportional odds model* atau *cumulative logit mixed model*.

Dari hasil pengujian data pendidikan dapat disimpulkan bahwa data yang digunakan memiliki struktur hierarki sehingga analisis untuk data tersebut menggunakan *cumulative logit mixed model.* Simpulan dari hasil analisis adalah variabel level rumah tangga yaitu Pendidikan Ibu dan Pendidikan Ayah untuk kategori PT tidak signifikan sedangkan variabel level kecamatan yaitu persentase petani tidak signifikan. Efek itersep semakin tinggi seiring semakin tingginya tingkat pendidikan dengan variasi tingkat pendidikan seseorang di antara kecamatan yang ada di Jawa Barat sebesar 0.2577.

DAFTAR PUSTAKA

[1] Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*, 2nd*Edition*. John Wiley & Sons, Inc.

[2] Agresti, Alan.2002.*Categorical Data Analysis*. *2nd edition*. New York: John Wiley & Sons, Inc.

[3] Bliese, P. 2006. *Multilevel Models in R* (2.2). R Development Core Team.

[4] Dobson, Annette J. 2002. *An Introduction to Generalized Linear Models* 2nd*edition*. London. Chapman & Hall.

[5] Goldstein, Harvey. 1995. *Multilevel Statistical Model2nd ed*., London, Arnold.

[6] Hedeker, Donald. 2007. Multilevel Models for Ordinal and Nominal Variables. *Handbook of Multilevel Analysi: edited by Leeuw and Meijer*. New York. Springer.

[7] Hesketh, S.,Rabe. 2003. *Multilevel modeling of ordered and unordered categorical Responses.* London. Institute of Child Health.

[8] Hox, J.J. 2002.*Multilevel Analysis: Techniques and Applications*. New Jersey. Lawrence Erlbaum Associates Publishers.

[9] Jones, B.S. & Steenbergen, M.R. 1997. *Modelling Multilevel Data Structures*. Paper prepared in 14th annual meeting of the political methodology society. Columbus. OH.

[10] Kramer, M. 2005. R^2 *Statistics for Mixed Models*. Published Paper in Biometrical Consulting Service, ARS (Beltsville, MD), USDA.

[11] McCullagh and Nelder. 1989. *Generalized Linear Models*. 2nd*edition*. , London. Chapman & Hall.

[12] Ringdal, K. 1992. Methods for Multilevel Analysis. *Acta Sosiologica* 35**:**235-243.

[13] Rodriguez, G., Goldman, N. 2001. Improved estimation procedures for multilevel models with binary response: a case-study, *Journal Royal Statist.Soc A,* **164**, Part 2 pp 339-355

[14] Snijder, Tom A. B., Bosker, Roel J. 1999.*Multilevel Analysis: An introduction to basic and advance multilevel modelling*. London. SAGE Publications.