Statistika, Vol. 16 No. 1, 29 – 39
Mei 2016

Masalah Overdispersi dalam Model Regresi Logistik Multinomial

ANNISA LISA NURJANAH, NUSAR HAJARISMAN, TETI SOFIA YANTI

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung,
Jl. Tamansari No. 1 Bandung 40116
e-mail: annisalisanur@gmail.com, nusarhajarisman@yahoo.com

ABSTRAK

Model regresi logistik multinomial merupakan pengembangan dari model regresi logistik binomial dimana variabel responnya mempunyai lebih dari dua kategori (politokomus). Model ini juga merupakan kelompok model linear terampat (generalized linear model), dimana komponen acaknya mengasumsikan bahwa distribusi dari variabel respon mengikuti distribusi multinomial. Salah satu asumsi yang harus dipenuhi dalam model regresi logistik multinomial ini adalah variabel responnya merupakan variabel acak yang saling bebas dan kategorinya bersifat mutually exclusive. Apabila asumsi ini dilanggar maka akan muncul masalah yang dikenal dengan masalah overdispersi. Konsekuensi dari adanya masalah overdispersi dalam data akan menghasilkan suatu model yang tidak valid. Salah satu cara untuk mengatasi masalah overdispersi dalam model regresi logistik multinomial yang akan dibahas dalam makalah ini adalah mengadopsi apa yang dilakukan oleh McCullagh dan Nelder (1989) dengan mengkoreksi matriks varians kovariansnya. Model regresi logistik multinomial ini kemudian akan diaplikasikan untuk mengetahui pengaruh dari jenis kelamin dan perilaku merokok orang tua terhadap perilaku merokok mahasiswa Unisba. Dari model regresi logistik multinomial biasa dan dengan model regresi logistik multinomial terkoreksi dapat disimpulkan bahwa variabel-variabel prediktor yang dianggap berarti dalam kedua pemodelan tersebut berbeda. Perbedaan lainnya terdapat pada nilai galat baku model regresi logistik multinomial biasa lebih kecil dari yang seharusnya dengan kata lain underestimate dibandingkan dengan model regresi logistik multinomial terkoreksi, dan selang kepercayaan untuk rasio odds menjadi pendek dibandingkan dengan model regresi logistik multinomial terkoreksi.

Kata Kunci : Data Politokomus, Distribusi Multinomial, Model Linear Terampat, Overdispersi.

1. PENDAHULUAN

Teknik pemodelan statistika merupakan suatu metode untuk mengeksplorasi informasi berupa data guna memahami lebih mendalam situasi yang dihadapi. Tak jarang para peneliti menggunakan model regresi logistik multinomial untuk memodelkan suatu data. Model regresi logistik multinomial merupakan model yang tujuannya adalah memprediksi variabel respon yang berskala nominal ataupun ordinal berdasarkan satu atau lebih variabel prediktor. Model ini merupakan pengembangan dari model regresi logistik binomial dimana variabel responnya mempunyai lebih dari dua kategori (politokomus). Sebagaimana model regresi diskrit lainnya, variabel prediktor yang digunakan dalam model regresi logistik multinomial dapat berupa nominal dan/atau kontinu, atau bahkan berbentuk interaksi antar variabel prediktor dalam memprediksi variabel respon. Perlu diketahui bahwa menurut McCullagh dan Nelder (1989) model ini juga merupakan kelompok model linear terampat (*generalized linear model*), dimana komponen acaknya mengasumsikan bahwa distribusi dari variabel respon mengikuti distribusi multinomial.

Pada saat memodelkan data dengan respon politokomus seperti ini, perlu diperhatikan apakah variabel responnya berskala ordinal atau nominal. Pada model regresi logistik biner atau binomial hal tersebut tidak menjadi perhatian. Sebagaimana yang dinyatakan oleh Agresti (2007) beberapa model hanya tepat digunakan pada respon ordinal (seperti model logit kumulatif, model kategorik *adjacent*, model rasio kontinu). Adapula beberapa model lainnya yang digunakan ketika variabel responnya berskala nominal ataupun ordinal, seperti model logit dasar (*baseline logit model*) dan model logit bersyarat (*conditional logit model*).

Ada beberapa asumsi dasar yang harus dipenuhi pada saat mengaplikasikan model regresi logistik multinomial pada gugus data tertentu. Asumsi-asumsi itu diantaranya adalah bahwa

(1) variabel respon merupakan variabel acak yang saling bebas dan kategorinya bersifat *mutually exclusive*; (2) tidak terdapat masalah multikolinearitas diantara variabel prediktor yang diamatinya; (3) adanya transformasi logit pada variabel respon; serta (4) tidak ada data pencilan yang berpotensi sebagai data yang berpengaruh. Apabila terdapat satu atau lebih asumsi yang tidak terpenuhi, maka akan diperoleh suatu model regresi logistik multinomial yang tidak valid. Jika hal ini terjadi, maka tentu perlu ada upaya untuk mengatasi masalah akibat adanya pelanggaran asumsi, sehingga akan diperoleh suatu model yang valid. Dalam hal ini akan difokuskan pada salah satu asumsi mengenai independensi dari variabel respon. Ketidakbebasan antar variabel respon dimaknai sebagai adanya korelasi diantara variabel respon, hal tersebut merupakan suatu indikasi adanya masalah overdispersi dalam data. Sebagaimana yang diungkapkan oleh McCullagh dan Nelder (1989) masalah overdispersi akan sering dijumpai dalam analisis data diskrit, baik variabel respon yang berbentuk biner (dikotomus), cacahan, maupun politokomus seperti dalam model regresi logistik multinomial ini.

Munculnya masalah overdispersi dalam pengamatan data diskrit dapat dijelaskan oleh dua hal, yaitu: adanya keragaman dalam peluang respon dan adanya korelasi antar variabel respon. Kedua kejadian tersebut merupakan kejadian yang saling berhubungan, artinya jika terdapat keragaman dalam peluang respon, maka terdapat korelasi antar variabel respon. Begitu juga sebaliknya, jika terdapat korelasi antara variabel respon, maka terdapat keragaman dalam peluang respon. McCullagh dan Nelder (1989) menyatakan bahwa kedua kejadian tersebut dapat terjadi karena adanya pengelompokan (*clustering*) dalam populasi. Sedangkan Collet (1991) menyebutkan bahwa kejadian-kejadian tersebut muncul karena sejumlah unit percobaan diamati beberapa kali pada kondisi yang sama, sehingga akan diperoleh suatu peluang respon yang berbeda dari satu percobaan ke percobaan yang lainnya.

Penggunaan metode statistika yang mengasumsikan ketidakbebasan antara variabel respon (seperti dalam regresi logistik binomial ataupun multinomial) dapat menjadi tidak tepat. Jika terdapat korelasi antar pengamatan, maka nilai penaksir parameter dari model tidak memberikan korelasi yang mungkin mempunyai galat baku yang bersifat *underestimated* jika terdapat korelasi yang positif (Cox and Snell, 1989). Konsekuensi lain dari adanya masalah overdispersi dalam data diskrit adalah pada nilai penaksir variansnya. Apabila penaksir varians ini digunakan untuk menghitung selang kepercayaan, maka akan diperoleh rata-rata yang terlalu kecil sehingga akan berakibat pada selang kepercayaan yang terlalu pendek. Apabila penaksir varians ini digunakan untuk mengerjakan pengujian hipotesis statistik, maka akan selalu menolak hipotesis H_0.

Berdasarkan hal tersebut, maka perlu dicari suatu metode untuk mendapatkan solusi statistika yang tepat dalam menentukan hubungan fungsional antara satu atau lebih variabel prediktor dengan satu variabel respon politokomus yang tidak saling bebas (berkorelasi). Salah satu cara untuk mengatasi masalah overdispersi dalam model regresi logistik multinomial yang akan dibahas dalam makalah ini adalah dengan mengkoreksi matriks kovariansnya sebagaimana yang diungkapkan oleh McCullagh dan Nelder (1989). Model regresi logistik multinomial yang dibahas dalam skripsi ini adalah model logit dasar (*baseline logit model*) dengan pertimbangan bahwa model tersebut dapat diaplikasikan pada respon yang berskala nominal ataupun ordinal. Model tersebut kemudian akan diaplikasikan untuk mengetahui pengaruh dari jenis kelamin dan perilaku merokok orang tua terhadap perilaku merokok mahasiswa Unisba. Adapun tujuan yang ingin dicapai dari penelitian ini adalah:

1. Mendekteksi masalah overdispersi dalam pemodelan regresi logistik multinomial.
2. Mengatasi masalah overdispersi dalam pemodelan regresi logistik multinomial.
3. Membandingkan model regresi logistik multinomial biasa dengan model regresi logistik multinomial terkoreksi.

2. LANDASAN TEORI

Distribusi Multinomial

Saat percobaan mempunyai lebih dari dua respon yang mungkin. Tinjau Y sebagai variabel acak dengan J kategori. Anggap $\pi_1, \pi_2, \dots, \pi_J$ menunjukan masing-masing peluang dengan $\pi_1 + \pi_2 + \dots + \pi_J = 1$. $\boldsymbol{y}$ membentuk vektor kolom dengan unsur $y_i,\ i = 1, \dots J$ dengan $\sum_{j=1}^{J} y_j = n$

Fungsi masa peluang dari variabel acak Y yang berdistribusi multinomial adalah

$$f(\boldsymbol{y}|n) = \frac{n!}{y_1!\, y_2! \ldots y_J!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_J^{y_J} \tag{2.1}$$

Jika $J = 2$ lalu $\pi_2 = 1 - \pi_1$, $y_2 = n - y_1$ dan persamaan (2.1) akan menjadi fungsi peluang distribusi binomial. Secara umum persamaan (2.1) tidak memenuhi persyaratan untuk menjadi anggota distribusi keluarga eksponensial. Namun hubungan dengan distribusi poisson berikut akan memastikan bahwa *Generalized Linear Model* (GLM) sesuai. Anggap $Y_1, \ldots, Y_J$ menunjukan variabel acak yang saling bebas dengan distribusi $Y_j \sim Poisson(\lambda_j)$.

Distribusi peluang bersama adalah

$$f(\boldsymbol{y}) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \tag{2.2}$$

dimana, $\boldsymbol{y}$ adalah vektor kolom dengan unsur y_i, $i = 1, \ldots J$. Anggap $n = Y_1 + Y_2 + \cdots + Y_J$, lalu n adalah variabel acak dengan distribusi $n \sim Poisson(\lambda_1, +\lambda_2 + \ldots + \lambda_J)$ (Kalbfleisch, 1985). Oleh karena itu distribusi dari y bergantung pada n

$f(\boldsymbol{y}|n) = \dfrac{\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}}{\frac{(\lambda_1+\lambda_2+\ldots+\lambda_J)^n e^{-((\lambda_1+\lambda_2+\ldots+\lambda_J))}}{n!}}$ yang mana dapat disederhanakan menjadi

$$f(\boldsymbol{y}|n) = \left(\frac{\lambda_1}{\sum \lambda_k}\right)^{y_1} \ldots \left(\frac{\lambda_J}{\sum \lambda_k}\right)^{y_J} \frac{n!}{y_1! \ldots y_J!} \tag{2.3}$$

Jika $\pi_j = \lambda_j \left(\sum_{k=1}^{K} \lambda_k\right)^{-1}$ untuk $j = 1, \ldots, J$. Lalu persamaan (2.3) sama seperti persamaan (2.1) dan $\sum_{j=1}^{J} \pi_j = 1$, sebagaimana yang disyaratkan. Oleh karena itu distribusi multinomial dapat dianggap sebagai distribusi gabungan dari variabel-variabel acak Poisson. Hasil ini membenarkan penggunaan *Generalized Linear Model* (GLM). Distribusi multinomial mempunyai:

$$E(Y_j) = n\pi_j$$
$$var(Y_j) = n\pi_j(1 - \pi_j)$$
$$Cov(Y_j, Y_k) = -n\pi_j\pi_k$$

Model Regresi Logistik Multinomial

Regresi logistik multinomial merupakan salah satu pendekatan pemodelan yang dapat digunakan untuk mendeskripsikan hubungan beberapa variabel kovariat X dengan suatu variabel respon politokomus. Misal J merupakan banyaknya kategori diskrit dari variabel respon, dimana $J \geq 2$. Variabel acak Y dapat mengambil salah satu dari nilai J yang mungkin. Setiap pengamatan saling bebas dan setiap y_i adalah variabel acak multinomial. Data dikumpulkan ke dalam setiap populasi yang merepresentasi satu kombinasi variabel-variabel prediktor.

Matriks $\boldsymbol{y}$ adalah matriks dengan N baris dan $J - 1$ kolom. Untuk setiap populasi, y_{ij} mewakili perhitungan pengamatan nilai ke j dari y_i. Demikian pula $\boldsymbol{\pi}$ adalah matriks dari dimensi yang sama seperti y dimana setiap unsur π_{ij} menunjukan peluang pengamatan nilai ke-j variabel respon dalam populasi ke-i. Desain matriks variabel prediktor $\boldsymbol{X}$, berukuran N baris dan $(K + 1)$ kolom, dimana K adalah banyaknya variabel prediktor dan unsur pertama pada setiap baris $x_{i0} = 1$ sebagai intersep. $\boldsymbol{\beta}$ menjadi matriks dengan $K + 1$ baris dan $J - 1$ kolom, sehingga setiap unsur β_{kj} mengandung penaksiran parameter untuk kovariat ke-k dan nilai variabel respon ke-j. Model regresi logistik multinomial menyamakan komponen linear dengan log dari odds pengamatan ke-j dibandingkan dengan pengamatan ke-J. Model regresi logistik multinomial adalah

$$log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = log\left(\frac{\pi_{ij}}{1 - \sum_{j=1}^{J-1} \pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_k \;\; ; i = 1, 2, \ldots, N \, ; j = 1, 2, \ldots J - 1 \tag{2.4}$$

dimana

$$\pi_{ij} = \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \qquad \text{Jika } j < J, \text{ dan} \qquad \pi_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_k}}$$

Penaksiran Parameter

Untuk setiap populasi, variabel respon mengikuti distribusi multinomial dengan tingkat J. Fungsi densitas gabungannya adalah

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\left[\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!}\prod_{j=1}^{J}\pi_{ij}{}^{y_{ij}}\right] \tag{2.5}$$

Kita ingin memaksimumkan persamaan (2.5) terhadap $\boldsymbol{\beta}$, istilah faktorial diberlakukan sebagai kosntanta yang diabaikan. Fungsi log likelihoodnya adalah sebagai berikut:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{J-1}\left(y_{ij}\sum_{k=0}^{K} x_{ik}\beta_k\right) - n_i log\left(1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}\right) \tag{2.8}$$

Turunan pertama dari fungsi log likelihood adalah

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i\, x_{ik} - n_i \pi_{ij} x_{ik}$$

Turunan pertama dari fungsi log likelihood merupakan sebuah matriks berukuran $(J-1)(K+1)$ yang akan diatur sama dengan nol. Turunan parsial kedua dari matriks untuk model regresi logistik multinomial mempunyai dua hasil turunan yaitu ketika $j' = j$ dan $j' \neq j$

$$\frac{\partial^2 l(\beta)}{\partial \beta_{kj}\partial \beta_{k'j'}} = \sum_{i=1}^{N} n_i \pi_{ij}(1-\pi_{ij}) x_{ik}\, x_{ik'} \qquad j' = j$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_{kj}\partial \beta_{k'j'}} = \sum_{i=1}^{N} n_i \pi_{ij}\pi_{ij'}\, x_{ik} x_{ik'} \qquad j' \neq j$$

Langkah selanjutnya memecahkan persamaan non-linear dari fungsi log-likelihood yang ditaksir dengan metode numerik menggunakan proses iterasi Newton-Raphson. Misalkan,

$$\underset{(N\times 1)}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \underset{(N\times p)}{\boldsymbol{X}} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}, \qquad \underset{(N\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$$

Didefinisikan bahwa

$$\underset{(N\times N)}{\boldsymbol{W}} = Diag\left(n_i \pi_i (1-\pi_i)\right)$$

Dengan menggunakan perkalian matriks turunan pertama dari fungsi log likelihood yang diatur sama dengan nol dan turunan kedua masing-masing dapat ditunjukan sebagai berikut:

$$l'(\beta) = \boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{\mu}) \qquad\qquad l''(\beta) = -\boldsymbol{X^T W X}$$

$l''(\beta)$ merupakan turunan kedua dari fungsi log likelihood yang berupa matriks berukuran $(K+1)(K+1)$. Sehingga,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left[\boldsymbol{X^T W^{(t)} X}\right]^{-1}\left(\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{\mu})\right) \tag{2.9}$$

Proses tersebut berlangsung secara terus menerus hingga konvergen, artinya tidak terdapat perubahan antara unsur $\boldsymbol{\beta}$ dari satu iterasi ke iterasi berikutnya. Apabila penaksiran kemungkinan maksimum dikatakan telah konvergen maka matriks $[\boldsymbol{X^T W X}]^{-1}$ akan menjadi matriks varians-kovarians dari penaksiran parameter.

Ukuran Kecocokan Model

Tabel 1. Ukuran Kecocokan Model Regresi Logistik Multinomial

Ukuran Kecocokan Model	Persamaan		Keterangan
Residual chi-kuadrat Pearson	$r_i = \sum_{i=1}^{N}\sum_{j}^{J-1} \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{V(\hat{\mu}_{ij})}}$	(2.10)	y_{ij}: pengamatan $\hat{\mu}_{ij}$: ekspektasi untuk distribusi multinomial $V(\hat{\mu}_{ij})$: penaksir varians untuk distribusi multinomial
Statistik chi-kuadrat	$\chi^2 = \sum_{i=1}^{N} r_i^2$	(2.11)	
Devians	$D = 2[l(\boldsymbol{b}_{maks}) - l(\boldsymbol{b})]$	(2.12)	$l(\boldsymbol{b})$: nilai maksimum fungsi log-likelihood untuk model dugaan
Rasio kemungkinan statistik chi-kuadrat	$C = 2[l(\boldsymbol{b}) - l(\boldsymbol{b}_{min})]$	(2.13)	$l(\boldsymbol{b}_{maks})$: nilaimaksimum fungsi log-likelihood untuk model
Pseudo R^2	$Pseudo\ R^2 = \frac{l(\boldsymbol{b}_{min}) - l(\boldsymbol{b})}{l(\boldsymbol{b}_{min})}$	(2.14)	$l(\boldsymbol{b}_{min})$: nilai maksimum fungsi log-likelihood untuk model minimal

Jika model baik maka χ^2 dan D mempunyai distribusi asimtotik $\chi^2_{(N-p)}$ dimana p adalah banyaknya parameter yang ditaksir. C mempunyai distribusi asimtotik $\chi^2_{[p-(J-1)]}$ karena model minimal akan mempunyai satu parameter untuk setiap definisi logit pada persamaan (2.4). Model dugaan yang baik akan mendekati atau sama dengan banyaknya derajat bebas.

Interpretasi Model

Seringkali dalam praktiknya lebih mudah untuk menginterpretasikan efek dari variabel prediktor dalam hal ini adalah rasio *odds* daripada parameter $\boldsymbol{\beta}$. Tinjau variabel respon dengan J kategori dan variabel prediktor x biner ($x = 1$ dan $x = 0$). Rasio *Odds* untuk penjelasan respon j ($j = 2, \dots, J$) relatif terhadap kategori referensi pertama, $j = 1$.

$$OR_j = \frac{\pi_{jp}/\pi_{ja}}{\pi_{1p}/\pi_{1a}} \tag{2.15}$$

dimana π_{jp} dan π_{ja} menunjukan peluang kategori respon j ($j = 2, \dots, J$) berdasarkan penjelasan dari masing-masing variabel prediktor ($x = 1$ dan $x = 0$). Untuk model

$$log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x. \quad j = 2, \dots, J$$

Log odds adalah sebagai berikut:

$$logOR_j = log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) = \beta_{1j}$$

Maka dari itu $OR_j = \exp(\beta_{1j})$, dimana $\exp(\beta_{1j})$ ditaksir oleh $\exp(b_{1j})$. Jika $\beta_{1j} = 0$ lalu $OR_j = 1$, maka faktor dalam variabel prediktor tidak berpengaruh. Sebagai contoh, batas kepercayaan 95% untuk OR_j diberikan oleh $exp[b_{1j} \pm 1{,}96.SE(b_{1j})]$ dimana $SE(b_{1j})$ menunjukan galat baku dari b_{1j}. Interval kepercayaan yang tidak termasuk dalam kesatuan yang sesuai dengan nilai-nilai β secara signifikan berbeda dari nol. Pilihan kategori sebagai kategori referensi untuk variabel respon akan mempengaruhi penaksiran parameter $\boldsymbol{b}$ tetapi tidak akan mempengaruhi taksiran peluang $\hat{\boldsymbol{\pi}}$ atau nilai kecocokan (Dobson, 2002).

Overdispersi

Istilah overdispersi dapat diartikan bahwa varians dari respon $\boldsymbol{Y}$ melebihi varians multinomial, $n\pi(1-\pi)$. Dalam praktiknya, tak jarang terjadi overdispersi. Cara yang paling sederhana menurut (McCullagh & Nelder, 1989) untuk mendeteksi adanya overdispersi adalah:

$\frac{nilai\ Devians}{derajat\ bebas} > 1$, dan $\frac{nilai\ chi-kuadrat\ Pearson}{derajat\ bebas} > 1$

yang artinya ketika hasil bagi antara nilai Devians dan chi-kuadrat Pearson dengan derajat bebas tersebut lebih besar dari 1 (satu) maka dalam kasus tersebut terindikasi adanya overdispersi. Jika $E(\pi_i) = \pi$ dan $var(\pi_i) = \emptyset\pi(1-\pi)$, ekspektasi dan varians tersebut mungkin menunjukan bahwa bentuk ekspektasi dan varians tanpa syarat dari Y adalah

$E(Y) = n\pi$

$$Var(Y) = n\pi(1-\pi)\{1 + (n-1)\emptyset\} = \sigma^2 n\pi(1-\pi) \tag{2.16}$$

dengan parameter dispersi $\sigma^2 = 1 + (n-1)\emptyset$ berdasarkan pada variabilitas π pada ukuran sampel n. Terjadinya overdispersi tidak akan berpengaruh pada ekspektasi, tapi berpengaruh pada varians yang terdapat pada persamaan (2.16) karena terjadi peningkatan oleh faktor σ^2 yang tidak diketahui. Matrik varians kovarians dari $\widehat{\boldsymbol{\beta}}$ yang diperoleh dari log likelihood multinomial diganti dengan

$cov(\widehat{\boldsymbol{\beta}}) \cong \sigma^2[\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}]^{-1}$

Penaksir untuk σ^2 mungkin akan berdasarkan jumlah kuadrat diboboti. Model dugaan dikatakan baik apabila

$$\tilde{\sigma}^2 = \frac{1}{N-p}\sum_i \frac{(y_i - n_i\tilde{\pi}_i)^2}{n_i\tilde{\pi}_i(1-\tilde{\pi}_i)} = \frac{\chi^2}{(N-p)} \tag{2.17}$$

dengan $\tilde{\sigma}^2$ adalah pendekatan tak bias untuk σ^2 asalkan p kecil dibandingkan dengan N. Selanjutnya, taksiran matriks varians-kovarians dari $\widehat{\boldsymbol{\beta}}$ adalah

$$\text{penaksiran } var(\widehat{\boldsymbol{\beta}}) = \tilde{\sigma}^2[\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}]^{-1} \tag{2.18}$$

Berdasarkan penaksiran parameter dengan metode iterasi Newton-Rapshon pada persamaan (2.9), proses tersebut akan berlangsung secara terus menerus hingga konvergen, artinya tidak terdapat perubahan antara unsur $\boldsymbol{\beta}$ dari satu iterasi ke iterasi berikutnya. Apabila penaksiran kemungkinan maksimum dikatakan telah konvergen maka matriks $[\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}]^{-1}$ akan menjadi matriks varians-kovarians dari penaksiran parameter. Namun, untuk memodelkan regresi logistik multinomial yang mengandung overdispersi, maka matriks varians-kovarians dari $\hat{\beta} = [\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}]^{-1}$ akan ditaksir melalui persamaan (2.18), dimana, $\tilde{\sigma}^2$ akan ditaksir oleh persamaan (2.17). Apabila nilai χ^2 dianggap masih lebih besar dibandingkan dengan derajat bebasnya, maka diperlukan iterasi kembali pada persamaan (2.9) hingga diperoleh nilai χ^2 yang mendekati nilai derajat bebasnya.

3. HASIL DAN PEMBAHASAN

Data pengamatan yang akan diteliti adalah data prilaku merokok mahasiswa dengan jenis kelamin ($\boldsymbol{X_1}$) dan prilaku merokok orang tua ($\boldsymbol{X_2}$) sebagai variabel prediktornya. Dengan bantuan *software* SAS 9.4, hasil penaksiran parameter untuk model regresi logistik multinomial menggunakan metode *maximum likelihood* disajikan dalam Tabel 2.

Berdasarkan hasil perhitungan yang tersaji pada Tabel 2 dapat diperoleh nilai-nilai taksiran parameter untuk model regresi logistik multinomial. Untuk model logit pertama (perokok sedang) nilai taksiran parameter *intercept* adalah -0.4378 (dengan galat baku 0.1232) dan nilai taksiran parameter koefisien jenis kelamin (X_1) untuk kategori 1 adalah -0.0247 (dengan galat baku 0.1247), sedangkan nilai taksiran parameter koefisien jenis kelamin kategori 0 tidak ada karena dijadikan kategori referensi atau *base level* untuk variabel jenis kelamin (X_1). Nilai taksiran parameter koefisien prilaku merokok orang tua (X_2) untuk kategori 2, dan 3 masing-masing adalah 0.2481 (dengan galat baku 0.1419), 0.1599 (dengan galat baku 0.1485), sedangkan nilai taksiran parameter koefisien prilaku merokok orang tua kategori 1 tidak ada karena dijadikan kategori referensi atau *base level* untuk variabel prilaku merokok orang tua (X_2).

Untuk model logit kedua (perokok berat) nilai taksiran parameter *intercept* adalah -2.074 (dengan galat baku 0.1736) nilai taksiran parameter koefisien jenis kelamin (X_1) untuk kategori 1 adalah -0.0247 (dengan galat baku 0.1247), sedangkan nilai taksiran parameter koefisien jenis kelamin kategori 0 tidak ada karena dijadikan kategori referensi atau *base level* untuk variabel jenis kelamin (X_1). Nilai taksiran parameter koefisien prilaku merokok orang tua (X_2) untuk kategori 2, dan 3 masing-masing adalah 0.2481 (dengan galat baku 0.1419), 0.1599 (dengan galat baku 0.1485), sedangkan nilai taksiran parameter koefisien prilaku merokok

orang tua kategori 1 tidak ada karena dijadikan kategori referensi atau *base level* untuk variabel prilaku merokok orang tua (X_2).

Tabel 2. Taksiran Parameter untuk Model Regresi Logistik Multinomial

Y	Parameter	Kode Kategori	Taksiran	Galat Baku	P-Value
Perokok Sedang	Intercept		-0.4378	0.1232	0.0004*
	Jenis Kelamin	1	-0.0247	0.1247	0.8433
	Prilaku Merokok Orang Tua	2	0.2481	0.1419	0.0803*
	Prilaku Merokok Orang Tua	3	0.1599	0.1485	0.2815
Perokok Berat	Intercept		2.074	0.1736	0.0001*
	Jenis Kelamin	1	-0.0247	0.1247	0.8433
	Prilaku Merokok Orang Tua	2	0.2481	0.1419	0.0803*
	Prilaku Merokok Orang Tua	3	0.1599	0.1485	0.2815

*)signifikan pada $\alpha = 10\%$

Berdasarkan nilai *P-value* bahwa variabel-variabel yang dapat dianalisis selanjutnya dengan taraf 10% pada model logit pertama dan model logit kedua adalah X_2 kategori 2, sedangkan X_1 kategori 1 dan X_2 kategori 3 tidak signifikan terhadap model. Model logit pertama dan kedua dapat ditulis sebagai berikut

$$log\left(\frac{\pi_1}{\pi_0}\right) = -0.4387 + 0.2481\, X_2(2)$$

$$log\left(\frac{\pi_2}{\pi_0}\right) = 2.0740 + 0.2481\, X_2(2)$$

Berdasarkan kedua fungsi logit di atas, dapat ditulis model peluang Prilaku Merokok Mahasiswa kategori ringan, sedang, dan berat masing-masing adalah sebagai berikut:

$$\pi_0(x) = \frac{1}{1 + \exp\left(-0.4387 + 0.2481\, X_2(2)\right) + exp\left(2.0740 + 0.2481\, X_2(2)\right)}$$

$$\pi_1(x) = \frac{exp\left(-0.4387 + 0.2481\, X_2(2)\right)}{1 + \exp\left(-0.4387 + 0.2481\, X_2(2)\right) + exp\left(2.0740 + 0.2481\, X_2(2)\right)}$$

$$\boldsymbol{\pi_2(x) = \frac{exp\left(2.0740 + 0.2481\, X_2(2)\right)}{1 + \exp\left(-0.4387 + 0.2481\, X_2(2)\right) + exp\left(2.0740 + 0.2481\, X_2(2)\right)}}$$

Tabel 3. *Output* Rasio Kemungkinan Chi-kuadrat

***Testing Global Null Hypothesis*: BETA=0**	
Pengujian	**P-value**
Rasio Kemungkinan	0.1552
***Pseudo R Square*=0.0125**	

Sumber: Hasil Pengolahan *Software* SAS 9.4

Salah satu ukuran kecocokan model berdasarkan Tabel 3 untuk nilai rasio kemungkinan chi-kuadrat dihasilkan nilai *P-value* 0.1552 lebih besar dari $\alpha = 10\%$ artinya tidak terdapat parameter yang signifikan dalam model regresi logistik multinomial dan *Pseudo-*R^2 hanya 1.25% variasi yang dijelaskan oleh faktor-faktor ini. Ukuran kecocokan model lainnya adalah sebagai berikut:

Tabel 4. *Output* Devians dan Chi-kuadrat Pearson

Kriteria	Nilai	db	Nilai/db	*P-value*
Devians	10.9106	7	1.5587	0.1426
Pearson	11.0443	7	1.5778	0.1367

Sumber: Hasil Pengolahan *Software* SAS 9.4

Berdasarkan Tabel 4 diperoleh nilai chi-kuadrat Pearson dan Devians masing-masing adalah 11.0443 dan 10.9106, serta *P-value* 0.1367 dan 0.1426. Keduanya mempunyai *P-value* lebih besar dari $\alpha = 10\%$ artinya model regresi logistik multinomial cocok memodelkan hubungan antara prilaku merokok mahasiswa dengan variabel prediktor jenis kelamin dan prilaku merokok orang tua. Dari perbandingan antara nilai chi-kuadrat Pearson dan Devians terhadap derajat bebasnya lebih dari 1 (satu) sehingga dapat diidentifikasikan bahwa pada pengamatan ini terjadi peristiwa overdispersi yang perlu adanya tindak lanjut. Selanjutnya akan dibahas mengenai rasio *odds* yang terdapat dalam Tabel 4.5 sebagai berikut:

Tabel 5. *Output* Rasio *Odds*

Penaksiran Rasio *Odds*			
Efek	**Penaksiran Titik**	**Penaksiran Interval**	
		Batas Atas	**Batas Bawah**
Jenis Kelamin	0.952	0.584	1.552
Prilaku Merokok Orang Tua(kategori 2)	1.927	1.08	3.439
Prilaku Merokok Orang Tua(kategori 3)	1.765	0.971	3.208

Sumber: Hasil Pengolahan *Software* SAS 9.4

Tabel 5 menampilkan output penaksir rasio *odds* dan selang kepercayaan rasio *odds*. Terlihat bahwa batas kepercayaan rasio odds untuk variabel prediktor jenis kelamin mencakup nilai 1 (satu) yang menunjukan bahwa variabel prediktor jenis kelamin dan variabel prilaku merokok orang tua pada kategori 3 (tiga) bukan merupakan variabel yang signifikan pada prilaku merokok mahasiswa. Penaksir rasio *odds* untuk mahasiswa dengan prilaku merokok orang tua dengan kategori 2 yaitu hanya terdapat salah satu diantara orang tuanya yang merokok mempunyai kemungkinan tingkat prilaku merokok yang lebih tinggi dibandingkan dengan mahasiswa dengan prilaku merokok orang tua dengan kategori 1 yaitu kedua orang tuanya merokok.

Saat overdispersi terjadi dalam pemodelan maka perlu adanya tindak lanjut, dalam hal ini adalah dengan mengoreksi matriks varians kovariansnya. Berikut adalah hasil pemodelan regresi logistik multinomial dengan penanganan overdispersi:

Tabel 6. Taksiran Parameter untuk Model Regresi Logistik Multinomial Terkoreksi

Y	Parameter	Kode Kategori	Taksiran	Galat Baku	P-Value
Perokok Sedang	Intercept		-0.4378	0.1548	0.0046*
	Jenis Kelamin	1	-0.0247	0.1567	0.8749
	Prilaku Merokok Orang Tua	2	0.2481	0.1782	0.1639
	Prilaku Merokok Orang Tua	3	0.1599	0.1865	0.3912
Perokok Berat	Intercept		2.074	0.2181	0.0001*
	Jenis Kelamin	1	-0.0247	0.1567	0.8749
	Prilaku Merokok Orang Tua	2	0.2481	0.1782	0.1639
	Prilaku Merokok Orang Tua	3	0.1599	0.1865	0.3912

*)signifikan pada $\alpha = 10\%$

Berdasarkan hasil perhitungan yang tersaji pada Tabel 6 dapat diperoleh nilai-nilai taksiran parameter untuk model regresi logistik multinomial terkoreksi. Nilai *p-value* bahwa variabel-variabel yang dapat dianalisis selanjutnya dengan taraf 10% pada model logit pertama dan model logit kedua adalah hanya *intercept*, sedangkan X_1 kategori 1, X_2 kategori 2 dan X_2 kategori 3 tidak signifikan terhadap model. Model logit pertama dan kedua dapat ditulis sebagai berikut:

$$log\left(\frac{\pi_1}{\pi_0}\right) = -0.4387$$
$$log\left(\frac{\pi_2}{\pi_0}\right) = 2.0740$$

Tabel 7. *Output* Rasio Kemungkinan Chi-kuadrat

Testing Global Null Hypothesis: BETA=0	
Pengujian	**P-value**
Rasio Kemungkinan	0.3449
Pseudo R Square=0.0008	

Sumber: Hasil Pengolahan *Software* SAS 9.4

Berdasarkan Tabel 7 untuk nilai rasio kemungkinan chi-kuadrat dihasilkan nilai *P-value* 0.3449 lebih besar dari $\alpha = 10\%$ artinya tidak terdapat parameter yang signifikan dalam model regresi logistik multinomial dan Pseudo-R^2 hanya 0.8% variasi yang dijelaskan oleh faktor-faktor ini. Selanjutnya akan dibahas mengenai rasio *odds* yang terdapat dalam Tabel 3.7 sebagai berikut:

Tabel 8.*Output* Rasio *Odds*

Penaksiran Rasio *Odds*			
Efek	**Penaksiran Titik**	**Penaksiran Interval**	
		Batas Atas	**Batas Bawah**
Jenis Kelamin	0.952	0.515	1.759
Prilaku Merokok Orang Tua(kategori 2)	1.927	0.931	3.988
Prilaku Merokok Orang Tua(kategori 3)	1.765	0.833	3.739

Sumber: Hasil Pengolahan *Software* SAS 9.4

Tabel 8 menampilkan output penaksir rasio *odds* dan selang kepercayaan rasio *odds*. Terlihat bahwa batas kepercayaan rasio odds untuk seluruh variabel prediktor mencakup nilai 1 (satu) yang menunjukan bahwa variabel-variabel prediktor bukan merupakan variabel yang signifikan pada prilaku merokok mahasiswa.

Dari model regresi logistik multinomial biasa dan dengan model regresi logistik multinomial terkoreksi dapat disimpulkan bahwa variabel-variabel prediktor yang dianggap berarti dalam kedua pemodelan tersebut berbeda dan perbedaannya terlihat pula pada nilai galat baku, dan selang kepercayaan untuk rasio *odds*.

Tabel 9. Galat Baku Model Regresi Logistik Multinomial Biasa dan Terkoreksi

Model Regresi Logistik Multinomial Biasa	**Model Regresi Logistik Multinomial Terkoreksi**
Galat Baku	**Galat Baku**
0.1232	0.1548
0.1247	0.1567
0.1419	0.1782
0.1485	0.1865
0.1736	0.2181
0.1247	0.1567
0.1419	0.1782
0.1485	0.1865

Sumber: Hasil Pengolahan *Software* SAS 9.4

Berdasarkan Tabel 9 terlihat bahwa pada saat data mengalami masalah overdispersi akan menyebabkan nilai galat baku yang lebih kecil dari yang seharusnya dengan kata lain *underestimate.* Hal ini akan berpengaruh pada penarikan kesimpulan yang kurang tepat, karena dapat membuat variabel prediktor yang pengaruhnya seharusnyata tidak nyata menjadi nyata.

Tabel 10. Penaksiran Interval Kepercayaan Rasio *Odds* Model Regresi Logistik Multinomial Biasa dan Terkoreksi

Penaksiran Rasio *Odds*			
Model Regresi Logistik Multinomial Biasa		**Model Regresi Logistik Multinomial Terkoreksi**	
Penaksiran Interval		**Penaksiran Interval**	
Batas Atas	**Batas Bawah**	**Batas Atas**	**Batas Bawah**
0.584	1.552	0.515	1.759
1.08	3.439	0.931	3.988
0.971	3.208	0.833	3.739

Sumber: Hasil Pengolahan *Software* SAS 9.4

Berdasarkan Tabel 10 terlihat bahwa pada saat data mengalami masalah overdispersi maka selang kepercayaan dari rasio *odds* akan menjadi pendek dibandingkan dengan model regresi logistik multinomial terkoreksi. Hal ini menunjukan bahwa hasil dari pemodelan dengam regresi logistik multinomial terkoreksi melalui penanganan overdispersi merupakan hasil dari pemodelan regresi logistik yang sebenarnya meskipun tidak terdapat variabel-variabel prediktor yang mempengaruhi variabel respon dalam kasus ini adalah jenis kelamin ($\boldsymbol{X_1}$) dan prilaku merokok orang tua ($\boldsymbol{X_2}$) tidak mempengaruhi prilaku merokok mahasiswa unisba.

4. KESIMPULAN

Hubungan fungsional dari variabel respon berbentuk politokomus dengan variabel prediktor yang dimodelkan melalui regresi logistik multinomial adalah:

$$log\left(\frac{\pi_1}{\pi_0}\right) = -0.4387 + 0.2481\, X_2(2)$$

$$log\left(\frac{\pi_2}{\pi_0}\right) = 2.0740 + 0.2481\, X_2(2)$$

Dari model regresi logistik multinomial ini diperoleh hasil perbandingan antara nilai chi-kuadrat Pearson dan Devians terhadap derajat bebasnya lebih dari 1 (satu) sehingga dapat diidentifikasikan bahwa pada pengamatan ini terjadi peristiwa overdispersi. Tindak lanjut untuk mengatasi masalah overdispersi dalam hal ini adalah dengan mengoreksi matriks varians kovariansnya. Pemodelan regresi logistik multinomial terkoreksi adalah sebagai berikut:

$$log\left(\frac{\pi_1}{\pi_0}\right) = -0.4387$$

$$log\left(\frac{\pi_2}{\pi_0}\right) = 2.0740$$

Dari model regresi logistik multinomial biasa dan dengan model regresi logistik multinomial terkoreksi dapat disimpulkan bahwa variabel-variabel prediktor yang dianggap berarti dalam kedua pemodelan tersebut berbeda. Perbedaan lainnya terdapat pada nilai galat baku model regresi logistik multinomial biasa lebih kecil dari yang seharusnya dengan kata lain *underestimate* dibandingkan dengan model regresi logistik multinomial terkoreksi, dan selang kepercayaan untuk rasio *odds* menjadi pendek dibandingkan dengan model regresi logistik multinomial terkoreksi. Hal ini menunjukan bahwa hasil dari pemodelan dengan regresi logistik multinomial terkoreksi merupakan hasil dari pemodelan regresi logistik yang sebenarnya meskipun tidak terdapat variabel-variabel prediktor yang mempengaruhi variabel respon dalam kasus ini adalah Jenis Kelamin ($\boldsymbol{X_1}$) dan Prilaku Merokok Orang Tua ($\boldsymbol{X_2}$) tidak mempengaruhi Prilaku Merokok Mahasiswa Unisba.

DAFTAR PUSTAKA

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. (Second Edition). New York: Wiley.

Collet, D. (1991). *Modeling Binary Data*. London: Chapman and Hall.

Cox, D. R., and Snell, E. J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Czepiel, S. A. (2011). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, (Online), (http://czep.net/stat/mlelr.pdf).

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. (Second Edition). New York: Capman and Hall.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

McCullagh, P., and J.A. Nelder (1989*). Generalized Linear Models*. (Second Edition). New York: Capman and Hall.

Rosiana, Dewi & Halimah, L. (2013). *Metode Intervensi Guna Menurunkan Intensi Merokok Pada Perempuan Perokok.* Makalah dipresentasikan dalam Seminar Nasional Proposal Penelitian (SNaPP) 2014, Fakultas Psikologi, Universitas Islam Bandung, Bandung.