PERBANDINGAN MODEL REGRESI POISSON DAN MODEL REGRESI BINOMIAL NEGATIF[1]

Kismiantini
Jurusan Pendidikan Matematika
FMIPA Universitas Negeri Yogyakarta

Abstrak

Dalam menganalisis hubungan antara beberapa peubah, terdapat sejumlah fenomena dimana peubah responnya bukan lagi kontinu melainkan berbentuk diskret. Fenomena peubah respon berbentuk diskret dengan data berupa cacahan biasanya dianalisis dengan regresi Poisson. Permasalahan yang sering muncul dari regresi Poisson adalah overdispersi (ragam melebihi rata-ratanya), untuk menanganinya dapat digunakan teknik regresi binomial negatif. Hipotesis parameter dispersi sama dengan nol atau tidak dapat digunakan untuk mengetahui model yang lebih baik diantara model regresi Poisson dan model regresi binomial negatif.

Kata kunci : Data cacahan, regresi Poisson, regresi binomial negatif

PENDAHULUAN

Seringkali penelitian mengkaji hubungan antara peubah respon (atau peubah tak bebas) dengan peubah bebas, dengan peubah respon dapat berupa kontinu maupun diskret. Hubungan fungsional antara peubah respon dengan peubah bebas dapat dijelaskan oleh teknik analisis regresi (Kutner *et al*., 2005). Analisis regresi klasik mengasumsikan bahwa peubah respon merupakan peubah kontinu dan mengikuti distribusi normal. Apabila peubah respon tidak lagi kontinu melainkan diskret maka analisis ini tidak dapat digunakan.

Salah satu fenomena dimana peubah responnya diskret adalah fenomena banyaknya kejadian yang jarang terjadi. Misalnya banyaknya kecelakaan mobil setiap bulan, banyaknya hujan badai setiap tahun, banyaknya kebakaran hutan setiap tahun, banyaknya barang yang cacat dalam suatu produksi tertentu. Data yang diperoleh berupa cacahan. Model regresi yang dapat digunakan untuk menjelaskan hubungan antara peubah bebas dengan peubah respon berupa cacahan adalah regresi Poisson dan regresi binomial negatif (Park, 2005). Regresi binomial negatif sering digunakan untuk mengatasi masalah overdispersi pada regresi Poisson (Berk & MacDonald, 2007). Overdispersi terjadi ketika ragam melebihi rataan pada kasus Poisson.

[1] Makalah ini disampaikan pada Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA yang diselenggarakan oleh FMIPA Universitas Negeri Yogyakarta pada tanggal 30 Mei 2008

PEMBAHASAN

Data cacahan merupakan data yang sering dijumpai pada penelitian kriminologi, kesehatan maupun biologi. Ketika peubah respon berupa cacahan, sangat umum untuk menggunakan regresi Poisson (kasus khusus dari model linear terampat). Masalah yang sering dihadapi dalam regresi Poisson adalah overdispersi, hal ini disebabkan diantaranya peubah bebas yang tidak termuat dalam model, sehingga masih dimungkinkan adanya keragaman dari peubah respon yang disebabkan oleh peubah lain.

Regresi Poisson

Model regresi untuk data cacahan diantaranya adalah model regresi Poisson. Pada model regresi ini, peubah respon berupa data cacahan yang mengikuti distribusi Poisson. Distribusi Poisson sering digunakan untuk kejadian-kejadian yang jarang terjadi dengan data berupa cacahan yang mempunyai nilai non negatif.

Peubah acak Y dikatakan berdistribusi Poisson dengan parameter μ dengan $y = 0, 1, 2, ...$ bila fungsi peluangnya adalah

$$p(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \mu > 0 \qquad (1)$$

Distribusi Poisson ini mempunyai rata-rata dan ragam berikut

$$E(Y) = Var(Y) = \mu \qquad (2)$$

Karena rata-rata sama dengan ragamnya, maka sembarang faktor akan berpengaruh terhadap lainnya, sehingga asumsi homogenitas tidak harus dipenuhi pada data Poisson (Rodriquez, 2001).

Selanjutnya untuk membangun model regresi Poisson, dimisalkan sampel acak $Y_i \sim Poisson(\mu_i)$, $i = 1, 2, ..., n$ dan rata-rata μ_i bergantung pada vektor peubah bebas (peubah penjelas) $\mathbf{x}_i$ dan vektor koefisien regresi $\boldsymbol{\beta}$, yaitu

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \qquad (3)$$

Tetapi model ini memiliki kelemahan yaitu prediktor linear ($\mathbf{x}_i^T \boldsymbol{\beta}$) dapat diasumsikan dengan sebarang nilai, padahal rata-rata Poisson merupakan harapan cacahan yang nilainya harus non negatif. Untuk mengatasi permasalahan ini digunakan log rata-rata dengan model linear sebagai berikut

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \qquad (4)$$

Regresi Binomial Negatif

Jika model regresi Poisson tidak *fit* dengan data cacahan dan ragam peubah respon melebihi rata-ratanya yang sering disebut sebagai overdispersi (hal ini dapat dilihat dari plot sisaan dengan prediktor linear dengan titik-titik berpola menyebar) maka model regresi binomial negatif dapat digunakan sebagai alternatif untuk mengatasi permasalahan tersebut (Cameron & Trivedi, 1999).

Langkah pertama dalam membangun model regresi binomial negatif adalah dengan mengasumsikan bahwa peubah respon Y_i merupakan peubah acak yang saling bebas dan identik yaitu $Y_i|\lambda_i \overset{iid}{\sim} Poisson(\lambda_i)$, dengan fungsi peluang $f(y_i|\lambda_i)=\dfrac{e^{-\lambda_i}\lambda_i^{\,y_i}}{y_i!}$, $y_i = 0,1,2,\ldots$ dan $\lambda_i > 0$.

Langkah kedua adalah dengan mengasumsikan bahwa $\lambda_i \sim Gamma(\alpha,\beta)$ dengan rata-rata $\alpha\beta$, ragam $\alpha\beta^2$ dan fungsi padat peluang berikut

$$m(\lambda_i)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda_i^{\alpha-1}\exp(-\lambda_i/\beta), & \lambda_i > 0\\ 0, & \lambda_i \text{ yang lain}\end{cases} \tag{5}$$

Maka diperoleh fungsi bersama adalah

$$f(y_i,\lambda_i)=\frac{e^{-\lambda_i}\lambda_i^{\,y_i}}{y_i!}\frac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda_i^{\alpha-1}\exp(-\lambda_i/\beta)\ ,\ y_i = 0,1,\ldots;\lambda_i > 0 \tag{6}$$

Selanjutnya diperoleh fungsi marjinal dapat diperoleh merupakan fungsi peluang dari distribusi binomial negatif sebagai berikut

$$\begin{aligned} m(y_i) &= \int_0^{\infty} f(y_i,\lambda_i)d\lambda_i \\ &= \frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha)y_i!}\left(\frac{\beta}{1+\beta}\right)^{y_i}\left(\frac{1}{1+\beta}\right)^{\alpha},\ y_i = 0,1,2,\ldots \end{aligned} \tag{7}$$

Distribusi binomial negatif dengan fungsi peluang pada (7) ini mempunyai rata-rata

$$E(Y_i) = E\left[E(Y_i|\lambda)\right] = E(\lambda) = \alpha\beta$$

dan ragam

$$\begin{aligned} Var(Y_i) &= E\left[Var(Y_i|\lambda)\right] + Var\left[E(Y_i|\lambda)\right] \\ &= Var(\lambda) + E(\lambda) \\ &= \alpha\beta + \alpha\beta^2 \end{aligned}$$

Selanjutnya dalam membangun model regresi binomial negatif, diasumsikan bahwa $\mu_i = \alpha\beta$ dan $\kappa = 1/\alpha$, sehingga $E(Y_i) = \mu_i$ dan $Var(Y_i) = \mu_i + \kappa\mu_i^2$, ragam ini merupakan fungsi kuadratik yang mengakomodasi parameter overdispersi κ >0. Sehingga distribusi Y_i menjadi

$$m(y_i) = \frac{\Gamma(\kappa^{-1} + y_i)}{\Gamma(\kappa^{-1})y_i!}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\left(\frac{1}{1+\kappa\mu_i}\right)^{1/\kappa} \quad (8)$$

Jika $\kappa \to 0$ maka distribusi ini mendekati Poisson(μ). Binomial negatif mampu mengakomodasi overdispersi ($\kappa > 0$) tetapi tidak underdispersi ($\kappa < 1$) pada model Poisson. Secara umum didefinisikan bahwa peubah respon merupakan peubah acak berdistribusi binomial negatif dengan parameter μ_i dan κ berikut

$$Y_i \sim BN(\mu_i, \kappa) \quad (9)$$

dan fungsi hubung log yaitu

$$\log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \quad (10)$$

dengan $\mathbf{x}_i$ vektor peubah bebas (peubah penjelas) dan $\boldsymbol{\beta}$ vektor koefisien regresi.

Perbandingan Model Regresi Poisson dan Model Regresi Binomial Negatif

Model regresi Poisson dan model regresi binomial negatif termasuk dalam model linear terampat (*Generalized Linear Model*). Ada tiga komponen utama dalam *GLM* yaitu (McCullagh & Nelder, 1989):

1. Komponen acak, yaitu komponen dari *Y* yang bebas dan fungsi padat peluang atau fungsi peluang *Y* termasuk dalam keluarga sebaran eksponensial dengan $E(Y) = \mu$.
2. Komponen sistematik, yaitu $x_1, x_2, \ldots, x_p$ yang menghasilkan penduga linear η dimana $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.
3. Fungsi penghubung (*link function*) $g(.)$, yang menggambarkan hubungan antara penduga linear η dengan nilai tengah μ. ($\eta = g(\mu)$).

Berikut adalah tabel yang menjelaskan tiga komponen utama *GLM* pada model regresi Poisson dan model regresi binomial negatif.

Tabel 1. Komponen *GLM*

Model Regresi	Komponen acak	Komponen Sistematik	Fungsi hubung
Poisson	$Y_i \overset{iid}{\sim} Poisson(\mu_i)$	$\mathbf{x}_i^T\boldsymbol{\beta}$	log
Binomial Negatif	$Y_i \overset{iid}{\sim} BN(\mu_i, \kappa)$	$\mathbf{x}_i^T\boldsymbol{\beta}$	log

Model regresi binomial negatif memuat parameter dispersi κ yang mengakomodasi overdispersi. Menurut Long (1997), uji *likelihood ratio* dapat digunakan untuk memeriksa hipotesis nol tidak ada overdispersi, yaitu hipotesis $H_0 : \kappa = 0$ lawan $H_1 : \kappa \neq 0$. Statistik uji yang digunakan $LR = 2(\ln L_{BN} - \ln L_{Poisson}) \sim \chi^2_{(1)}$. Jika H_0 ditolak maka terjadi overdispersi dengan kata lain model regresi binomial negatif lebih baik digunakan daripada model regresi Poisson.

Tabel 2. Perbandingan Model Regresi Poisson dan Model Regresi Binomial Negatif

	Model Regresi Poisson	Model Regresi Binomial Negatif
Peubah respon	$Y_i \overset{iid}{\sim} Poisson(\mu_i)$	$Y_i \overset{iid}{\sim} BN(\mu_i, \kappa)$
Rata-rata dan ragam dari peubah respon Y_i	$E(Y_i) = Var(Y_i) = \mu_i$	$E(Y_i) = \mu_i$, $Var(Y_i) = \mu_i + \kappa\mu_i^2$
Parameter dispersi (κ)	Tidak ada	Ada
Hipotesis $H_0 : \kappa = 0$ $H_1 : \kappa \neq 0$	H_0 diterima maka model regresi Poisson lebih baik daripada model regresi binomial negatif.	H_0 ditolak maka model regresi binomial negatif lebih baik daripada model Poisson.

Tabel 2 menjelaskan secara garis besar perbedaan dari model regresi Poisson dan model regresi binomial negatif, walaupun kedua model ini sama-sama digunakan untuk memodelkan data berupa cacahan.

Ilustrasi

Data yang digunakan dalam makalah ini adalah dua data sekunder. Data pertama diambil dari Gail (1978) dalam Stokes *et al*. (2000) yaitu tentang penderita melanoma pada pria berkulit putih dari tahun 1969-1971 di dua wilayah. Data ini berupa banyaknya penderita melanoma (sebagai peubah respon), wilayah, kelompok usia (sebagai peubah bebas), dan banyaknya penduduk yang beresiko pada wilayah dan kelompok usia tertentu. Input data melanoma pada SAS versi 9.1,

```
data melanoma;
input age $ region $ cases total;
ltotal=log(total);
datalines;
35-44 south  75  220407
45-54 south  68  198119
55-64 south  63  134084
65-74 south  45   70708
75+   south  27   34233
<35   south  64 1074246
35-44 north  76  564535
45-54 north  98  592983
```

```
55-64 north 104  450740
65-74 north  63  270908
75+   north  80  161850
<35   north  61 2880262
;
proc genmod data=melanoma order=data;
class age region;
model cases = age region
/ dist=poisson link=log offset=ltotal;
run;
```

Berikut output SAS versi 9.1 dari data melanoma dengan model regresi Poisson.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	6.2149	1.2430
Scaled Deviance	5	6.2149	1.2430
Pearson Chi-Square	5	6.1151	1.2230
Scaled Pearson X2	5	6.1151	1.2230
Log Likelihood		2694.9262	

Selanjutnya untuk mendapatkan *Likelihood Ratio* dari model regresi binomial negatif pada data melanoma adalah dengan mengganti distribusi pada input data, yaitu semula `dist=poisson` menjadi `dist=negbin`, sehingga diperoleh output berikut :

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	20.0285	4.0057
Scaled Deviance	5	20.0285	4.0057
Pearson Chi-Square	5	18.4675	3.6935
Scaled Pearson X2	5	18.4675	3.6935
Log Likelihood		2697.4922	

Berdasarkan kedua output SAS ini diperoleh bahwa $LR = 2\times(2697.4922 - 2694.9262) = 2.566$. Bila dipilih taraf nyata $\alpha = 0.05$, $\chi^2_{0.05(1)} = 3{,}841$, maka $LR < 3.841$ sehingga H_0 diterima ($\kappa = 0$), yang berarti tidak terjadi overdispersi atau dengan kata lain model regresi Poisson lebih baik digunakan daripada model regresi binomial negatif.

Data kedua diambil dari LaVange *et al*. (1994) tentang infeksi pernapasan pendek. Data ini berupa banyaknya penderita pernapasan pendek setiap tahun (sebagai peubah respon), banyaknya perokok pasif dalam rumahtangga, status sosial ekonomi, *crowding*, ras dan kelompok usia (sebagai peubah bebas), dengan jumlah pengamatan ada sebanyak 284 anak. Dalam kasus ini, sangat masuk akal bahwa anak yang terserang batuk kebanyakan disebabkan oleh hal lain, sehingga dimungkinkan tambahan keragaman atau terjadi overdispersi pada data ini. Input data infeksi pernapasan pendek pada SAS versi 9.1,

```
data lri;
input id count risk passive crowding ses agegroup race @@;
logrisk =log(risk/52);
datalines;
```

```
1 0 42 1 0 2 2 0 96 1 41 1 0 1 2 0 191 0 44 1 0 0 2 0
2 0 43 1 0 0 2 0 97 1 26 1 1 2 2 0 192 0 45 0 0 0 2 1
3 0 41 1 0 1 2 0 98 0 36 0 0 0 2 0 193 0 42 0 0 0 2 0
4 1 36 0 1 0 2 0 99 0 34 0 0 0 2 0 194 1 31 0 0 0 2 1
. . .
92 1 3 1 0 1 3 1 187 0 42 0 0 0 2 0 282 1 32 1 0 2 2 0
93 0 26 1 0 0 2 1 188 0 38 0 0 0 2 0 283 0 22 1 1 2 2 1
94 0 35 1 0 0 2 0 189 0 36 1 0 0 2 0 284 0 35 0 0 0 2 1
95 3 37 1 0 0 2 0 190 0 39 0 1 0 2 0
;
proc genmod data=lri;
class ses id race agegroup;
model count = passive crowding ses race agegroup /
dist=negbin offset=logrisk type3;
run;
```

Berikut output SAS versi 9.1 dari data infeksi pernapasan pendek dengan model regresi Poisson.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	276	408.1549	1.4788
Scaled Deviance	276	408.1549	1.4788
Pearson Chi-Square	276	495.4493	1.7951
Scaled Pearson X2	276	495.4493	1.7951
Log Likelihood		-260.4117	

Berdasarkan output ini, diperoleh nilai 1.4788 untuk deviance/df dan 1.7951 untuk Perason/df, nilai ini mengindikasikan terjadinya overdispersi. Selanjutnya dengan cara yang sama pada data pertama, untuk mendapatkan *Likelihood Ratio* dari model regresi binomial negatif pada data infeksi pernapasan pendek ini adalah dengan mengganti distribusi pada input data, yaitu semula `dist=poisson` menjadi `dist=negbin`, sehingga diperoleh output berikut :

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	276	256.9688	0.9310
Scaled Deviance	276	256.9688	0.9310
Pearson Chi-Square	276	298.2410	1.0806
Scaled Pearson X2	276	298.2410	1.0806
Log Likelihood		-242.2932	

Berdasarkan output ini, nilai 0.9310 untuk deviance/df dan 1.0806 untuk Perason/df, nilai ini mengindikasikan tidak terjadinya overdispersi. Dari kedua output SAS ini diperoleh bahwa $LR = 2\times(-242.2932 - (-260.4117) = 18.1185$. Bila dipilih taraf nyata α

= 0.05, $\chi^2_{0.05(1)} = 3{,}841$, maka $LR > 3.841$ sehingga H_0 ditolak ($\kappa \neq 0$), yang berarti terjadi overdispersi atau dengan kata lain model regresi binomial negatif lebih baik digunakan daripada model regresi Poisson.

PENUTUP

Model regresi binomial negatif memiliki parameter dispersi κ yang mampu mengakomodasi permasalahan overdispersi pada model regresi Poisson. Bila hipotesis nol tidak terjadi overdispersi diterima maka model regresi Poisson lebih baik daripada model regresi binomial negatif dan sebaliknya bila hipotesis nol tidak terjadi overdispersi ditolak maka model regresi binomial negatif lebih baik digunakan daripada model regresi Poisson. Bila nilai deviance/df dan Pearson/df pada *goodness of fit* mendekati satu maka tidak mengindikasikan terjadinya overdispersi.

DAFTAR PUSTAKA

Berk, D. & MacDonald, J. 2007. Overdispersion and Poisson regression. Department of Statistics, Department of Criminology, University of Pennsylvania.

Cameron, A.C. & Trivedi, P.K. 1999. Essentials of count data regression. A Companion to Theoretical Econometrics, Blackwell.

Gail, M. 1978. The analysis of heterogeneity for indirect standardized mortality ratios. *Journal of the Royal Statistical Society A* 141: 224-234.

Kutner, M.H., Nachtsheim, C.J., Neter, J. & Li, W. 2005. *Applied Linear Statistical Models*. New York: McGraw-Hill.

Lavange, L.M., Keyes, L.L., Koch, G.G. & Margolis, P.E. 1994. Application sample survey methods for modelling ratios to incidence densities. *Statistics in Medicine* 13: 343-355.

Long, J.S. 1997. Regression models for categorical and limit dependent variables. *Advanced Quantitative Techniques in the Social Sciences*. Sage Publications.

McCullagh, P. & Nelder, J.A. 1989. *Generalized Linear Models*. London: Chapman & Hall.

Park, H.M. 2005. Regression models for event count data using SAS, STATA, and LIMDEP. Indiana: The Trustees of Indiana University.

Rodriguez, G. 2001. Poisson models for count data. [terhubung berkala] http://data.pricenton.edu/wws509/notes/c4.pdf [13 Juni 2006].

Stokes, M.E., Davis, C.S. & Koch, G.G. 2000. *Categorical data analysis using the SAS® system second edition*. North Carolina: John Wiley & Sons.