Penerapan Hurdle Negative Binomial pada Data Tersensor

Resa Septiani Pontoh, Defi Yusti Faidah.
Departemen Statistika FMIPA Universitas Padjadjaran
resa.pontoh@gmail.com

Abstrak— Pada kasus data tercacah, model regresi poisson biasa digunakan untuk memodelkan jenis data tersebut. Namun, apabila terdapat overdispersi didalamnya, maka model regresi poisson menjadi kurang tepat menggambarkan kondisi data yang ada. Jika hal ini terjadi, maka *Negative binomial Model* biasanya digunakan sebagai salah satu alternatif solusi. Untuk kasus yang sangat jarang terjadi, maka pada variabel responnya akan ditemukan nilai nol yang berlebih. Hal ini tentunya menjadi indikator yang sangat kuat akan terjadinya kejadian overdispersi. Kejadian nilai nol berlebih tersebut diartikan sebagai data tersensor yang kemudian akan dimodelkan. Pada penelitian ini, data tersensor tersebut akan dimodelkan dengan pendekatan *Hurdle Negative binomial*. Maximum Likelihood akan digunakan untuk menaksir parameter regresi yang ada.

Kata kunci: Hurdle Negative Binomial, data tersensor

I. PENDAHULUAN

Pada kasus data tercacah, model regresi poisson biasa digunakan untuk memodelkan jenis data tersebut (Agresti, 2002). Namun, apabila terdapat overdispersi didalamnya, maka model regresi poisson menjadi kurang tepat menggambarkan kondisi data yang ada. Jika hal ini terjadi, maka *Negative binomial Model* biasanya digunakan sebagai salah satu alternatif solusi. Untuk kasus yang sangat jarang terjadi, maka pada variabel responnya akan ditemukan nilai nol yang berlebih. Hal ini tentunya menjadi indikator yang sangat kuat akan terjadinya kejadian overdispersi. Kejadian nilai nol berlebih tersebut diartikan sebagai data tersensor yang kemudian akan dimodelkan.

Untuk nilai nol yang berlebih ini, beberapasolusi dapat digunakan seperti mengaplikasikan zero inflated model seperti zero inflated poisson, zero inflated generalized poisson dan zero inflated negative binomial. Bahkan, untuk beberapa kasus, negative binomial model menghasilkan model yang lebih fit dibandingkan dengan sero inflated model (Pontoh, 2014). *Zero Inflated Poisson* model mempunyai ciri khas pada dua jenis pemodelan didalamnya yaitu memodelkan observasi dengan nilai 0 menggunakan model logistik dan memodelkan observasi dengan nilai positif menggunakan model poisson. Model *hurdle* pada dasarnya hampir mirip dengan model ZIP yang melakukan dua pemodelan. Namun, pada pemodelan kedua, model *Hurdle Poisson* menggunakan *Truncated Poisson* untuk data yang tidak bernilai 0 dan positif. Penaksiran parameter pada kedua model ini menggunakan metode maksimum likelihood. Keunggulan dari model *Hurdle Poisson* adalah kedua model didalamnya dapat dilakukan penaksiran parameter secara terpisah atau dengan kata lain dimaksimumkan secara terpisah sehingga diharapkan dapat lebih mudah dalam penginterpretasiannya (Cantoni dan Zedini, 2010). Selain Hurdle poisson, *Hurdle Negative Binomial* pun populer digunakan yaitu dengan memodelkan nilai bukan nol menggunakan *truncated binomial negative*. Pada penelitian ini, peneliti mengaplikasikan Hurdle Binomial Negative Model untuk digunakan pada pemodelan faktor-faktor yang berpengaruh terhadap kejadian kasus difteri di Provinsi Jawa Barat.

ISBN. 978-602-73403-0-5

II. METODE PENELITIAN

Pada penelitian ini, Hudle Negative Binomial akan diaplikasikan untuk memodelkan faktor-faktor yang berpengaruh terhadap kejadian kasus difteri di Provinsi Jawa Barat. Metode penelitian ini khusus bagi makalah hasil penelitian. Bagian ini memuat rancangan, bahan, subjek penelitian, prosedur, instrumen, dan teknik analisis data, serta hal-hal yang terkait dengan cara-cara penelitian.

A. Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data *cross sectional* yaitu data sekunder yang diperoleh dari Dinas Kesehatan Provinsi Jawa Barat pada untuk 26 kabupaten/kota di Provinsi Jawa Barat serta dari Badan Pusat Statistik Jawa Barat. Pada data ini, nilai 0 untuk kejadian TB berada di atas nilai 50% sehingga dapat disimpulkan terjadi *excess zeros* yang mengakibatkan over dispersi. Variabel dependen atau variabel respon yang dijadikan studi kasus berupa banyak kasus penyakit difteri yang dialami penduduk dan terdaftar di pusat kesehatan di setiap kabupaten/kota Provinsi Jawa Barat (Y) dengan Variabel Independen atau variabel prediktor yang digunakan sebagai berikut :

a. Persentase balita gizi buruk (X_1)
b. Jumlah cakupan DPT1+HB1 (dalam ribuan) (X_2)
c. Jumlah cakupan DPT3 + HB3 (dalam ribuan) (X_3)
d. Persentase rumah sehat (X_4)
e. Rata-rata kepadatan tiap rumah (X_5)
f. Persentase penduduk miskin (X_6)
g. Pola hidup bersih dan sehat (X_7)
h. Rumah Sehat (X_7)
i. Kepadatan (X_8)
j. Angka Partisipasi Kasar (X_9)
k. Air Bersih (X_{10})
l. Angka Melek Huruf (X_{11})
m. Pendapatan Perkapita (X_{12})

Tabel 1 Distribusi Frekuensi Kasus Difteri di Provinsi Jawa Barat Tahun 2013

No	Kabupaten/Kota	Frekuensi	No	Kabupaten/Kota	Frekuensi
1	Kab. Bogor	1	15	Kab. Karawang	3
2	Kab. Sukabumi	1	16	Kab. Bekasi	8
3	Kab. Cianjur	6	17	Kab. Bandung Barat	0
4	Kab. Bandung	0	18	Kab. Pangandaran	0
5	Kab. Garut	0	19	Kota Bogor	0
6	Kab. Tasikmalaya	2	20	Kota Sukabumi	0
7	Kab. Ciamis	0	21	Kota Bandung	5
8	Kab. Kuningan	0	22	Kota Cirebon	0
9	Kab. Cirebon	0	23	Kota Bekasi	1
10	Kab. Majalengka	0	24	Kota Depok	1
11	Kab. Sumedang	0	25	Kota Cimahi	3
12	Kab. Indramayu	0	26	Kota Tasikmalaya	0
13	Kab. Subang	0	27	Kota Banjar	0
14	Kab. Purwakarta	1			

Seperti yang terlihat pada Tabel 1, bahwa frekuensi banyaknya kasus difteri untuk tiap Kabupaten/Kota di Jawa Barat banyak yang bernilai nol (*excess zeros*). Terdapat 15 Kabupaten/Kota yang frekuensinya 0 dari 27 Kabupaten/Kota di Jawa Barat pada tahun 2013.

Kejadian difteri akan menghasilkan data yang berbentuk diskrit non negatif. Pengamatan dengan variabel respon berbentuk data frekuensi/ cacahan (*count*) dengan kejadian yang sangat langka, tetapi pasti terjadi pada selang waktu tertentu. Model regresi yang dapat digunakan jika variabel respon berupa data *count* adalah model regresi *Poisson* (Agresti, 2002). Asumsi penting pada analisis regresi *Poisson* adalah nilai ragam harus sama dengan nilai rata-ratanya yang disebut dengan *equidispersion* (Famoye *et al*, 2004), jika berbeda maka akanterjadi overdispersi atau underdispersi. Banyaknya kasus difteri di Jawa Barat pada data profil Dinas Kesehatan Jabwa Barat menunjukkan ciri-ciri terjadinya overdispersi akibat banyaknya hasil observasi yang bernilai nol, sehingga model regresi *Hurdle Negative Binomial* (HNB) merupakan salah satu alternatif untuk memodelkan data difteri di Provinsi Jawa Barat.

B. Model Regresi Hurdle Negative Binomial

Analisis faktor yang mempengaruhi banyaknya kasus difteri di Jawa Barat pada data profil Dinas Kesehatan Jabar menunjukkan ciri-ciri terjadinya overdispersi akibat banyaknya hasil observasi yang bernilai nol (*excess zeros*), selain itu hasil pada regresi HNB lebih mudah diinterpretasikan, sehingga metode regresi *Hurdle Negative Binomial* (HNB) merupakan salah satu alternatif untuk memodelkan banyaknya kasus difteri di Provinsi Jawa Barat.

Di misalkan *Yi* adalah variabel random yang diskrit dengan *i* adalah bilangan bulat non negatif (*i* = 1,2,..., *n*) dan *Yi* merupakan variabel respon dari model regresi HNB, maka nilai dari variabel respon tersebut terjadi dalam dua keadaan. Keadaan pertama disebut *zero state* dan menghasilkan hanya pengamatan bernilai nol, sementara keadaan kedua disebut *negative binomial state* yang memiliki sebaran *Binomial Negative*. Fungsi peluang dari *Yi* adalah (Saffari *et al*, 2012):

$$P(Y_i = y_i)= \begin{cases} p_i & y_i = 0 \\ (1-pi)\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\frac{(1+\alpha\mu_i)^{-\alpha^{-1}}\alpha^{y_i}\mu_i^{y_i}}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}} & y_i > 0 \end{cases}, \quad (1)$$

dimana :

$\tau(.)$: parameter gamma

y_i : variabel random independen yang diskrit

p_i : probabilitas keadaan zero state menghasilkan hanya observasi bernilai nol

$(1-p_i)$: probabilitas keadaan Negative Binomial state berdistribusi binomial negatif

μ : mean dari distribusi binomial negatif

α : parameter dispersi

di mana *i* = 1,2,3,.. ,n ; 0 < *pi* <1 ; $\mu i \geq 0$ dan $\alpha \geq 0$

Diasumsikan bahwa μ_i dan pi bergantung pada vektor dari variabel prediktor *xi* yang dapat didefinisikan sebagai berikut :

$$p_i = \frac{e^{x_i^T\delta}}{1+e^{x_i^T\delta}}, \quad (2)$$

$$(1-p_i) = \frac{1}{(1+e^{x_i^T\delta})}, \quad (3)$$

$$\mu_i = e^{x_i^T\beta}. \quad (4)$$

Model regresi HNB dapat dinyatakan sebagai berikut :

Model untuk truncated negative binomial dengan log link adalah:

$$\ln(\mu_i) = \hat{\beta}_0 + \sum_{i=1}^{p}\hat{\beta}_j x_{ij,}\, i = 1,\dots,n \text{ dan } j = 1,\dots,p \quad (5)$$

Model untuk hurdle (binomial dengan logit link)

$$\text{logit } p_i = \hat{\delta}_0 + \sum_{i=1}^{p}\hat{\delta}_j x_{ij,}\, i = 1,\dots,n \text{ dan } j = 1,\dots,p \quad (6)$$

dimana :

p : jumlah variabel prediktor

n : jumlah pengamatan

β : parameter model regresi HNB yang diestimasi

δ : parameter model regresi HNB yang diestimasi

C. Pengujian Signifikansi Parameter Regresi HNB

Uji Simultan

Pengujian signifikansi parameter secara simultan didasarkan pada *Likelihood Rasio Test* dengan statistik uji G.

$H_0 : \beta_{1=} \beta_{1=} ... = \delta_1 = \delta_2 = \cdots = \delta_p = 0$ (model regresi HNB tidak dapat digunakan sebagai model)

H_1: paling sedikit ada satu $\beta_j \neq 0$ atau $\delta_j \neq 0$ dimana j=1,2,3,...p

(model regresi HNB dapat digunakan sebagai model)

Statistik uji pada LRT adalah G^2 yang dirumuskan sebagai berikut:

$$G^2 = -2ln\left[\frac{L_0}{L_1}\right] = -2(ln\, L_0 - \ln L_1) \quad , \qquad (7)$$

dimana :

L_0 : likelihood tanpa variabel bebas (model konstan)

L_1 : likelihood dengan variabel bebas (model penuh)

p : selisih derajat bebas pada model penuh dan model konstan

α : tingkat signifikansi

Kriteria uji:

Tolak H_0 jika $G^2 \geq X^2_{\alpha,2p}$ dan terima untuk hal lainnya.

Uji Parsial

Jika uji simultan memberikan hasil penolakan terhadap H_0 yang berarti model HNB dapat digunakan sebagai model, maka dilanjutkan ke uji parsial. Pengujian signifikansi parameter secara parsial digunakan untuk mengetahui pengaruh masing-masing variabel prediktor terhadap variabel respon. Hasil pengujian secara parsial berdasarkan statistik uji Wald.

A. *Uji signifikansi parameter model* $log(\mu) = \boldsymbol{X\beta}$

H_0: $\beta_j = 0$ (koefisien tidak signifikan)

H_1: $\beta_j \neq 0$ (koefisien signifikan)

Statistik uji Wald dirumuskan sebagai berikut :

$$W_j = \left(\frac{\widehat{\beta_j}}{SE\left(\widehat{\beta_j}\right)}\right)^2 \quad , \qquad (8)$$

dimana :

$\widehat{\beta_j}$: taksiran koefisien pada model $\log(\mu) = \boldsymbol{X\beta}$ variabel prediktor ke-j

$SE\left(\widehat{\beta_j}\right)$: *standard error* dari taksiran koefisien pada model $logit\ (\omega) = \boldsymbol{X\delta}$ variabel prediktor ke-j

α : tingkat signifikansi

Kriteria uji : Tolak Ho jika $W_j \geq X^2_{\alpha;v}$, terima dalam hal lainnya.

B. Uji signifikansi parameter model $logit\ (\omega) = \boldsymbol{X\delta}$

H_0: $\delta_j = 0$ (koefisien tidak signifikan)

H_1: $\delta_j \neq 0$ (koefisien signifikan)

Statistik uji Wald dirumuskan sebagai berikut :

$$W_j = \left(\frac{\widehat{\delta_j}}{SE\left(\widehat{\delta_j}\right)}\right)^2 \quad , \qquad (9)$$

dimana :

$\widehat{\delta_j}$: taksiran koefisien model $logit\ (\omega) = \boldsymbol{X\delta}$ variabel prediktor ke-j

$SE\left(\widehat{\delta_j}\right)$: *standard error* dari taksiran koefisien pada model $logit\ (\omega) = \boldsymbol{X\delta}$ variabel prediktor ke-j

α : tingkat signifikansi

Kriteria uji : Tolak Ho jika $W_j \geq X^2_{\alpha;v}$, terima dalam hal lainnya.

III. HASIL DAN PEMBAHASAN

Setelah dilakukan analisis data,dengan juga mengatasi masalah multikoleniaritas, maka diperoleh hasil seperti pada tabel 2.

TABEL 2. OUTPUT HURDLE NEGATIVE BINOMIAL MODEL

```
Count model coefficients (truncated negbin with log link):
              Estimate Std. Error   z value Pr(>|z|)
(Intercept) -5.912e+01  2.667e+01    -2.217   0.0266 *
x2           5.982e-02  3.234e-02     1.850   0.0643 .
x7          -1.402e-01  8.136e-02    -1.724   0.0848 .
x10         -1.474e-01  7.426e-02    -1.985   0.0472 *
x11          8.479e-01  5.078e-01     1.670   0.0950 .
Log(theta)   2.672e+01  7.382e-04 36200.901   <2e-16 ***
Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -68.04536   30.46035  -2.234   0.0255 *
x2            0.07000    0.02833   2.471   0.0135 *
x5            3.12245    1.50123   2.080   0.0375 *
x6            0.42514    0.22849   1.861   0.0628 .
x11           0.50702    0.25824   1.963   0.0496 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 403226917619.398
Number of iterations in BFGS optimization: 94
Log-likelihood: -44.76
```

Tabel 2 memperlihatkan bahwa uji parsial sudah signifikan dengan α=10%. Hasil penaksiran parameter dari model *hurdle* terdiri dari model logit dan model *truncated poisson*. Pengujian secara serentak model*hurdle* dapat dilihat dari nilai *chi-square* hitung dibandingkan dengan tabel *chi-square.* Nilai *chi-square* hitung adalah 25,768. Hal ini berarti bahwa minimal ada satu parameter yang berpengaruh secara signifikan terhadap model. Kemudian, terlihat bahwa dengan menggunakan $\alpha = 10\%$, faktor-faktor yang mempengaruhi turunnya kejadian difteri di daerah Jawa Barat dengan model tersensor menggunakan truncated negative binomial dan log link adalah sebagai berikut :

1. jumlah cakupan DPT1+HB1sebesar 0.05982 (1.06), dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut telah mendapatkan imunisasi DPT1+HB1, maka kemungkinan anak tersebut tidak terkena difteri 1,07 kali.
1. pola Hidup Bersih dan Sehat (0.87), dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut mempunyai pola hidup bersih, maka kemungkinan anak tersebut terkena difteri 1,15 (1/0.87) kali.
2. air bersih (0.86), dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut tinggal di lingkungan dengan kualitas air yang baik, maka kemungkinan anak tersebut terkena difteri 1,16 (1/0.86) kali.
2. dan angka melek huruf (2.33), dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut melek huruf, maka kemungkinan anak tersebut tidak terkena difteri 2.33 kali.

sedangkan untuk data tersensor atau tidak adanya kejadian difteri, faktor-faktor yang mempengaruhi adalah:

3. Jumlah cakupan DPT1+HB1 (dalam ribuan) (1.07), dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut telah mendapatkan imunisasi DPT1+HB1, maka kemungkinan anak tersebut tidak terkena difteri 1,07 kali.
4. kepadatan rumah (22.7), dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut tinggal di daerah dengan tingkat kepadatan yang lebih baik, maka kemungkinan anak tersebut tidak terkena difteri 22.7 kali.
5. Persentase penduduk miskin (1.53), dapat dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut hidup di lingkungan dengan konsisi air yang lebih bersih, maka kemungkinan anak tersebut tidak terkena difteri 1,07 kali.
6. dan angka melek huruf (1.66)., dapat dikatakan jika sebagin besar anak-anak di kabupaten-kabupaten tersebut hidup di daerah yang tingkat melek hurufnya lebih tinggi, maka kemungkinan anak tersebut tidak terkena difteri 1,66 kali.

IV. SIMPULAN DAN SARAN

Dari hasil tersebut, dapat disimpulkan bahwa Hurdle negative Binomial memberikan wawasan baru terhadap penyelesaian kasus data yang mempunyai nilai nol yang cukup banyak. Terlihat bahwa cakupan DPT1+HB1dan angka melek huruf merupakan faktor serupa untuk kedua model terhadap kejadian difteri. Faktor-faktor yang berpengaruh terhadap difteri mempunyai arah yang positif sehingga dengan meningkatkan kuantitas dan kualitas faktor-faktor yang berpengaruh diharapkan akan menurunkan angka kejadian difteri. Dengan hasil log theta yang signifikan, dapat dikatakan bahwa negative hurdle model memang cocok untuk digunakan. Namun, apakah model ini lebih cocok dibandingkan model lainnya perlu dilakukan . Untuk itu, pada penelitian selanjutnya, peneliti akan membandingkan model ini dengan model-model lainnya.

DAFTAR PUSTAKA

Agresti, A. 2002. Categorical Data nalysis, Second Edition. New York : Jihn Wiley & Sons

Kemenkes RI. 2010-2014. Profil Kesehatan Indonesia 2009-2013

Saffari, Seyed Ehsan, Adnan, Robiah, & Greene, William. (2012). *Hurdle negative binomial regression model with right censored count data*. (Saffari, Seyed Ehsan; Adnan, Robiah; Greene, William. Hurdle negative binomial regression model with right censored count data. "SORT", vol. 36, núm. 2, p. 181-194.) Institut d'Estadística de Catalunya. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

Cantoni, E., & Zedini, A. (January 01, 2011). A robust version of the hurdlemodel. *Journal of Statistical Planning and Inference, 141*, 3, 1214-1223. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.