Pendekatan *Hurdle Poisson* Pada *Excess Zero Data*

Defi Yusti Faidah, Resa Septiani Pontoh
Departemen Statistika FMIPA Universitas Padjadjaran
defi.yusti@unpad.ac.id

***Abstrak*—**Model *HurdlePoisson* digunakan untuk menjelaskan hubungan antara variabel respon yang berupa *count data* dengan variabel prediktor yang berupa kontinu, diskrit atau campuran antara keduanya. Pada beberapa kasus tertentu, variabel respon memiliki nilai nol yang berlebih sehingga menyebabkan terjadinya overdispersi. Oleh karena itu, untuk mengatasi overdispersi dapat digunakan pendekatan *hurdle poisson*. Model *Hurdlepoisson*merupakan kombinasi antara model logit dan *truncated poisson*. Penaksiran parameter dapat dilakukan secara terpisah sehingga interpretasi lebih mudah. Kajian penelitian ini adalah memodelkan jumlah kasus difteri di Provinsi Jawa Barat yang merupakan provinsi dengan penderita difteri terbanyak kelima di Indonesia. Berdasarkan hasil pemodelan *hurdle* dapat diketahui faktor-faktor yang berpengaruh secara signifikan terhadap difteri. Jawa Barat. Diharapkan dengan diketahui faktor-faktor yang berpengaruh dapat dijasikan sebagai masukan kepada pemerintah untuk menekan jumlah penderita difteri di Jawa Barat.

Kata kunci:difteri, excess zeros, hurdle poisson

I. PENDAHULUAN

Pada kasus tertentu, variabel penelitian mengandung *excess zeros data*. *Excess zeros data* menghasilkan nilai nol yang cukup banyak. Penggunaan metode analisis regresi linier klasik dengan metode *Ordinary Least Square* (OLS) pada *excess zeros data* akan menimbulkan bias dalam data [1]. Hal ini disebabkan observasi yang bernilai nol tidak disertakan dalam persamaan regresi, sehingga tidak akan didapatkan hasil yang optimal. Estimasi yang dihasilkan menjadi tidak konsisten [2].Namun demikian dalam kasus tertentu terdapat sejumlah fenomena dimana variabel respon berbentuk diskrit, sehingga analisis dengan regresi linier ganda atau regresi klasik tidak lagi memberikan hasil yang tepat dan mengakibatkan kesalahan dalam penarikan kesimpulan.

Salah satu model regresi yang dapat digunakan untuk menjelaskan hubungan antara variabel respon yang berupa *count* data dengan variabel prediktor yang berupal kontinu, diskrit atau campuran antara keduanya adalah regresi poisson.Akan tetapi pada beberapa kasus tertentu terdapat nilai nol yang berlebih pada variabel respon sehingga menyebabkan terjadinya overdispersi. Untuk mengatasi masalah overdispersi maka digunakan pendekatan model *hurdle*.

Pendekatan model *hurdle* akan diaplikasikan pada kasus penyakit difteri. Penyakit difteri merupakan salah satu penyakit menupakan salah satu penyakit menular yang oleh bakteri *Corynebacterium diphtheriae*. Difteri menyerang sistem pernapasan bagian atas anak-anak usia 1-10 tahun [3]. Gejala penyakit ini adalah sakit tenggorokan, demam, sulit bernapas dan menelan, mengeluarkan lendir dari mulut dan hidung, dan sangat lemah. Kuman difteri disebarkan melalui cairan dari mulut atau hidung orang yang terinfeksi, jari-jari atau handuk yang terkontaminasi, dan dari susu yang terkontaminasi penderita. Difteri dapat dicegah dengan imunisasi DPT (*Difteri Pertuisis Tetanus*).

Penderita difteri di Provinsi Jawa Barat menempati posisi kelima besar di Indonesia [4]. Berdasarkan referensi [4], banyak penderita difteri di Jawa Barat mengalami peningkatan yang dilaporkan di tahun sebelumnya tidak ditemukan penderita difteri. Diharapkan dengan penelitian ini dapat menjadi prior riset mengenai kasus difteri di Jawa Barat.

II. METODE PENELITIAN

A. *Model Hurdle*

Salah satu pendekatan yang dapat digunakan untuk mengatasi overdispersi adalah model *Hurdle*. Overdipersi terjadi ketika varians lebih besar dari rata-rata. Salah satu penyebab terjadi overdispersi adalah banyaknya nilai nol pada variabel respon [5]. Pada model *hurdle* dilakukan dua jenis pemodelan.

Pemodelan pertama memodelkan observasi yang bernilai nol dengan menggunakan model logistik. Fungsi hubung model logistik adalah logit sesuai dengan (1).

$$\text{logit}\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i^T\alpha \qquad (1)$$

Pemodelan kedua memodelkan observasi yang bernilai positif dengan menggunakan *truncated poisson*. Fungsi hubung yang digunakan adalah log yang ditunjukkan pada (2)

$$\log(\mu_i) = \mathbf{x}_i^T\beta \qquad (2)$$

`Misalkan variabel respon y_i dengan $i = 1,2,\ldots,n$. Dimana $\mathbf{z}_i^T$ dan $\mathbf{x}_i^T$ vector kovariat pada variabel prediktor. Sementara α adalah parameter dari model logit dan β adalah parameter koefisien regresi untuk model *truncated*. Berdasarkan (1) dan (2) maka fungsi peluang model hurdle secara umum adalah

(3)
$$P(Y_i = y_i) = \begin{cases} \dfrac{1}{1+\exp(\mathbf{z}_i^T\alpha)} \\ [C]\left[\dfrac{(\exp(\mathbf{x}_i^T\beta))y_i}{\{(\exp(\mathbf{x}_i^T\beta))-1\}+y_i!}\right] \end{cases}$$

Fungsi peluang pada model *hurdle* merupakan gabungan antara peluang pada model logit dan model *truncated poisson* [6].

Penaksiran parameter model *hurdle* menggunakan metode *Maximum Likelihood Estimation* (MLE). Nilai maksimum fungsi likelihood dapat diperoleh dengan cara menurunkan fungsi likelihoodnya terhadap parameter yang dicari yang kemudian disama dengankan nol. Fungsi yang dihasilkan tidak linear sehingga diselesaikan dengan algoritma *Fisher Scoring*. Fungsi likelihood model *hurdle* adalah

$$L(\alpha,\beta) = \prod_0 \frac{1}{1+\exp(\mathbf{z}_i^T\alpha)} \prod_{y_i>0}\left[\frac{\exp(\mathbf{z}_i^T\alpha)}{1+\exp(\mathbf{z}_i^T\alpha)}\right]\left[\frac{(\exp(\mathbf{x}_i^T\beta))y_i}{\{(\exp(\mathbf{x}_i^T\beta))-1\}+y_i!}\right] \qquad (4)$$

B. Pengujian Parameter Model Hurdle

Persamaan yang mengandung beberapa variabel prediktor dan berpengaruh terhadap variabel respon dapat dilakukan pengujian dengan *likelihood ratio test* [1]. *Likelihood ratiotest* digunakan untuk menguji estimasi parameter secara serentak, sedangkan uji *wald* digunakan untuk pengujian secara individu.

Uji Serentak

Uji serentak digunakan untuk menguji parameter secara bersama-sama. Hipotesis yang digunakan adalah sebagai berikut :

$H_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_1 =$ minimal ada salah satu β yang tidak sama dengan 0

Statistik Uji yang digunakan adalah

$$G^2 = -2\ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) \quad (5)$$

Dimana $L(\hat{\omega})$ = nilai maksimum likelihood tanpa variabel prediktor tertentu

$L(\hat{\Omega})$ = nilai maksimum likelihood dengan variabel prediktor tertentu

H_0 ditolak jika $G^2 > \chi^2_{(\alpha,k)}$, karena G^2 secara *asymptotically* mengikui distribusi *chi-square*. Dimana k adalah banyaknya variabel prediktor model atau jika $p - value < \alpha$ yang berarti ada salah satu atau lebih β_k yang berpengaruh pada model.

Uji Parsial

Uji Parsial digunakan untuk untuk pengujian individu yang menunjukkan apakah suatu variabel bebas signifikan atau layak untuk masuk model. Pengujian parameter parsial untuk masing-masing bagian logit dan Truncated Poisson digunakan untuk menguji masing-masing parameter Pengujian yang digunakan adalah *Wald test* [6].

Hipotesis model logit

$H_0 : \alpha_j = 0$

$H_1 : \alpha_j \neq 0$, dimana $j = 1,2,\cdots,k$

Statistik uji Wald yang digunakan adalah

$$W = \frac{\hat{\alpha}_j}{SE(\hat{\alpha}_j)} \quad (6)$$

Hipotesis model truncated Poisson

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$, dimana $j = 1,2,\cdots,k$

Statistik uji Wald yang digunakan adalah

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (7)$$

Berdasarkan (6) dan (7) H_0 ditolak jika $|W| > Z_{\alpha/2}$, atau jika $p - value < \alpha$ yang berarti bahwa parameter berpengaruh . Sampel besar mengikuti sebaran normal, maka kriteria pengujian dibandingkan dengan tabel normal Z.

C. Data Penelitian

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari Dinas Kesehatan Provinsi Jawa Barat tahun 2012. Variabel penelitian yang digunakan terdiri dari variabel respon (Y) yaitu jumlah kasusdifteri di kabupaten/kota Provinsi Jawa Barat dan beberapa variabel prediktor yang diduga berpengaruh antara lain persentase balita gizi buruk (X_1), jumlah cakupan Imunisasi DPT1+HB1 (X_2), jumlah cakupan Imunisasi DPT3+HB3 (X_3), persentase Rumah Sehat (X_4), rata-rata kepadatan penghuni rumah (X_5) dan persentase keluarga dengan sumber air minum terlindung (X_6)

D. Langkah Penelitian

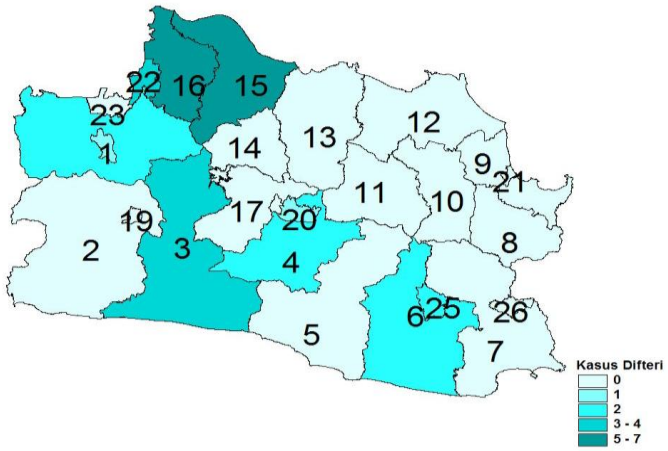
Langkah-langkah yang dilakukan pada penelitian ini adalah melakukan analisis deskriptif variabel penelitian. Kemudian melakukan penaksiran parameter model hurdle. Setelah diperoleh parameter model hurdle maka langkah berikutnya adalah melakukan pengujian parameter model hurdle secara serentak dan parsial. Berdasarkan hasil pengujian parameter hurdle maka dapat dilakukan analisis dan menyimpulkan faktor-faktor yang berpengaruh terhadap penyakit difteri.

III. HASIL DAN PEMBAHASAN

Pada bagian ini diuraikan tentang deskripsi kasus penyakit difteri di Jawa Barat. Selain itu juga diuraikan persebaran kasus penyakit difteri beserta fakor-faktor yang mempengaruhinya dengan menggunkaan model *hurdle*.

A. Deskripsi Variabel Penelitian

Jumlah kasus penyakit difteri di Jawa Barat sebanyak 31 kasus yang tersebar pada 26 kabupaten/kota. Terdapat 16 kabupaten/kota yang tidak ditemukan kasus difteri. Sementara itu jumlah kasus penyakit difteri paling banyak di Kabupaten Karawang dan Kabupaten Bekasi sebanyak 7 kasus. Persebaran kasus difteri di Jawa Barat tahun 2012 disajikan pada Gambar 1.



B. Pemodelan Penyakit Difteri

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 23,827 yang lebih besar dari $\chi^2_{(0,05;6)}$= 12.59. Hal ini berarti bahwa minimal ada satu parameter yang berpengaruh secara signifikan terhadap model. Penaksiran model logit disajikan pada Tabel 1.

TABEL 1. ESTIMASI PARAMETER MODEL LOGIT

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
α_0	-7,7784	5,2354	-1,486	0,137
α_1	0,2833	1,4088	0,201	0,841
α_2	0,2070	0,7665	0,27	0,787
α_3	-0,1675	0,7824	-0,214	0,83
α_4	0,0010	0,0411	0,023	0,981
α_5	2,4492	1,2025	2,0367	0,026*
α_6	0,0025	0,0344	0,072	0,943

Berdasarkan tabel 1 dapat diketahui bahawa variabel prediktor yang signifikan dengan tingkat kesalahan 5% pada model logit adalah rata-rata kepadatan penghuni rumah (X_5). Model logit dapat dikatakan sebagai indikator apakah suatu kabupaten/kota di Provinsi Jawa Barat memiliki kecenderungan ditemukan kejadian difteri atau tidak. Persamaan model logit berdasarkan tabel 1 adalah

$$\text{logit}\left(\frac{\pi_i}{1-\pi_i}\right) = -7{,}7784 + 2{,}4492\,X_5 \qquad (8)$$

Dari hasil (8), hal ini berarti bahwa semakin padat penghuni rumah maka kecenderungan suatu kabupaten/kota di Provinsi Jawa Barat ditemukan kasus difteri semakin tinggi. Setiap penambahan satu penghuni rumah maka cenderung akan ditemukan kasus difteri di kabupaten/kota sebesar exp (2,4492) = 11,58 kali.

Hasil penaksiran parameter dari model *truncated poisson* disajikan dari tabel 2. Berdasarkan tabel 2 dapat diketahui bahwa variabel prediktor yang signifikan pada taraf kesalahan 5% adalah rata-rata kepadatan penghuni rumah (X_5) dan persentase keluarga dengan sumber air minum terlindung (X_6).

TABEL 2. ESTIMASI PARAMETER MODEL *TRUNCATED POISSON*

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,68359	3,55118	-0,474	0,6354
x1	1,16939	1,37182	0,852	0,394
x2	0,25269	0,42451	0,595	0,5517
x3	-0,28619	0,45152	-0,634	0,5262
x4	-0,04633	0,03448	-1,344	0,179
x5	2,03476	0,85283	2,386	0,017
x6	0,03105	0,01209	2,568	0,0102

Persamaan model *truncated poisson* adalah

$$\mu_i = \exp(-1{,}68359 + 2{,}03476\,X_5 + 0{,}03105\,X_6) \qquad (9)$$

Berdasarkan (9), hal ini berarti bahwa semakin padat penghuni rumah dan semakin tinggi persentase sumber air minum terlindung maka kecenderungan suatu kabupaten/kota di Provinsi Jawa Barat ditemukan kasus difteri semakin tinggi. Setiap penambahan satu penghuni rumah maka akan meningkatkan rata-rata terjadi kasus difteri di kabupaten/kota adalah exp (2,03476) = 7,6504. Selain itu setiap penambahan satu persen sember air minum terlindung maka akan meningkatkan rata-rata terjadi kasus difteri adalah exp (0,03105) = 1,03154.

IV. SIMPULAN DAN SARAN

Berdasarkan hasil pemodelan diperoleh variabel yang berpengaruh terhadap jumlah kasus difteri adalah rata-rata kepadatan penghuni rumah pada model logit sedangkan rata-rata kepadatan penghuni rumah dan persentase keluarga dengan sumber air minum terlindung pada model *hurdle*. Penelitian ini belum memperhatikan adanya pencilan (*outlier*) dalam pemodelan. Perlu dilakukan pemodelan Hurdle Poisson yang mempertimbangkan adanya pencilan dengan metode robust.

DAFTAR PUSTAKA

[1]W.H.Greene, "Econometrics Analysis, 6thedition," New Jersey: Prentice Hall, 2008.
[2] J.S.Long, "Regression Models for Categorical and Limited Dependent Variables", California: Sage Publications Inc, 1997.
[3] Kemenkes. "Laporan Riset Kesehatan Dasar", Jakarta: BPPK Kemenkes RI, 2013.
[4] Dinas Kesehatan Provinsi Jawa Barat. "Profil Kesehatan Provinsi Jawa Barat Tahun 2012", Bandung: Dinas Kesehatan Provinsi Jawa Barat,2012.

[5] C.J.W. Zorn, "Evaluating Zero Inflated and Hurdle Poisson Specifications", Ohio State University: Midwest Political Science Assosiation, 1996.
[6] E.CantoniandA.Zedini, "A Robust Version of the Hurdle Model", Journal of Statistical Planning and Inference, Vol.141(3), pp:1214-1223, 2010.