PEMODELAN JUMLAH HARI SAKIT MENGGUNAKAN ZERO INFLATED NEGATIVE BINOMIAL DAN HURDLE NEGATIVE BINOMIAL

Kusni Rohani Rumahorbo[1], Budi Susetyo[2‡], Kusman Sadik[3]

[1]Badan Pusat Statistik Kabupaten Aceh Tengah, Indonesia, kusnirohani11@gmail.com
[2]Department of Statistics, IPB University, Indonesia, buset008@yahoo.com
[3]Department of Statistics, IPB University, Indonesia, kusmansadik@gmail.com
‡corresponding author

Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)
Vol 3 No 2 (2019), 184 - 201

Abstract

Health is a very important thing for humanity. One way to look at a person's health condition is through the number of unhealthy days which can also shows the productivity of the community in a region. Modeling the number of unhealthy days which are examples of count data can be done using Poisson regression. Problems that are often faced in data counts are overdispersion and excess zero. Poisson regression cannot be applied to data that experiences both of these. Zero Inflated Negative Binomial and Hurdle Negative Binomial modeling was performed on data with 2 conditions, uncensored and censored. The explanatory variables used are gender, age, marital status, education level, home ownership status and rural-urban status. According to the results of the AIC and RMSE calculation, Zero Inflated Negative Binomial on censored data showed the best performance for estimating the number of unhealthy days.

Keywords: zero-inflated, hurdle, CZINB, CHNB, unhealthy-days.

1. Pendahuluan

Kesehatan adalah hal yang sangat penting bagi umat manusia. Kondisi kesehatan seseorang mempengaruhi kualitas hidupnya. Seseorang yang sehat berkesempatan memanfaatkan segala sumber daya yang ada pada dirinya untuk menjadi pribadi yang produktif. Hal ini sesuai dengan definisi kesehatan dalam UU RI No. 36 Tahun 2009 dalam Pemerintah Republik Indonesia (2009), keadaan sehat adalah kondisi baik secara fisik, mental, spritual yang memungkinkan setiap orang untuk hidup produktif secara sosial dan ekonomis.

Menilai kondisi kesehatan seseorang dapat dilakukan, salah satunya melalui pendekatan morbiditas, yaitu ”jumlah hari sakit-terganggunya pekerjaan, sekolah atau kegiatan sehari-hari lainnya akibat adanya keluhan kesehatan”. Semakin lama jumlah hari terganggunya kegiatan sehari-hari maka semakin rendah kondisi kesehatannya. Jika hal ini terjadi pada banyak orang di suatu wilayah, maka derajat kesehatan di wilayah tersebut juga akan rendah.

Hasil Susenas bulan Maret 2017, menunjukkan bahwa persentase penduduk Indonesia yang mengalami keluhan kesehatan sebesar 28.62 persen. Persentase penduduk yang mengalami keluhan kesehatan dan mengakibatkan terganggunya kegiatan sehari-hari sebesar 14.31 persen, dengan rata-rata lama sakit adalah 5.42 hari (BPS 2017).

Penelitian di bidang kesehatan seperti morbiditas ini sangat penting dalam rangka meningkatkan kualitas hidup manusia. Penelitian tersebut dapat digunakan untuk melihat karakteristik penduduk yang mengalami keluhan kesehatan atau memprediksi bagaimana kondisi kesehatan pada wilayah tertentu di masa mendatang.

Fenomena morbiditas seperti ”jumlah hari terganggunya pekerjaan, sekolah atau kegiatan sehari-hari lainnya akibat adanya keluhan kesehatan” merupakan contoh data cacah (*count*). Metode dasar untuk memodelkannya adalah regresi Poisson (Hu *et al.* 2011; Hofstetter *et al.* 2016). Poisson memiliki asumsi ekuidispersi yaitu nilai ragam yang sama besar dengan rataan. Hal ini merupakan kelemahan Poisson karena data cacah pada praktiknya sering mengalami overdispersi (ragam lebih besar daripada rataan). Jika regresi Poisson tetap diterapkan pada data cacah yang overdispersi maka kesalahan baku dari pendugaan parameter regresi yang dihasilkan akan terlalu kecil. Kondisi tersebut mengakibatkan kesimpulan yang diambil tidak sesuai dengan data sebenarnya (Coxe *et al.* 2009).

Secara umum, masalah yang terjadi pada data cacah selain overdispersi adalah nilai nol berlebih (*excess zero*). Regresi Poisson tidak mampu mengatasi permasalahan ini. Beberapa model regresi alternatif dikembangkan oleh banyak peneliti antara lain *Zero Inflated* (Lambert 1992) dan *Hurdle* (Mullahy 1986). Kedua model ini mengakomodasi nilai nol berlebih dalam data dengan cara mengkombinasikan sebaran Bernoulli dan sebaran lainnya (biasanya Poisson dan *Negative Binomial*) pada dua proses yang berbeda secara sistematis. Proses pertama adalah model logit yang digunakan untuk menentukan peluang dari peubah respon suatu amatan bernilai nol sedangkan proses kedua adalah model log yang digunakan untuk menentukan peluang dari peubah respon suatu amatan bernilai selain nol.

Desjardin (2013) mengatakan bahwa untuk mengatasi overdispersi pada data yang memiliki banyak amatan bernilai nol (*excess zero*), sebaran *Negative Binomial* sangat tepat digunakan dalam model regresi *Zero Inflated* dan *Hurdle*. Penelitian Rose *et al.* (2006), Hu *et al.* (2011) dan Hofstetter *et al.* (2016) pada aplikasi di bidang kesehatan, membandingkan beberapa model regresi dengan *Zero Inflated Negative Binomial* (ZINB) dan *Hurdle Negative Binomial* (HNB). Data yang digunakan merupakan data cacah yang mengalami overdispersi dan *excess zero*.

Hasilnya menunjukkan bahwa ZINB dan HNB paling baik akurasinya dibanding model regresi Poisson, *Negative Binomial* (NB), *Zero Inflated Poisson* (ZIP), dan *Hurdle Poisson* (HP).

Penelitian pada data hari sakit yang overdispersi dan *excess zero* juga telah dilakukan oleh beberapa peneliti. Yang *et al.* (2017) meneliti hari sakit di Negara Bagian Rhode Island, Amerika Serikat dengan membandingkan model regresi ZIP, ZINB, HP, dan HNB. Hasilnya menunjukkan bahwa model regresi ZINB dan HNB menghasilkan performa yang lebih baik daripada model regresi lainnya. Sumarni (2009) meneliti banyaknya gangguan aktivitas primer yang disebabkan sakit menggunakan model regresi *Zero Inflated Generalized Poisson* di antara beberapa kelompok sosial di Kabupaten/Kota di Provinsi Bali. Beberapa peubah penjelas yang signifikan diantaranya jenis kelamin, umur, pendidikan, daerah tempat tinggal, sumber air minum dan pengeluaran/konsumsi.

Pada perkembangan pemodelan, penyensoran dilakukan pada data cacah seperti pada Famoye dan Wang (2003). Pembatasan atau penyensoran pada peubah respon untuk beberapa kasus dengan tujuan tertentu seperti data yang terlalu menceng perlu dilakukan (Frone 1997). Penyensoran dilakukan dengan terlebih dulu menentukan titik-titik sensor pada data cacah. Pada sensor kanan, nilai-nilai yang berada di atas nilai (titik) sensor akan diakumulasikan ke titik sensor tersebut (Greene 2005). Penelitian Saffari dan Adnan (2011) dan Saffari *et al.* (2012) yang melakukan pemodelan ZINB dan HNB pada data tersensor kanan (*Censored* ZINB (CZINB) dan *Censored* HNB (CHNB)) menghasilkan pemodelan yang lebih baik daripada data yang tidak disensor.

Berdasarkan hal tersebut, maka penelitian ini bertujuan untuk menerapkan model regresi *Zero Inflated Negative Binomial* (ZINB), *Hurdle Negative Binomial* (HNB), *Zero Inflated Negative Binomial* dengan data tersensor (CZINB) dan *Hurdle Negative Binomial* dengan data tersensor (CHNB) terhadap data jumlah hari terganggunya pekerjaan, sekolah atau kegiatan sehari-hari lainnya akibat adanya keluhan kesehatan di Provinsi Gorontalo tahun 2017. Provinsi Gorontalo terpilih sebagai tempat penelitian karena memiliki permasalahan morbiditas yang tinggi. Hasil dari keempat pemodelan tersebut akan dibandingkan untuk memperoleh model terbaik.

2. Metodologi

2.1 Bahan dan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang berasal dari hasil Survei Sosial Ekonomi Nasional (Susenas) Maret 2017 untuk Provinsi Gorontalo. Unit penelitiannya adalah penduduk berusia 35 tahun ke atas. Peubah penjelas yang digunakan dalam penelitian ini disajikan pada tabel berikut:

Tabel 1 Peubah yang digunakan dalam penelitian

Peubah	Deskripsi	Kategori
Y	Model *zero inflation*	
	Terganggunya Kegiatan Sehari-hari Akibat Keluhan Kesehatan	0 = Tidak terganggunya kegiatan sehari-hari akibat keluhan kesehatan 1 = terganggunya kegiatan sehari-hari akibat keluhan kesehatan
	Model *count*	
	Rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan	-
X_1	Jenis Kelamin	0 = Perempuan, 1 = Laki-laki
X_2	Umur (tahun)	-
X_3	Status Perkawinan	0 = Belum menikah/Cerai Hidup/Cerai Mati, 1 = Menikah
X_4	Tingkat Pendidikan	0 = di bawah SMA sederajat, 1 = SMA sederajat ke atas,
X_5	Status Kepemilikan Rumah	0 = Selain milik sendiri, 1 = Milik sendiri
X_6	Daerah tempat tinggal	0 = Perdesaan, 1 = Perkotaan

2.2 Metode Penelitian

Tahapan-tahapan penelitian yang digunakan untuk mencapai tujuan penelitian ini adalah sebagai berikut:

1. Melakukan eksplorasi data (analisis deskriptif) pada peubah respon dan peubah penjelas.
2. Melihat ada atau tidaknya multikolinieritas antar peubah penjelas.
   Menghitung besar nilai korelasi antar peubah penjelas digunakan ukuran yang sesuai dengan skala peubah penjelas. Koefisien korelasi *tetrachoric* digunakan antar peubah kategorik (dikotomi/biner), sedangkan *point biserial* digunakan antara peubah kategorik (dikotomi) dengan peubah kontinu (Olsson *et al*. 1982; Das D dan Das A 2017).
   Koefisien korelasi *tetrachoric* dapat dihitung dengan rumus berikut:

$$r_T = \cos\frac{180^0\sqrt{BC}}{\sqrt{AD}+\sqrt{BC}} \quad (1)$$

   dengan AD/BC menunjukkan nilai *odds ratio* pada tabel 2x2.
   Rumus berikut menghitung nilai r_{pbi} yang merupakan koefisien korelasi *point biserial* dengan p dan q adalah proporsi kasus atau individu dalam dua kelas pada peubah biner, $\bar{X}_p$ dan $\bar{X}_q$ adalah skor rata-rata dari peubah numerik (kontinu) untuk kasus atau individu yang termasuk dalam kelas masing-

masing peubah biner, dan S_X adalah simpangan baku dari seluruh skor pada peubah kontinu.

$$r_{pbi} = \frac{\overline{X}_p - \overline{X}_q}{S_X}\sqrt{\frac{p}{q}} \quad (2)$$

3. Mengidentifikasi overdispersi pada peubah respon
   McCullagh dan Nelder (1998) memberikan cara untuk mendeteksi overdispersi, yaitu dengan *Likelihood Ratio Goodness of Statistics*:

$$\mu = \frac{Devians}{db} \quad (3)$$

dengan $Devians = -2ln\left(\frac{L(y_i,y_i)}{L(\hat{\mu}_i,y_i)}\right) = 2\sum_{i=1}^{n} y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)$ (4)

   dimana db menyatakan derajat bebas. Jika nilai μ>1, maka model regresi Poisson dikatakan mengalami gejala overdispersi. Jika hal ini terjadi, maka regresi Poisson menjadi tidak tepat menggambarkan data yang sebenarnya.
4. Membagi data ke dalam data *training* dan *testing* secara acak. Penghitungan AIC dilakukan pada data *training*, sedangkan penghitungan *Root Mean Square Error Prediction* (RMSEP) dilakukan pada data *testing*.
5. Melakukan pendugaan dan pengujian hipotesis pada data *training* untuk model regresi ZINB, CZINB (10), HNB dan CHNB (10). Titik sensor 10 dipilih karena pertimbangan rentang yang akan disensor.
6. Melakukan penyeleksian peubah dengan metode *backward elemination* pada model regresi ZINB, CZINB (10), HNB, dan CHNB (10). Metode penyeleksian peubah ini memulai analisis pada model penuh. Seluruh peubah penjelas dimasukkan ke dalam model kemudian peubah yang memberikan pengaruh kecil atau memiliki nilai peluang paling besar dieliminasi secara bertahap.
7. Memilih model terbaik pada data *training* dengan menggunakan AIC.
   Membandingkan model berdasarkan *Maximum Likelihood Estimation*, Akaike dalam Cameron dan Trivedi (1998) mengusulkan kriteria pemilihan berdasarkan fungsi *ln likelihood*. Cara mendapatkan nilai AIC adalah sebagai berikut:

$$AIC = -2l(\alpha,\gamma,\boldsymbol{\beta}) + 2p \quad (5)$$

   dengan *l* adalah nilai *ln likelihood* dari model dan *p* adalah banyaknya parameter dalam model. Pemilihan model terbaik dilihat dari nilai terkecil dari AIC.
8. Melakukan penghitungan *Root Mean Square Error Prediction* (RMSEP) pada data *testing*.

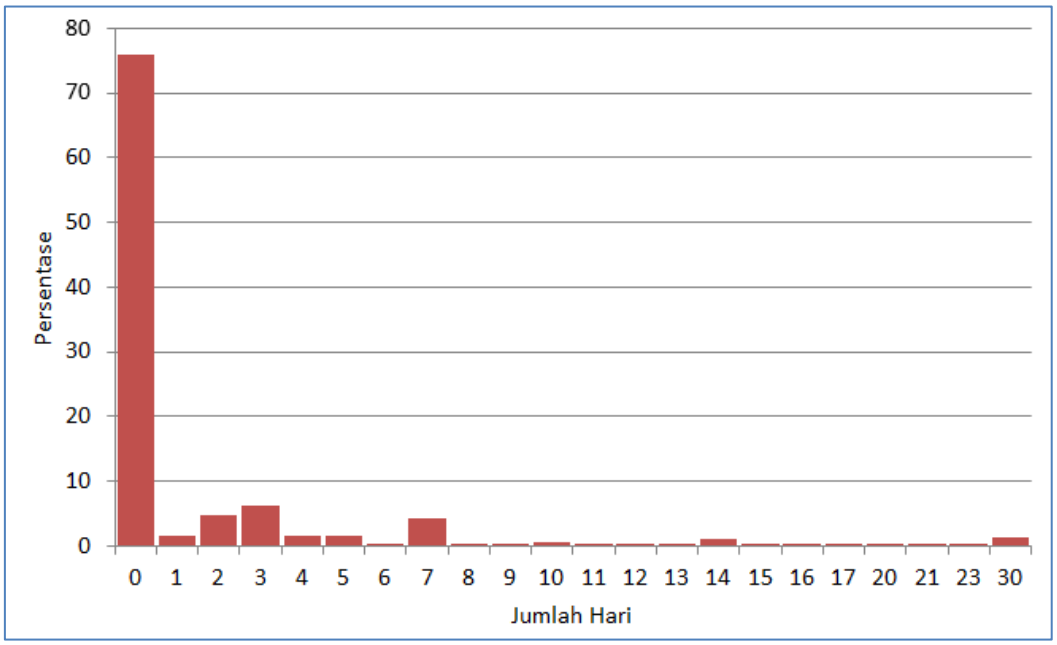
$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (6)$$

dengan y_i = data awal, $\hat{y}_i$ = data hasil pendugaan dan n = jumlah data. Data dianalisis menggunakan software SAS versi 9.2 dan Microsoft Excel.

3. Hasil dan Pembahasan

3.1 Deskripsi Data

Eksplorasi data berguna untuk mempelajari karakteristik data agar lebih mudah dalam menentukan model analisis statistik yang sesuai. Gambar 1 merupakan histogram jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017. Gambar tersebut memperlihatkan bentuk yang tidak simetris dan cenderung membentuk pola eksponen. Pada gambar juga terlihat bahwa persentase amatan bernilai nol sangat tinggi, yaitu mencapai 75.85%. Hal inilah yang menjadi salah satu fokus dalam penelitian ini, yaitu banyaknya amatan yang bernilai nol (*excess zero*).



Gambar 1 Histogram dari persentase frekuensi penduduk menurut jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan Provinsi Gorontalo tahun 2017

Deskripsi tentang responden djelaskan dalam Tabel 2. Pada tabel diketahui terdapat 50.89% responden perempuan dan laki-laki 49.11%. Responden yang berstatus menikah adalah sebanyak 82.99%, sedangkan yang belum menikah, cerai hidup, dan cerai mati sebanyak 17.01%. Dari keseluruhan responden, yang berpendidikan minimal SMA sederajat adalah sebanyak 24.15%, sedangkan sisanya, yaitu 75.85% memiliki pendidikan kurang dari SMA sederajat. Responden yang

memiliki rumah yang ditempati (milik sendiri) adalah sebanyak 85.85%, sedangkan selainnya (sewa/kontrak dan lain sebagainya) adalah sebanyak 14.15 %. Responden yang tinggal di daerah perkotaan sebanyak 34.17% sedangkan sisanya, 65.83% tinggal di daerah perdesaan.

Tabel 2: Proporsi responden menurut peubah penjelas kategorik

Kategori	Jenis Kelamin (X_1)	Status Perkawinan (X_3)	Pendidikan (X_4)	Status Kepemilikan Rumah (X_5)	Daerah Tempat Tinggal (X_6)
0	50.89	17.01	75.85	14.15	65.83
1	49.11	82.99	24.15	85.85	34.17

Pemeriksaaan Korelasi Antar Peubah Penjelas

Sebelum melakukan pemodelan regresi berganda, dilakukan pemeriksaan terhadap korelasi antar peubah penjelas. Penghitungan nilai korelasi antar peubah penjelas dalam penelitian inin dilakukan dengan menggunakan koefisien korelasi *Tetrachoric* dan *Point Biserial* (Tabel 3 dan Tabel 4). Koefisien korelasi *tetrachoric* digunakan antar peubah penjelas yang kategorik (dikotomi), sedangkan *point biserial* digunakan antara peubah kategorik (dikotomi) dengan peubah kontinu. Hasilnya menunjukkan bahwa hubungan antar peubah penjelas tidak ada yang kuat.

Tabel 3: Korelasi antar peubah penjelas kategorik dikotomi

Peubah	X_1	X_3	X_4	X_5	X_6
X_1	-	0.3024	0.0550	0.0007	-0.0368
X_3	0.3024	-	0.0082	0.0787	-0.1763
X_4	0.0550	0.0082	-	0.2066	-0.4277
X_5	0.0007	0.0787	0.2066	-	-0.2445
X_6	-0.0368	-0.1763	-0.4277	-0.2445	-

Tabel 4: Korelasi antara peubah penjelas kategorik dikotomi dan kontinu

Peubah	X_1	X_3	X_4	X_5	X_6
X_2	-0.0334	-0.3288	-0.0933	0.0874	-0.0045

Pemeriksaan Overdispersi

Menguji asumsi ekuidispersi dilakukan pada model regresi Poisson yaitu melalui penghitungan nilai dispersi. Nilai dispersi didapat yaitu dengan cara membagi nilai *deviance* dengan derajat bebasnya. Jika nilai *deviance* dibagi dengan derajat bebasnya sama dengan satu, maka asumsi ekuidispersinya terpenuhi. Apabila

nilainya lebih besar dari satu, maka diindikasikan terjadi overdispersi (nilai ragam lebih besar daripada rataan), sedangkankan apabila nilainya kurang dari satu, maka diindikasikan terjadi underdispersi (nilai ragam lebih kecil daripada rataan). Berdasarkan hasil analisis diperoleh nilai dispersi seperti yang disajikan dalam Tabel 5 berikut:

Tabel 5: *Dispersi* model regresi Poisson

Kriteria	Nilai
db	4 305.00
Dispersi	4.97

Nilai *dispersi* sebesar 4.97 (>1) mengindikasikan overdispersi. Telah terjadi pelanggaran terhadap asumsi yang harusnya dipenuhi saat menerapkan model regresi Poisson. Saat terjadi overdispersi dan mengalami masalah dengan banyaknya amatan bernilai nol, maka pemodelan dapat dilakukan dengan menggunakan *Zero Inflated Negative Binomial* atau *Hurdle Negative Binomial.* Kedua model tersebut akan diterapkan pada data yang tidak disensor (ZINB dan HNB) dan yang disensor di titik 10 (CZINB (10) dan CHNB (10)).

Pemodelan dengan Model Regresi ZINB

Setelah dilakukan pemeriksaan overdispersi dan korelasi antar peubah penjelas, selanjutnya dilakukan pemodelan menggunakan model regresi *Zero Inflated Negative Binomial* (ZINB). Hasil analisis disajikan pada Tabel 6. Model regresi ZINB memberikan hasil akhir dari dua proses yang berbeda secara sistematis. Model pertama adalah model logit yang digunakan untuk menentukan peluang dari peubah respon suatu amatan bernilai nol sedangkan model kedua adalah model log yang digunakan untuk menentukan besarnya peubah respon suatu amatan. Misal dalam kasus terganggunya kegiatan sehari-hari akibat keluhan kesehatan, model *zero inflation* menjelaskan tentang terganggu atau tidaknya kegiatan sehari-hari seseorang akibat keluhan kesehatan sedangkan model *count* menjelaskan tentang rata-rata lama (hari) terganggunya kegiatan sehari-hari seseorang akibat keluhan kesehatan.

Hasil pemodelan menggunakan model regresi ZINB pada Tabel 6 menunjukkan bahwa terdapat peubah penjelas yang tidak signifikan. Untuk itu tahapan selanjutnya adalah penyeleksian peubah penjelas. Penyeleksian peubah penjelas menggunakan metode *backward elimination*. Analisis dimulai dengan model penuh yaitu memasukkan seluruh peubah penjelas ke dalam model kemudian mengeliminasi peubah penjelas dari masing-masing model secara bertahap. Peubah yang dieliminasi adalah peubah yang memberikan pengaruh paling kecil atau memiliki nilai peluang paling besar. Berikut ringkasan hasil eliminasi menggunakan metode *backward elimination*.

Tabel 6: Pendugaan parameter model regresi ZINB

Parameter	Nilai Dugaan	Galat Baku	Nilai-p
Model *Zero Inflation*			
γ_0	2.6310	0.2650	<.0001
γ_1	-0.0787	0.0808	0.3301
γ_2	-0.0336	0.0038	<.0001
γ_3	-0.1867	0.1121	0.0958
γ_4	0.2776	0.1000	0.0055
γ_5	0.0940	0.1161	0.4179
γ_6	0.0174	0.0870	0.8416
Model *Count*			
β_0	0.6443	0.2305	0.0052
β_1	0.0918	0.0685	0.1803
β_2	0.0193	0.0032	<.0001
β_3	-0.0606	0.0943	0.5207
β_4	-0.0680	0.0872	0.4356
β_5	-0.1081	0.0985	0.2725
β_6	0.1277	0.0725	0.0782
1/k	0.9482	0.0818	<.0001
AIC	10 261.0		

Penyeleksian peubah

a. Tahap pertama: Eliminasi $X_6|X_3$
Pada tahap pertama ini, dijelaskan pada Tabel 7, peubah penjelas yang pertama kali dieliminasi adalah daerah tempat tinggal (perkotaan-perdesaan) (X_6) di model *zero inflation* dan status perkawinan (X_3) di model *count*. Nilai-p dari kedua peubah penjelas tersebut dalam Tabel 6 adalah sebesar 0.8416 untuk daerah tempat tinggal (perkotaan-perdesaan) (X_6) dan 0.5207 untuk status perkawinan (X_3). Pada tahap ini terlihat nilai AIC mengalami penurunan dari 10 261.0 menjadi 10 257.0.

b. Tahapan kedua: Eliminasi $X_6, X_1|X_3, X_4$
Pada tahap kedua ini, dijeaskan pada Tabel 8, peubah penjelas yang dieliminasi adalah jenis kelamin (X_1) di model *zero inflation* dan pendidikan (X_4) di model *count*. Pada Tabel 7, keduanya memiliki nilai-p yang paling besar di antara peubah penjelas yang lain, yaitu 0.3962 untuk peubah penjelas jenis kelamin (X_1) dan 0.4513 untuk peubah penjelas pendidikan (X_4). Nilai AIC menunjukkan penurunan dari 10 257.0 di tahapan sebelumnya menjadi 10 254.0

Tabel 7: Pendugaan parameter penyeleksian peubah tahap pertama model ZINB

Parameter	Nilai Dugaan	Galat Baku	Nilai-p
Model *Zero Inflation*			
γ_0	2.6293	0.2597	<.0001
γ_1	-0.0684	0.0806	0.3962
γ_2	-0.0337	0.0038	<.0001
γ_3	-0.1929	0.1100	0.0796
γ_4	0.2851	0.0971	0.0033
γ_5	0.1010	0.1153	0.3811
Model *Count*			
β_0	0.6349	0.1876	0.0007
β_1	0.0751	0.0673	0.2644
β_2	0.0189	0.0029	<.0001
β_4	-0.0660	0.0875	0.4513
β_5	-0.1165	0.0986	0.2376
β_6	0.1311	0.0713	0.0661
1/k	0.9522	0.0824	<.0001
AIC	10 257.0		

Tabel 8: Pendugaan parameter penyeleksian peubah tahap kedua model ZINB

Parameter	Nilai Dugaan	Galat Baku	Nilai-p
Model *Zero Inflation*			
γ_0	2.6169	0.2590	<.0001
γ_2	-0.0339	0.0038	<.0001
γ_3	-0.2182	0.1085	0.0444
γ_4	0.3088	0.0949	0.0011
γ_5	0.1092	0.1153	0.3437
Model *Count*			
β_0	0.6060	0.1842	0.0010
β_1	0.0834	0.0658	0.2053
β_2	0.0191	0.0029	<.0001
β_5	-0.1137	0.0979	0.2452
β_6	0.1233	0.0704	0.0800
1/k	0.9553	0.0829	<.0001
AIC	10 254.0		

c. Tahapan ketiga: Eliminasi $X_6, X_1, X_5 | X_3, X_4, X_5$

Pada tahap ketiga ini, dijelaskan dalam Tabel 9, status kepemilikan rumah (X_5) dieliminasi bersamaan di kedua model. Nilai-p dari status kepemilikan rumah di kedua model pada Tabel 8 adalah yang paling besar di antara peubah penjelas yang tersisa, yaitu masing-masing sebesar 0.3437 dan 0.2452. Nilai AIC pada

model hasil seleksi tahap ini adalah sebesar 10 253.0, menurun dibanding tahap sebelumnya.

Tabel 9: Pendugaan parameter penyeleksian peubah tahap ketiga model ZINB

Parameter	Nilai Dugaan	Galat Baku	Nilai-p
Model *Zero Inflation*			
γ_0	2.6509	0.2479	<.0001
γ_2	-0.0330	0.0038	<.0001
γ_3	-0.2024	0.1081	0.0611
γ_4	0.3004	0.0944	0.0015
Model *Count*			
β_0	0.4832	0.1664	0.0037
β_1	0.0904	0.0660	0.1708
β_2	0.0194	0.0029	<.0001
β_6	0.1373	0.0699	0.0494
1/k	0.9610	0.0837	<.0001
AIC	10 253.0		

d. Tahapan keempat: Eliminasi $X_6, X_1, X_5 | X_3, X_4, X_5, X_1$

Pada tahapan sebelumnya yang dijelaskan dalam Tabel 9, masih ditemukan peubah penjelas yang tidak signifikan di model *count*, yaitu peubah penjelas jenis kelamin (X_1). Untuk itu peubah tersebut dieliminasi pada tahap ini, dengan nilai-p sebesar 0.1708. Nilai AIC pada hasil seleksi tahap terakhir ini (Tabel 10) tidak mengalami perubahan dibanding tahap sebelumnya, yaitu tetap sebesar 10 253.0

Tabel 10: Pendugaan parameter penyeleksian peubah tahap keempat model ZINB

Parameter	Nilai Dugaan	Galat Baku	Nilai-p
Model *Zero Inflation*			
γ_0	2.6644	0.2478	<.0001
γ_2	-0.0332	0.0038	<.0001
γ_3	-0.2081	0.1080	0.0540
γ_4	0.3008	0.0944	0.0015
Model *Count*			
β_0	0.5489	0.1593	0.0006
β_2	0.0190	0.0028	<.0001
β_6	0.1369	0.0699	0.0501
1/k	0.9623	0.0838	<.0001
AIC	10 253.0		

Proses penyeleksian peubah penjelas yang tidak signifikan terhadap terganggunya kegiatan sehari-hari akibat keluhan kesehatan pada model regresi

ZINB di Provinsi Gorontalo melalui empat tahapan, dengan ringkasan sebagai berikut:

Tabel 11: Ringkasan penyeleksian peubah penjelas model ZINB

Peubah yang dieliminasi		Nilai AIC	Parameter yang signifikan
Model *Zero Inflation*	Model *Count*		
-	-	10 261.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6	X_3	10 257.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6 dan X_1	X_3 dan X_4	10 254.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_3, X_4 dan X_5	10 253.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_3, X_4, X_5 dan X_1	10 253.0	1/k, γ_2, γ_3, γ_4, β_2, β_6

Hasil akhir dari proses penyeleksian peubah penjelas tersebut yaitu pada model *zero inflation* peubah yang signifikan berpengaruh terhadap terganggu atau tidaknya kegiatan sehari-hari akibat keluhan kesehatan adalah umur (X_2), status perkawinan (X_3) dan pendidikan (X_4) sedangkan pada model *count* umur (X_2) dan daerah tempat tinggal (X_6) yang signifikan berpengaruh terhadap rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan.

Dari Tabel 11 tersebut juga terlihat bahwa nilai AIC pada setiap tahapan penyeleksian mengalami penurunan. Pada saat belum adanya proses eliminasi pada peubah penjelas, nilai AIC sebesar 10 261.0, kemudian turun hingga mencapai 10 253.0 pada tahap keempat. Hingga tahap keempat ini 3 peubah penjelas telah dieliminasi pada model *zero inflation* (daerah tempat tinggal, jenis kelamin dan status kepemilikan rumah) dan 4 peubah penjelas (status perkawinan, pendidikan, status kepemilikan rumah dan jenis kelamin) pada model *count*.

Pemodelan dengan Model Regresi CZINB (10), HNB, dan CHNB (10)

Hasil pemodelan menggunakan model regresi CZINB (10), HNB, dan CHNB (10) memiliki kesamaan dengan hasil pada pemodelan ZINB. Proses penyeleksian peubah penjelas pada ketiga model regresi tersebut juga menggunakan metode *backward elimination*. Penyeleksian peubah penjelas pada model regresi CZINB (10), HNB dan CHNB (10) menghasilkan 3 peubah penjelas yang signifikan pada model *zero inflation* dan 2 peubah penjelas pada model *count*, persis sama dengan hasil pemodelan ZINB.

Umur (X_2) dan pendidikan (X_4) signifikan berpengaruh dalam model *zero inflation*, yaitu terganggu atau tidaknya kegiatan sehari-hari akibat keluhan kesehatan, sedangkan umur (X_2) dan daerah tempat tinggal (X_6) berpengaruh pada rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan. Ringkasan proses penyeleksian peubah penjelas model regresi CZINB (10), HNB dan CHNB (10) disajikan pada Tabel 12, Tabel 13 dan Tabel 14 berikut.

Tabel 12: Ringkasan penyeleksian peubah penjelas CZINB (10)

Peubah yang dieliminasi		Nilai AIC	Parameter yang signifikan
Model *Zero Inflation*	Model *Count*		
-	-	8 497.2	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6	X_3	8 493.3	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6 dan X_1	X_3 dan X_4	8 490.5	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_3, X_4 dan X_5	8 489.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_3, X_4, X_5 dan X_1	8 488.2	1/k, γ_2, γ_3, γ_4, β_2, β_6

Sama dengan proses penyeleksian peubah penjelas pada ZINB, proses pneyeleksian pada CZINB (10) melalui 4 tahap, dengan hasil akhir yaitu peubah yang signifikan berpengaruh terhadap terganggu atau tidaknya kegiatan sehari-hari akibat keluhan kesehatan (model *zero inflation*) adalah umur (X_2), status perkawinan (X_3) dan pendidikan (γ_4), sedangkan peubah yang signifikan berpengaruh terhadap rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan (model *count*) adalah umur (X_2) dan daerah tempat tinggal (X_6). Dari Tabel 12 juga terlihat bahwa nilai AIC pada setiap tahapan penyeleksian mengalami penurunan. Nilai AIC sebesar awal sebesar 8 497.2, kemudian turun hingga mencapai 8 488.2 pada tahap keempat.

Tabel 13: Ringkasan penyeleksian peubah penjelas HNB

Peubah yang dieliminasi		Nilai AIC	Parameter yang signifikan
Model Zero Inflation	Model Count		
-	-	10 260.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6	X_1	10 258.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6 dan X_1	X_1 dan X_4	10 256.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_1, X_4 dan X_3	10 254.0	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_1, X_4, X_3 dan X_5	10 253.0	1/k, γ_2, γ_3, γ_4, β_2, β_6

Sama dengan proses penyeleksian peubah penjelas pada pemodelan ZINB dan CZINB (10), peubah yang signifikan berpengaruh terhadap terganggu atau tidaknya kegiatan sehari-hari akibat keluhan kesehatan (model *zero inflation*) adalah umur (X_2), status perkawinan (X_3) dan pendidikan (γ_4), sedangkan peubah penjelas yang signifikan berpengaruh terhadap rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan (model *count*) adalah umur (X_2) dan daerah tempat tinggal (X_6). Pada Tabel 13 juga terlihat bahwa nilai AIC pada setiap tahapan terus mengalami penurunan, mulai dari nilai 10 260.0 pada tahap pertama hingga mencapai 10 253.0 pada tahap keempat.

Tabel 14: Ringkasan penyeleksian peubah penjelas CHNB (10)

Peubah yang dieliminasi		Nilai AIC	Parameter yang signifikan
Model Zero Inflation	Model Count		
-	-	8497.3	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6	X_3	8493.4	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6 dan X_1	X_3 dan X_4	8490.8	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_3, X_4 dan X_1	8489.5	1/k, γ_2, γ_3, γ_4, β_2, β_6
X_6, X_1 dan X_5	X_3, X_4, X_1 dan X_5	8488.3	1/k, γ_2, γ_3, γ_4, β_2, β_6

Sama dengan proses penyeleksian peubah penjelas pada ZINB, CZINB (10) dan HNB, proses penyeleksian peubah penjelas pada pemodelan CHNB (10) dilakukan dalam 4 tahap. Hasilnya didapat bahwa peubah yang signifikan berpengaruh terhadap terganggu atau tidaknya kegiatan sehari-hari akibat keluhan kesehatan (model *zero inflation*) adalah umur (X_2), status perkawinan (X_3) dan pendidikan (γ_4), sedangkan peubah penjelas yang signifikan berpengaruh terhadap rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan (model *count*) adalah umur (X_2) dan daerah tempat tinggal (X_6). Pada Tabel 14 juga terlihat bahwa nilai AIC pada setiap tahapan pada pemodelan CHNB (10) juga mengalami penurunan hingga mencapai 8 488.3 pada tahap keempat. Hasil akhir proses pemodelan, meliputi penyeleksian peubah penjelas pada ketiga model regresi tersebut disajikan pada Tabel 15 berikut.

Tabel 15: Pendugaan parameter pada model regresi CZINB, HNB, dan CHNB.

Parameter	CZINB (10)		HNB		CHNB (10)	
	Nilai Dugaan	Nilai-P	Nilai Dugaan	Nilai-P	Nilai Dugaan	Nilai-P
Model *Zero Inflation*						
γ_0	2.9260	<.0001	3.0033	<.0001	3.0033	<.0001
γ_2	-0.0346	<.0001	-0.0350	<.0001	-0.0350	<.0001
γ_3	-0.1943	0.0579	-0.1891	0.0608	-0.1892	0.0608
γ_4	0.2826	0.0017	0.2746	0.0021	0.2746	0.0021
Model *Count*						
β_0	0.9794	<.0001	0.5526	0.0005	0.9808	<.0001
β_2	0.0059	0.0010	0.0190	<.0001	0.0059	0.0010
β_6	0.0873	0.0465	0.1361	0.0560	0.0865	0.0496
1/k	0.1044	<.0001	0.9619	<.0001	0.1044	<.0001
AIC	8 488.2		10 253.0		8 488.3	

Pemilihan Model Terbaik

Dari pemodelan dengan berbagai model regresi yang sesuai dengan karakteristik data jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017, didapat nilai *Akaike's Information Criterion* (AIC) untuk setiap model sebagai berikut:

Tabel 16: Nilai AIC

Model Regresi	Nilai AIC
ZINB	10 253.0
CZINB (10)	8 488.2
HNB	10 253.0
CHNB (10)	8 488.3

Pada Tabel 16 diketahui bahwa model regresi ZINB memiliki nilai AIC yang sama dengan HNB, yaitu sebesar 10 253.0. Model regresi CZINB (10) dan CHNB (10) juga memiliki nilai AIC yang relatif sama, yaitu sebesar 8 488.2 dan 8 488.3. Hal ini menunjukkan bahwa baik itu ZINB dan HNB maupun CZINB (10) dan CHNB (10) memiliki performa yang sama pada data jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017.

Hal lain yang dapat kita simpulkan dari tabel tersebut yaitu penyensoran pada data memiliki hasil pemodelan yang lebih baik dalam menduga peubah respon yang diteliti. Nilai AIC pada model regresi CZINB (10) dan CHNB (10) lebih kecil daripada ZINB dan HNB. Nilai AIC CZINB (10) dan CHNB (10) adalah sebesar 8 488.2 dan 8 488.3, sedangkan pada ZINB dan HNB adalah sama-sama sebesar 10 253.0.

Validasi terhadap hasil pemodelan juga dilakukan melalui penghitungan *Root Means Square Error Prediction* (RMSEP) pada data *testing*. RMSEP adalah metode alternatif untuk mengevaluasi teknik peramalan yang digunakan untuk mengukur tingkat akurasi hasil prakiraan suatu model. Dari Tabel 17 terlihat bahwa nilai RMSEP dari model regresi ZINB paling tinggi, diikuti oleh HNB, CHNB (10) dan yang paling kecil CZINB (10), yaitu sebesar 1.57. Hal ini menunjukkan bahwa model CZINB (10) memiliki tingkat akurasi yang paling baik dalam menduga jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017.

Tabel 17: Nilai RMSEP

Model Regresi	Nilai RMSEP
ZINB	3.87
CZINB (10)	1.57
HNB	3.81
CHNB (10)	1.58

Berdasarkan penghitungan nilai AIC dan RMSEP, model terbaik dalam menduga jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017 adalah CZINB (10), dengan model sebagai berikut:

Model *Zero Inflation*

$$\hat{\pi} = \frac{\exp\left(2.9260 - 0.0346X_2 - 0.1943X_3 + 0.2826X_4\right)}{1 + \exp\left(2.9260 - 0.0346X_2 - 0.1943X_3 + 0.2826X_4\right)}$$

Model *Count*

$$\hat{\mu} = \exp\left(0.9794 + 0.0059X_2 + 0.0873X_6\right)$$

Simpulan

Hasil penerapan model regresi ZINB, CZINB (10), HNB dan CHNB (10) pada jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017 menunjukkan bahwa penyensoran pada data menghasilkan pemodelan yang lebih baik. Hal itu terlihat pada nilai AIC dan RMSEP CZINB (10) yang lebih kecil daripada ZINB. Begitu juga nilai AIC dan RMSEP CHNB (10) yang lebih kecil daripada HNB.

Dari nilai AIC dan RMSEP keempat model terlihat bahwa performa model regresi ZINB dan HNB relatif sama, begitu juga pada CZINB (10) dan CHNB (10). Hal ini ditunjukkan dengan nilai AIC dan RMSEP ZINB dan HNB yang besarannya hampir sama, begitu juga dengan nilai AIC dan RMSEP CZINB (10) dan CHNB (20). Hal ini menyimpulkan bahwa dalam data jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan di Provinsi Gorontalo tahun 2017 ini tidak mengandung *structural zeros*.

Model regresi terbaik berdasarkan nilai AIC dan RMSEP adalah CZINB (10) dengan peubah penjelas yang signifikan berpengaruh terhadap jumlah hari terganggunya kegiatan sehari-hari akibat keluhan kesehatan adalah umur (X_2), tingkat pendidikan (X_4) dan status kepemilikan rumah (X_5) pada model *zero inflation* dan peubah penjelas umur (X_2) pada model *count.* Pada model *zero inflation* umur (X_2) dan pendidikan (γ_4) signifikan pada taraf 5%, sedangkan status perkawinan (X_3) pada taraf 10%. Pada model *count*, umur (X_2) dan daerah tempat tinggal (X_6) signifikan pada taraf 5%.

Dari hasil pemodelan dapat diinterpretasikan bahwa umur (X_2), status perkawinan (X_3) dan pendidikan (γ_4) signifikan berpengaruh terhadap terganggu atau tidaknya kegiatan sehari-hari akibat keluhan kesehatan, sedangkan umur (X_2) dan daerah tempat tinggal (X_6) terhadap rata-rata lama terganggunya kegiatan sehari-hari akibat keluhan kesehatan.

Daftar Pustaka

[BPS] Badan Pusat Statistik. (2017). Statistik Kesejahteraan Rakyat 2017. Jakarta (ID): Badan Pusat Statistik.

Cameron, A.C., Trivedi, P.K. (1998). *Regression Analysis of Count Data*. London (UK): Cambridge University Press

Coxe, S., West, S.G., Aiken, L.S. (2009). The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment*. 91(2):121-136.

Das, D., Das, A. (2017). *Statistics in Biology and Psychology*. West Bengal (IN). Academics Publisher.

Famoye, F., Wang, W. (2003). Censored Generalized Poisson Regression Model. *Computational Statistics & Data Analysis*. 46:547–560.

Frone, M. (1997). *Regression Models for Discrete and Limited Dependent Variables*. New York (US): Research Methods Forum.

Greene, W. (2005). Censored Data and Truncated Distributions. *Theoretical Econometrics*. 20(1).

Hofstetter, H., Dusseldorp, E., Zeileis, A., Schuller, A.A. (2016). Modeling Caries Experience: Advantages of the Use of the *Hurdle* Model. *Caries Res*. 50:517-526.

Hu, M.C., Pavlicova, M., Nunes, E.V. (2011). Zero-Inflated and *Hurdle* Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *The American Journal of Drug and Alcohol Abuse*. 37(1):367-375.

Lambert, D. (1992). Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*. 34:1-14. doi: 10.2307/1269547.

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

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*. 33(3):341–365.

Olsson, U., Drasgow, F., Dorans, N.J. (1982). The Polyserial Correlation Coefficient. *Psychometrika*. 47:337. http://doi.org/10.1007/BF02294164.

Pemerintah Republik Indonesia. (2009). Undang-Undang Republik Indonesia No. 36 Tahun 2009 tentang Kesehatan. Jakarta (ID): Sekretariat Negara.

Rose, C.E., Martin, S.W., Wannemuehler, K.A., Plikaytis, B.D. (2006). On The Use of Zero-Inflated and *Hurdle* Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics*. 16: 463–481. doi: 10.1080/10543400600719384.

Saffari, S.E., Adnan, R. (2011). Zero-Inflated *Negative Binomial* Regression Model with Right Censoring Count Data. *Journal of Materials Science and Engineering B*. 1:551-554.

Saffari, S.E., Robiah, A., Greene, W. (2012). *Hurdle Negative Binomial* Regression Model with Right Censored Count Data. *Journal of Statistics and Operations Research Transactions*. 36(2): 181-194.

Sumarni, C. (2009). Uji kesamaan parameter model regresi zero inflated generalized poisson diantara beberapa kelompok sosial [tesis]. Surabaya (ID): Institut Teknologi Sepuluh November.

Yang, S., Harlow, L.L., Puggioni, G., Redding, C.A. (2017). A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys. *Journal of Modern Applied Statistical Methods*. 16(1):518-543. doi: 10.22237/jmasm/1493598600.