PEMODELAN REGRESI *ZERO INFLATED POISSON* (APLIKASI PADA DATA PEKERJA SEKS KOMERSIAL DI KLINIK REPRODUKSI PUTAT JAYA SURABAYA)

Alia Lestari[1], Purhadi[2], Madu Ratna[2]
[1]Jurusan Teknik Sipil Universitas Andi Djemma Palopo
[2]Jurusan Statistika Institut Teknologi Sepuluh November Surabaya

Abstrak

Dalam menganalisis hubungan antara beberapa variabel, terdapat sejumlah fenomena dimana variabel responsnya berbentuk biner ataupun berbentuk diskrit. Fenomena dimana variabel responsnya berbentuk diskrit tapi tidak biner, biasanya dianalisis dengan Regresi Poisson. Namun demikian dalam kasus tertentu sering dihadapi suatu peristiwa yang sangat jarang terjadi atau responsnya mempunyai data nol yang sangat banyak, sehingga analisis dengan pendekatan distribusi Poisson seringkali tidak lagi memberikan kesimpulan yang tepat.

Pada penelitian ini akan dikaji suatu metode untuk mengatasi banyaknya respons bernilai nol yang telah dikembangkan oleh Lambert (1992) yaitu Regresi *Zero-Inflated* Poisson (*ZIP*). Estimasi parameter model ini menggunakan Algoritma EM dan pengujian hipotesisnya menggunakan *Likelihood Ratio Test*. Aplikasi pada data Pekerja Seks Komersial di Klinik Reproduksi Putat Jaya Surabaya menunjukkan bahwa variabel yang mempengaruhi *zero state* atau peluang y_i bernilai nol sama dengan variabel yang mempengaruhi *poisson state* atau peluang y_i berdistribusi Poisson, yaitu lamanya seorang PSK menjalani profesinya dan proporsi pemakaian kondom. Statistik Vuong yang dihasilkan menunjukkan bahwa Pemodelan Regresi *ZIP* menghasilkan model yang lebih baik daripada Regresi Poisson.

Kata kunci : Algoritma *EM*, Pekerja Seks Komersial (PSK), Penyakit Menular Seksual (PMS), Regresi Poisson, *Zero Inflated Poisson* (*ZIP*).

PENDAHULUAN

Dalam kasus tertentu sering dihadapi suatu peristiwa yang sangat jarang terjadi atau responsnya mempunyai data nol yang sangat banyak. Misalnya dalam proses produksi suatu item, argumen yang umum menyatakan bahwa ketika suatu proses *manufacturing* yang baik berada dalam kendali, banyaknya item yang cacat akan berdistribusi Poisson. Jika *mean* dari distribusi Poisson adalah λ, maka pada n sampel yang besar, akan mempunyai sekitar $ne^{-\lambda}$ item yang tidak cacat. Tetapi kadang-kadang, dari item yang tidak sempurna, terdapat lebih banyak lagi item tanpa cacat dibandingkan dengan yang

bisa diprediksi. Hal ini seringkali mengakibatkan kesalahan pada analisis. Apalagi data dalam jumlah yang banyak, terlalu banyaknya respons yang bernilai nol dan jumlah data yang sangat besar, membuat regresi Poisson kurang tepat untuk menganalisis kasus ini.

Beberapa peneliti telah mengembangkan metode untuk mengatasi masalah tersebut. Feuerverger (1979) merangkaikan peluang respons bernilai nol dengan mean dari distribusi gamma untuk memodelkan curah hujan. Farewell (1986) dan Meeker (1987) memberlakukan nilai-nilai nol tersebut sebagai data tersensor kanan (*right cencored continuous distribution*) untuk memodelkan data *survival*. Heilbron (1989) mengusulkan *zero altered Poisson* (*ZAP*) dan *negative binomial regression* secara bersama-sama untuk memodelkan data tentang resiko yang besar pada perilaku pasangan *gay*. Kemudian Lambert (1992) memperkenalkan *zero-inflated Poisson Regression* (*ZIP Regression*), dimana akronim ini merupakan modifikasi dari akronim yang diperkenalkan oleh Heilborn (*ZAP*), tetapi pengembangan yang dilakukan oleh Lambert berbeda dengan Heilborn.

Penelitian ini bermaksud untuk mengkaji Regresi *Zero-Inflated* Poisson (*ZIP*) tersebut, tentang estimasi parameter dan uji hipotesisnya, kemudian akan diaplikasikan pada data Pekerja Seks Komersial di Klinik Reproduksi Putat Jaya Surabaya.

MODEL REGRESI *ZIP*

Untuk setiap pengamatan Y_i yang saling bebas $i = 1, ..., n$, dan

$$Y_i \sim \begin{cases} 0, & \text{dengan peluang } \pi_i; \\ Poisson(\lambda_i), & \text{dengan peluang } (1-\pi_i) \end{cases} \quad (1)$$

maka :

$$\text{P}(Y = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i} & , \text{untuk } y_i = 0 \\ \dfrac{(1-\pi_i)\,e^{-\lambda_i}\lambda_i^{\,y_i}}{y_i!} & , \text{untuk } y_i > 0 \end{cases} \quad (2)$$

dengan parameter $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$ dan $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$ yang memenuhi :

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X\beta} \quad (3)$$

dan

$$\text{logit}(\boldsymbol{\pi}) = \boldsymbol{X\gamma} \quad (4)$$

dimana logit $(\pi_i) = \ln\left(\frac{\pi}{(1-\pi)}\right)$,

$\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ adalah parameter regresi yang akan ditaksir, sedangkan $\boldsymbol{x}$ adalah matriks variabel yang memuat himpunan-himpunan yang berbeda dari faktor eksperimen yang berhubungan dengan peluang pada *zero state* dan rata-rata Poisson pada *poisson state*.

Variabel-variabel yang mempengaruhi *mean* Poisson pada *poisson state* bisa sama dan bisa juga berbeda dengan variabel yang mempengaruhi peluang pada *zero state*. Pada saat variabel yang mempengaruhi *mean* Poisson pada *poisson state* dan *zero state* sama serta setiap parameter ($\boldsymbol{\lambda}$ dan $\boldsymbol{\pi}$) bukan merupakan fungsi dari yang lainnya, maka pemodelan regresi *ZIP* dilakukan sebanyak dua kali sesuai dengan banyaknya parameter pada regresi Poisson. Pada kasus lain, dimana variabel yang mempengaruhi kedua state berbeda, atau peluang dari *zero state* tidak bergantung pada variabel, sehingga matriks *x* yang berhubungan dengan *zero state* tersebut merupakan matriks kolom yang elemen-elemennya adalah 1, maka regresi *ZIP* hanya dimodelkan satu kali atau sama seperti memodelkan regresi Poisson.

Jika variabel yang mempengaruhi $\boldsymbol{\lambda}$ dan $\boldsymbol{\pi}$ sama, serta $\boldsymbol{\pi}$ merupakan fungsi dari $\boldsymbol{\lambda}$ atau sebaliknya, maka jumlah parameter yang akan ditaksir dapat dikurangi, dengan asumsi bahwa fungsi tersebut diketahui sebanyak suatu konstanta yang mendekati setengah dari jumlah parameter yang dibutuhkan untuk regresi *ZIP* dan secara nyata dapat mempercepat perhitungannya.

Tetapi, dalam banyak aplikasi hanya terdapat sedikit informasi awal tentang bagaimana $\boldsymbol{\pi}$ berhubungan dengan $\boldsymbol{\lambda}$. Sehingga untuk menaksir parameternya digunakan *link function* :

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X\beta} \quad \text{dan} \quad \text{logit}(\boldsymbol{\pi}) = -\tau \boldsymbol{X\beta} \tag{5}$$

dimana nilai sebenarnya dari parameter τ yang tidak diketahui, mengakibatkan

$$\pi_i = (1+\lambda_i^{\tau})^{-1} \tag{6}$$

Dalam persamaan pada *generalized linear models*, $\log(\boldsymbol{\lambda})$ dan $\text{logit}(\boldsymbol{\pi})$ adalah *link function* atau transformasi yang umumnya digunakan untuk melinearkan mean dari Poisson dan peluang sukses pada Bernoulli. *ZIP* model (20) kemudian akan dituliskan sebagai *ZIP*(τ).

Link function logit untuk parameter π akan simetrik disekitar nilai 0,5. Dua *link function asymetric* yang sering digunakan adalah log-log *link* yang didefinisikan sebagai :

$$\log(-\log(\pi_i)) = \tau\, \boldsymbol{x_i^T \beta} \quad \text{ekuivalen dengan } \pi_i = \exp(-\lambda_i^{\tau})$$

dan *complementary* log-log *link* yng didefinisikan sebagai :

$$\log(-\log(1-\pi_i)) = -\tau\, \boldsymbol{x_i^T \beta} \quad \text{atau } \pi_i = 1-\exp(-\lambda_i^{-\tau})$$

ESTIMASI PARAMETER REGRESI *ZIP*

Untuk menaksir parameter Regresi *ZIP*, digunakan metode *Maksimum Likelihood Estimation* (MLE).

Dari (3) dan (4) diperoleh :

$$\lambda_i = \exp(\boldsymbol{x_i^T \beta}) \tag{7}$$

$$\pi_i = \frac{e^{x_i^T \gamma}}{1+e^{x_i^T \gamma}} \tag{8}$$

$$(1-\pi_i) = \frac{1}{1+e^{x_i^T \gamma}} \tag{9}$$

Subtitusi (7), (8), dan (9) ke (2) :

$$P(Y=y_i) = \begin{cases} \dfrac{e^{x_i^T\gamma} + \left(\exp\left(-e^{x_i^T\beta}\right)\right)}{1+e^{x_i^T\gamma}} & \text{untuk } y_i = 0 \\ \dfrac{\dfrac{1}{1+e^{x_i^T\gamma}}\left(\exp\left(\left(e^{x_i^T\beta}\right)+(\boldsymbol{x_i^T\beta})^{y_i}\right)\right)}{y_i!} & \text{untuk } y_i > 0 \end{cases}$$

Sehingga fungsi likelihoodnya adalah :

$$L(\gamma,\beta \mid y_i) = \begin{cases} \displaystyle\prod_{i=1}^{n} \frac{e^{x_i^T\gamma} + \left(\exp\left(-e^{x_i^T\beta}\right)\right)}{1+e^{x_i^T\gamma}} & \text{untuk } y_i = 0 \\ \dfrac{\displaystyle\prod_{i=1}^{n} \frac{1}{1+e^{x_i^T\gamma}}\left(\exp\left(\left(e^{x_i^T\beta}\right)+(\boldsymbol{x_i^T\beta})^{y_i}\right)\right)}{\displaystyle\prod_{i=1}^{n} y_i!} & \text{untuk } y_i > 0 \end{cases}$$

dan fungsi log-natural likelihoodnya adalah :

$$= \sum_{\substack{i=1 \\ y_i=0}}^{n} \ln\left(e^{x_i^T\gamma} + \exp\left(-e^{x_i^T\beta}\right)\right) - \sum_{i=1}^{n} \ln(1+e^{x_i^T\gamma}) + \sum_{\substack{i=1 \\ y_i>0}}^{n} \left(\left(y_i \boldsymbol{x}_i^T \boldsymbol{\beta}\right) - e^{x_i^T\beta}\right) - \sum_{\substack{i=1 \\ y_i>0}}^{n} \ln y_i! \quad (10)$$

Penjumlahan fungsi eksponensial suku pertama pada (10) akan menyulitkan perhitungan. karena tidak diketahuinya nilai 0 mana yang berasal dari *zero state* dan mana yang berasal dari *poisson state* membuat fungsi likelihood ini tidak dapat diselesaikan dengan metode numerik biasa. Oleh karena itu (10) biasa disebut juga *incomplete data likelihood*.

Misalkan setiap variabel Y_i berkaitan dengan variabel indikator Z_i, yaitu :

$$Z_i = \begin{cases} 1, & \text{jika } y_i \text{ berasal dari } \textit{zero state} \\ 0, & \text{jika } y_i \text{ berasal dari } \textit{poisson state} \end{cases} \quad (11)$$

Permasalahannya adalah jika nilai variabel respon y_i = 1, 2, 3, ... , n, maka nilai z_i = 1. Sedangkan jika nilai variabel respon y_i = 0, maka nilai z_i mungkin 0 mungkin juga 1. Oleh karena itu nilai z_i dianggap hilang sebagian.

Untuk mengatasi hal ini maka estimasi parameter akan dilakukan dengan algoritma EM. Langkah langkahnya adalah sebagai berikut :

1. Menentukan distribusi dari variabel Z_i :

Berdasarkan (1) dan (11), maka :

$$P(Z_i = 1) = \pi_i$$

$$P(Z_i = 0) = P(Y_i \sim Poisson(\lambda)) = 1 - \pi_i$$

Jadi $Z_i \sim$ Binomial $(1, \pi_i)$, E(Z_i) = π_i , Var (Z_i) = $\pi_i(1-\pi_i)$

2. Membentuk distribusi gabungan antara $\boldsymbol{y}$ dan $\boldsymbol{z}$, yaitu

$$f(y, z \mid \pi, \lambda) = f(z_i \mid 1, \pi_i)\, f(y_i \mid z_i, \lambda_i)$$

$$= (1-\pi_i)^{1-z_i} (\pi_i)^{z_i} \left(\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right)^{1-z_i} \quad (12)$$

Subtitusi kembali (7), (8), dan (9) ke (12), diperoleh

$$f(y, z \mid \beta, \gamma) = \left(\frac{1}{1+e^{x_i^T\gamma}}\right)(e^{x_i^T\gamma})^{z_i} \left(\frac{\exp(-e^{x_i^T\beta})\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})^{y_i}}{y_i!}\right)^{1-z_i}$$

Sehingga diperoleh fungsi likelihood :

$$L(\beta,\gamma \mid y,z)=\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{x_i^T\gamma}}\right)(e^{x_i^T\gamma})^{z_i}\left(\frac{\exp(-e^{x_i^T\beta})\exp(\boldsymbol{x_i^T\beta})^{y_i}}{y_i!}\right)^{1-z_i}\right]$$

dan log-natural likelihoodnya:

$$\begin{aligned}\ln L(\beta,\gamma \mid y,z)=&\sum_{i=1}^{n}\left(z_i\boldsymbol{x_i^T\gamma}-\ln(1+e^{x_i^T\gamma})\right)+\sum_{i=1}^{n}(1-z_i)\left(y_i\boldsymbol{x_i^T\beta}-e^{x_i^T\beta}\right)+\\&-\sum_{i=1}^{n}(1-z_i)\ln(y_i!)\end{aligned} \tag{13}$$

Fungsi likelihood ini biasa disebut *complete data likelihood*. Fungsi inilah yang akan dimaksimumkan menggunakan algoritma EM, dimana vektor parameter $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ dapat diestimasi secara terpisah, dengan menuliskan (13) menjadi :

$$\ln L(\beta,\gamma,y,z)=\ln L(\beta,y,z)+\ln L(\gamma,y,z)+\sum_{i=1}^{n}z_i\ln(y_i!)$$

dengan $\ln L(\gamma;y,z)=\sum_{i=1}^{n}\left(z_i\boldsymbol{x_i^T\gamma}-\ln(1+e^{x_i^T\gamma})\right)$ (14)

dan $\ln L(\beta;y,z)=\sum_{i=1}^{n}(1-z_i)\left(y_i\boldsymbol{x_i^T\beta}-e^{x_i^T\beta}\right)$ (15)

Sedangkan $\sum_{i=1}^{n}z_i\ln(y_i!)$ dapat diabaikan karena tidak mengandung vektor parameter $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$.

3. Selanjutnya dilakukan tahap Ekspektasi dan Maksimalisasi:
   a. Tahap Ekspektasi :

      Ganti variabel Z_i dengan $Z_i^{(k)}$, dengan

$$\begin{aligned}Z_i^{(k)}&=E\left(Z_i \mid y_i,\gamma^{(k)},\beta^{(k)}\right)\\&=P(Z_i=1\mid y_i,\gamma^{(k)},\beta^{(k)})\\&=\begin{cases}P(Z_i=1\mid y_i=0,\gamma^{(k)},\beta^{(k)}), & \text{jika } y_i=0\\0, & \text{jika } y_i>0\end{cases}\end{aligned}$$

$$P(Z_i=1\mid y_i=0,\gamma^{(k)},\beta^{(k)})=\frac{1}{1+\exp\left(-\boldsymbol{x_i^T\gamma}-\exp(\boldsymbol{x_i^T\beta})\right)}$$

      Sehingga (14) dan (15) menjadi :

$$\ln L(\gamma; y, z^{(k)}) = \sum_{i=1}^{n} \left(z_i^{(k)} \boldsymbol{x_i^T \gamma} - \ln(1 + \exp(\boldsymbol{x_i^T \gamma})) \right) \quad (16)$$

dan

$$\ln L(\beta; y, z^{(k)}) = \sum_{i=1}^{n} (1 - z_i^{(k)}) \left(y_i \boldsymbol{x_i^T \beta} - e^{x_i^T \beta} \right) \quad (17)$$

b. Tahap Maksimalisasi, terdiri dari :

- Maksimalisasi untuk $\boldsymbol{\beta}$, dengan menghitung $\boldsymbol{\beta}^{(k+1)}$ yang diperoleh dari memaksimumkan (17) dengan metode Newton Raphson.
- Maksimalisasi untuk $\boldsymbol{\gamma}$, dengan memaksimalkan (16), dimana untuk setiap $y_i > 0$, nilai $z_i^{(k)}$ akan sama dengan 0, sehingga (16) dapat juga ditulis:

$$\ln L(\gamma; y, z^{(k)}) = \sum_{y_i=0} z_i^{(k)} x_i^T \gamma - \sum_{y_i=0} z_i^{(k)} \ln(1 + e^{x_i^T \gamma}) +$$
$$- \sum_{i=1}^{n} \left(1 - z_i^{(k)}\right) \ln(1 + e^{x_i^T \gamma}) \ln(1 + e^{x_i^T \gamma})$$

Misalkan nilai y_i sampai ke in_0 adalah 0, atau dapat dituliskan :

$y_{i1}, y_{i2}, \ldots, y_{in_0} = 0$

Definisikan $y_*^T = (y_1, \ldots y_n, y_{i1}, \ldots, y_{in_0})$

$$x_{i*}^T = (x_1^T, \ldots x_n^T, x_{i1}^T, \ldots, x_{in_0}^T)$$

$$P_*^T = (p_1^T, \ldots p_n^T, p_{i1}^T, \ldots, p_{in_0}^T)$$

Definisikan juga matrix diagonal $W^{(k)}$ dengan elemen diagonal :

$$w_*^T = (1 - z_i^{(k)}, \ldots 1 - z_n^{(k)}, 1 - z_{i1}^{(k)}, \ldots, z_{in_0}^{(k)})^T$$

Sehingga (16) dapat ditulis kembali sebagai :

$$\ln L(\gamma; y, z^{(k)}) = \sum_{i=1}^{n+n_0} y_{*i} w_i^{(k)} x_{i*}^T \gamma - \sum_{i=1}^{n+n_0} w_i^{(k)} \ln(1 + e^{x_i^T \gamma}) \quad (18)$$

sedangkan vektor gradien $\boldsymbol{g^T} = x_{i*}^T W^{(k)} (y_* - P_*) = 0$

dan matrix Hessian adalah $\boldsymbol{H} = X_*^T W^{(k)} Q_* X_*$ dimana Q_* adalah matriks diagonal dengan $P_*(1 - P_*)$ sebagai elemen diagonalnya.

4. Ganti $\beta^{(k)}$ dan $\gamma^{(k)}$ dengan $\beta^{(k+1)}$ dan $\gamma^{(k+1)}$, kemudian lakukan kembali tahap ekspektasi.
5. Tahap ke-3 dan ke-4 ini dilakukan berulang-ulang sampai diperoleh penaksir parameter yang konvergen.

PENGUJIAN PARAMETER MODEL REGRESI *ZIP*

Untuk menguji kelayakan model yang diperoleh dari estimasi parameter, dilakukan pengujian parameter model Regresi *ZIP* dengan menguji hipotesis-hipotesis berikut menggunakan metode *Likelihood Ratio Test*.

1. H_0 : $\beta_1 = \beta_2 = \ldots = \beta_k = \gamma_1 = \gamma_2 = \ldots = \gamma_k = 0$

   H_1 : Paling sedikit ada satu $\beta_i \neq 0$ atau $\gamma_i \neq 0$

   - Himpunan parameter dibawah populasi (Ω): $\Omega = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$
   - Himpunan parameter jika H_0 benar (ω) : $\omega = \{\beta_0, \gamma_0\}$
   - Fungsi likelihood di bawah populasi ($L(\Omega)$):

$$L(\Omega) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) = = \prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\boldsymbol{x}_i^T \boldsymbol{\gamma})} \right) (\exp(\boldsymbol{x}_i^T \boldsymbol{\gamma}))^{z_i} \left(\frac{\exp(-e^{x_i^T \boldsymbol{\beta}}) (\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}))^{y_i}}{y_i!} \right)^{1-z_i} \right]$$

   - Fungsi likelihood jika H_0 benar : ($L(\omega)$)

$$L(\omega) = \prod_{i=1}^{n} f(y_i; \beta_0, \gamma_0) = \prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\gamma_0)} \right) (\exp(\gamma_0))^{z_i} \left(\frac{\exp(-e^{\beta_0}) (\exp(\beta_0))^{y_i}}{y_i!} \right)^{1-z_i} \right]$$

   - $$\frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\hat{\gamma}_0)} \right) (\exp(\hat{\gamma}_0))^{z_i} \left(\exp(-e^{\hat{\beta}_0}) (\exp(\hat{\beta}_0))^{y_i} \right)^{1-z_i} \right]}{\prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\gamma}})} \right) (\exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\gamma}}))^{z_i} \left(\exp(-e^{x_i^T \hat{\boldsymbol{\beta}}}) (\exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}))^{y_i} \right)^{1-z_i} \right]}$$

   - Nilai $\hat{\boldsymbol{\beta}}$ dan $\hat{\boldsymbol{\gamma}}$ diperoleh dari bagian 3
   - Tolak H_0 jika $\frac{L(\hat{\omega})}{L(\hat{\Omega})} < \mu_0 < 1$, dimana $0 < \mu_0 < 1$
   - $G = -2 \ln \left[\frac{L(\hat{\omega})}{L(\hat{\Omega})} \right]$ yang berdistribusi χ_v^2, tolak H_0 jika $G_{\text{hitung}} > \chi_{\alpha,v}^2$

2. H_0 : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

   H_1 : Paling sedikit ada satu $\beta_i \neq 0$

   - Himpunan parameter dibawah populasi (Ω): $\Omega = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$
   - Himpunan parameter jika H_0 benar (ω) : $\omega = \{\beta_0\}$
   - Fungsi likelihood di bawah populasi ($L(\Omega)$):

     $$L(\Omega) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) \ \exp(\boldsymbol{x_i^T \beta})$$

   - Fungsi likelihood di bawah H_0 ($L(\omega)$)

     $$L(\omega) = \prod_{i=1}^{n} f(y_i; \beta_0) = \prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T \gamma})} \right) \left(\exp(\boldsymbol{x_i^T \gamma}) \right)^{z_i} \left(\frac{\exp(-e^{x_i^T \beta}) \left(\exp(\boldsymbol{x_i^T \beta}) \right)^{y_i}}{y_i!} \right)^{1-z_i} \right]$$

   - $$\frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\prod_{i=1}^{n} \left(\exp(-e^{\hat{\beta}_0}) \left(\exp(\hat{\beta}_0) \right)^{y_i} \right)^{1-z_i}}{\prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T \hat{\gamma}})} \right) \left(\exp(\boldsymbol{x_i^T \hat{\gamma}}) \right)^{z_i} \left(\exp(-e^{x_i^T \hat{\beta}}) \left(\exp(\boldsymbol{x_i^T \hat{\beta}}) \right)^{y_i} \right)^{1-z_i} \right]}$$

   - Nilai $\hat{\boldsymbol{\beta}}$ dan $\hat{\boldsymbol{\gamma}}$ diperoleh dari bagian 3
   - Tolak H_0 jika $\frac{L(\hat{\omega})}{L(\hat{\Omega})} < \mu_0 < 1$, dimana $0 < \mu_0 < 1$
   - $G = -2 \ln \left[\frac{L(\hat{\omega})}{L(\hat{\Omega})} \right]$ yang berdistribusi χ_v^2, Tolak H_0 jika $G_{\text{hitung}} > \chi_{\alpha,v}^2$

3. H_0 : $\gamma_1 = \gamma_2 = \ldots = \gamma_k = 0$

   H_1 : Paling sedikit ada satu $\gamma_i \neq 0$

   - Himpunan parameter dibawah populasi (Ω): $\Omega = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$
   - Himpunan parameter di bawah H_0 (ω) : $\omega = \{\gamma_0\}$
   - Fungsi likelihood di bawah populasi ($L(\Omega)$):

     $$L(\Omega) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T \gamma})} \right) \left(\exp(\boldsymbol{x_i^T \gamma}) \right)^{z_i} \left(\frac{\exp(-e^{x_i^T \beta}) \left(\exp(\boldsymbol{x_i^T \beta}) \right)^{y_i}}{y_i!} \right)^{1-z_i} \right]$$

   - Fungsi likelihood di bawah H_0 ($L(\omega)$)

$$L(\omega)=\prod_{i=1}^{n} f(y_i;\gamma_0)=\prod_{i=1}^{n}\left(\frac{1}{1+\exp(\gamma_0)}\right)(\exp(\gamma_0))^{z_i}$$

- $$\frac{L(\hat{\omega})}{L(\hat{\Omega})}=\frac{\prod_{i=1}^{n}\left(\frac{1}{1+\exp(\hat{\gamma}_0)}\right)(\exp(\hat{\gamma}_0))^{z_i}}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T}\hat{\boldsymbol{\gamma}})}\right)(\exp(\boldsymbol{x_i^T}\hat{\boldsymbol{\gamma}})^{z_i}\left(\frac{\exp(-e^{x_i^T\hat{\boldsymbol{\beta}}})\left(\exp(\boldsymbol{x_i^T}\hat{\boldsymbol{\beta}})\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}$$

- Nilai $\hat{\boldsymbol{\beta}}$ dan $\hat{\boldsymbol{\gamma}}$ diperoleh dari bagian 3
- Tolak H_0 jika $\frac{L(\hat{\omega})}{L(\hat{\Omega})}<\mu_0<1$, dimana $0<\mu_0<1$
- $G=-2\ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$ yang berdistribusi χ_v^2, tolak H_0 jika $G_{\text{hitung}} > \chi_{\alpha,v}^2$

4. H_0 : $\beta_i = 0$

   H_1 : $\beta_i \neq 0$

   - Himpunan parameter dibawah populasi (Ω): $\Omega=\{\boldsymbol{\beta},\boldsymbol{\gamma}\}$
   - Himpunan parameter di bawah H_0 (ω) : $\omega=\{\beta_i\}$
   - Fungsi likelihood di bawah populasi ($L(\Omega)$):

$$L(\Omega)=\prod_{i=1}^{n} f(y_i;\boldsymbol{\beta},\boldsymbol{\gamma})=\prod_{i=1}^{n}\left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T}\boldsymbol{\gamma})}\right)(\exp(\boldsymbol{x_i^T}\boldsymbol{\gamma}))^{z_i}\left(\frac{\exp(-e^{x_i^T\boldsymbol{\beta}})\left(\exp(\boldsymbol{x_i^T}\boldsymbol{\beta})\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]$$

   - Fungsi likelihood di bawah H_0 ($L(\omega)$)

$$L(\omega)=\prod_{i=1}^{n} f(y_i;\beta_i)=\prod_{i=1}^{n}\left(\frac{\exp(-e^{x_i\beta_i})(\exp(x_i\beta_i))^{y_i}}{y_i!}\right)^{1-z_i}$$

- $$\frac{L(\hat{\omega})}{L(\hat{\Omega})}=\frac{\prod_{i=1}^{n}\left(\exp(-e^{x_i\hat{\beta}_i})\left(\exp(x_i\hat{\beta}_i)\right)^{y_i}\right)^{1-z_i}}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T}\hat{\boldsymbol{\gamma}})}\right)\left(\exp(\boldsymbol{x_i^T}\hat{\boldsymbol{\gamma}})\right)^{z_i}\left(\exp(-e^{x_i^T\hat{\boldsymbol{\beta}}})\left(\exp(\boldsymbol{x_i^T}\hat{\boldsymbol{\beta}})\right)^{y_i}\right)^{1-z_i}\right]}$$

   - Nilai $\hat{\boldsymbol{\beta}}$ dan $\hat{\boldsymbol{\gamma}}$ diperoleh dari bagian 3
   - Tolak H_0 jika $\frac{L(\hat{\omega})}{L(\hat{\Omega})}<\mu_0<1$, dimana $0<\mu_0<1$

- $G = -2\ln\left[\dfrac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$ yang berdistribusi χ_v^2, tolak H_0 jika $G_{\text{hitung}} > \chi_{\alpha,v}^2$

5. H_0 : $\gamma_i = 0$

   H_1 : $\gamma_i \neq 0$

   - Himpunan parameter dibawah populasi (Ω): $\Omega = \{\boldsymbol{\beta},\boldsymbol{\gamma}\}$
   - Himpunan parameter di bawah H_0 (ω) : $\omega = \{\gamma_i\}$
   - Fungsi likelihood di bawah populasi ($L(\Omega)$):

$$L(\Omega) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T\gamma})}\right)\left(\exp(\boldsymbol{x_i^T\gamma})\right)^{z_i}\left(\frac{\exp(-e^{\boldsymbol{x_i^T\beta}})\left(\exp(\boldsymbol{x_i^T\beta})\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]$$

   - Fungsi likelihood di bawah H_0 ($L(\omega)$)

$$L(\omega) = \prod_{i=1}^{n} f(y_i;\gamma_i) = \prod_{i=1}^{n}\left(\frac{1}{1+\exp(x_i\gamma_i)}\right)\left(\exp(x_i\gamma_i)\right)^{z_i}$$

   - $$\frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\prod_{i=1}^{n}\left(\frac{1}{1+\exp(x_i\hat{\gamma}_i)}\right)\left(\exp(x_i\hat{\gamma}_i)\right)^{z_i}}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+\exp(\boldsymbol{x_i^T\hat{\gamma}})}\right)\left(\exp(\boldsymbol{x_i^T\hat{\gamma}})\right)^{z_i}\left(\frac{\exp(-e^{\boldsymbol{x_i^T\hat{\beta}}})\left(\exp(\boldsymbol{x_i^T\hat{\beta}})\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}$$

   - Nilai $\hat{\boldsymbol{\beta}}$ dan $\hat{\boldsymbol{\gamma}}$ diperoleh dari bagian 3
   - Tolak H_0 jika $\dfrac{L(\hat{\omega})}{L(\hat{\Omega})} < \mu_0 < 1$, dimana $0 < \mu_0 < 1$
   - $G = -2\ln\left[\dfrac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$ yang berdistribusi χ_v^2, tolak H_0 jika $G_{\text{hitung}} > \chi_{\alpha,v}^2$

APLIKASI REGRESI *ZIP* PADA DATA PSK DI KLINIK REPRODUKSI PUTAT JAYA SURABAYA.

Pemodelan Regresi *ZIP* akan diaplikasikan pada data PSK yang diperoleh dari Klinik Reproduksi Putat Jaya, Jl. Kupang Gunung IV/25 Surabaya. Sebagai gambaran awal, deskripsi statistik dari data dapat dilihat pada tabel berikut :

Tabel 4.1. Deskripsi Statistik Data PSK di Klinik Reproduksi Putat Jaya Surabaya

Variable	Mean	StDev	Minimum	Maximum
Y	0,0507	0,2512	0	2
X_1	6,576	1,671	1	12
X_2	3,238	1,905	0,518	9,604
X_3	28,747	3,833	20,9	39
X_4	0,26674	0,10117	0	0,54

Terlihat bahwa banyaknya PSK yang menderita PMS pada setiap rumah maksimal berjumlah 2 orang. Lama pendidikan yang pernah ditempuh oleh PSK maksimal 12 tahun atau tamat SMA, dan terdapat pula PSK yang hanya pernah menempuh pendidikan selama 1 tahun. Rata-rata PSK telah menjalani profesinya selama 3,238 tahun, atau sekitar 38 bulan. Usia PSK berkisar antara 20,9 sampai 39 tahun, dengan rata-rata usia 28 tahun. Sedangkan rata-rata proporsi pemakaian kondom adalah 0,267%, dengan proporsi pemakaian paling banyak adalah sekitar 0,54%.

Setelah dilakukan pemodelan dengan Regresi *ZIP* menggunakan paket program SAS 9.1, diperoleh hasil sebagai berikut :

Pengujian parameter secara serentak dengan hipotesis :

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_4 = \gamma_1 = \gamma_2 = \ldots = \gamma_4 = 0$

H_1 : Paling sedikit ada satu $\beta_i \neq 0$ atau $\gamma_i \neq 0$

menghasilkan nilai G_{hitung} = 42,7, sedangkan G = 2,71 ($\chi^2_{0,1;1}$) memberikan alasan yang cukup kuat untuk menolak H_0, yang berarti bahwa setiap variabel bebas memberikan efek yang berbeda terhadap variabel respon, oleh karena itu diperlukan pengujian secara parsial untuk mengetahui efek yang diberikan masing-masing variabel bebas tersebut.

Pengujian parameter secara parsial dengan hipotesis :

H_0 : $\beta_i = 0$

H_1 : $\beta_i \neq 0$

Hasil estimasi parameter dapat dilihat pada tabel berikut:

Tabel 4.2. Nilai Estimasi Parameter Model Regresi *ZIP*

Parameter	Estimasi	SE	DF	t	P-Value	Lower	Upper
γ_0	-1,1168	2,8186	138	-0,4	0,6926	-6,69	4,4564
γ_1	2,9013	3,1892	138	0,91	0,3646	-3,4047	9,2072
γ_2	-3,088	1,7825	138	-1,73	0,0854	-6,6125	0,4365
γ_4	-2,9959	1,7043	138	-1,76	0,081	-6,3658	0,3741
β_0	-2,0961	0,9479	138	-2,21	0,0287	-3,9703	-0,2218
β_1	2,3526	1,2032	138	1,96	0,0526	0,02648	4,7317
β_2	-1,1848	0,6197	138	-1,91	0,0579	-2,4101	0,04046
β_4	-0,8069	0,8971	138	-0,9	0,37	-2,5807	0,967

Pada tabel dapat dilihat bahwa parameter yang signifikan adalah β_1 dan β_2. Hal ini terlihat pada nilai p-value masing-masing 0,0526 dan 0,0579 yang lebih kecil dari α = 0,1. Penggunaan α = 0,1 karena penelitian ini merupakan penelitian sosial.

Selain itu, pengujian parameter secara parsial dengan hipotesis :

H_0 : $\gamma_i = 0$

H_1 : $\gamma_i \neq 0$

menghasilkan γ_2 dan γ_4 sebagai parameter yang signifikan pada α = 0,1, terlihat pada nilai p-value masing-masing 0,0854 dan 0,081. Tetapi karena nilai estimasi parameter berubah ketika parameter yang tidak signifikan dikeluarkan dari model, maka semua parameter dimasukkan ke dalam model. Sehingga model Regresi *ZIP*-nya adalah :

$$\log(\lambda_i) = -2{,}0961 + 2{,}3526X_{1i} - 1{,}1848X_{2i} - 0{,}8069X_{4i}$$

dan

$$\text{logit}(\pi_i) = -0{,}1168 - 2{,}9013X_{1i} + 3{,}088X_{2i} + 2{,}9959X_{4i}$$

Dimana X_1 menyatakan lama pendidikan (tahun), X_2 adalah lama menjadi PSK (tahun), dan X_4 adalah proporsi pemakaian kondom(%).

Model yang dihasilkan memperlihatkan bahwa variabel yang mempengaruhi *zero state* sama dengan variabel yang mempengaruhi *poisson state*. Model logit menjelaskan peluang respon (y_i) bernilai nol dipengaruhi oleh lamanya seorang PSK menjalani

profesinya dan proporsi pemakaian kondom, sedangkan lama pendidikan PSK meskipun dimasukkan ke dalam model tetapi pengaruhnya tidak signifikan. Model log menjelaskan bahwa semakin tinggi proporsi pemakaian kondom dan semakin lama seorang PSK menjalani profesinya, akan menurunkan rata-rata jumlah PSK penderita PMS *Trikomoniasis Vaginalis* di setiap rumah. Model ini juga dapat diartikan bahwa semakin lama seorang PSK menjalani profesinya maka kesadaran PSK dalam memakai kondom akan semakin tinggi sehingga peluang PSK tersebut tertular PMS *Trikomoniasis Vaginalis* semakin kecil.

PERBANDINGAN PEMODELAN REGRESI *ZIP* DENGAN REGRESI POISSON

Untuk membandingkan model yang diperoleh dengan Regresi *ZIP* dengan Regresi Poisson, Vuong (1989) dalam Greene (2000) mengusulkan statistik uji untuk membandingkan model *zero inflated* dengan model awalnya, dengan hipotesis:

H_0 : Regresi *ZIP* cocok untuk data ini

H_1 : Regresi *ZIP* tidak cocok untuk data ini

Misalkan $m_i = \log\left(\frac{f_1(y_i \mid x_i)}{f_2(y_i \mid x_i)}\right)$

dimana $f_1(y_i \mid x_i)$ adalah fungsi kepadatan peluang dari model regresi *ZIP* dan $f_2(y_i \mid x_i)$ adalah fungsi kepadatan peluang dari model regresi Poisson, maka statistik ujinya adalah :

$$v = \frac{\sqrt{n}\left[(1/n)\sum_{i=1}^{n} m_i\right]}{\sqrt{(1/n)\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \frac{\sqrt{n}(\overline{m})}{S_m} \qquad (18)$$

dimana $\overline{m}$ adalah *mean*, S_m adalah standar deviasi, dan *n* adalah jumlah sampel. Nilai *v* berdistribusi normal standar secara asimptotik. Jika $|v|$ lebih kecil 1,645 (t *test* dengan *interval convidence* 90 %), maka tidak cukup alasan untuk memodelkan data tersebut dengan Regresi *ZIP* (H_0 ditolak)(Greene, 2000).

. Setelah dilakukan analisis, diperoleh nilai Statistik Vuong = 1,8743465871 yang lebih besar dari t_{hitung} = 1,645. Hal ini menunjukkan pemodelan Regresi *ZIP* lebih baik daripada pemodelan Regresi Poisson, untuk α = 0,1%.

SIMPULAN

Estimasi parameter model Regresi *ZIP* menghasilkan estimator yang berbentuk *implisit*, sehingga diperlukan prosedur iteratif untuk memperoleh estimasi parameternya. Metode yang digunakan adalah Algoritma EM, yang terdiri dari dua tahap yaitu tahap ekspektasi dan tahap maksimalisasi. Pada tahap maksimalisasi digunakan metode Newton Rapson untuk memaksimalkan fungsi Likelihood yang diperoleh pada tahap Ekspektasi. Statistik uji untuk pengujian secara serentak dan parsial diperoleh dengan metode *Likelihood Ratio Test*.

Pemodelan Regresi *ZIP* memberikan hasil bahwa pada data PSK di Klinik Reproduksi Putat Jaya Surabaya, variabel yang mempengaruhi *zero state* atau peluang y_i bernilai nol sama dengan variabel yang mempengaruhi *poisson state* atau peluang y_i berdistribusi Poisson, yaitu lamanya seorang PSK menjalani profesinya dan proporsi pemakaian kondom. Semakin lama seorang PSK menjalani profesinya dan semakin tinggi proporsi pemakaian kondomnya akan menurunkan rata-rata jumlah PSK yang mengidap PMS *Trikomoniasis Vaginalis* di setiap rumah bordil di Kelurahan Putat Jaya Surabaya. Statistik Vuong yang dihasilkan dari perbandingan antara pemodelan dengan Regresi *ZIP* dan Regresi Poisson menunjukkan bahwa pemodelan Regresi *ZIP* pada data PSK di Klinik Reproduksi Putat Jaya Surabaya lebih baik daripada pemodelan dengan Regresi Poisson.

DAFTAR PUSTAKA

Greene, W.H. 2000. *Econometrics Analysis* 4th Edition. London: Prentice Hall.

Dempster, A.P., Laird, N.M., & Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38.

Fauzi, A., Lucianawaty, M., Hanifah, L., & Bernadette, N. 2007. Penyakit Menular Seksual dan HIV/AIDS,. yminti@mweb.co.id.

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

Maharani, R. 2004. *Kajian Penyakit Menular Seksual Pekerja Seks Komersil (PSK) dengan Model Regresi Poisson (Studi Kasus di Lokalisasi Dolly Jarak, Kelurahan*

Putat Jaya Kecamatan Sawahan Kodya Surabaya). Skripsi. Surabaya: Jurusan Statistika FMIPA ITS.

McCullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models.* 2nd Edition. London: Chapman & Hall.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*. New York: PWS Kent Publishing Company.

Vuong, H.Q. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrics*, 57 (2), 307-333.