Jurnal Matematika UNAND
Vol. **1** No. **1** Hal. 71 – 76
ISSN : 2303–291X
©Jurusan Matematika FMIPA UNAND

Regresi Poisson dan Penerapannya Untuk Memodelkan Hubungan Usia dan Perilaku Merokok Terhadap Jumlah Kematian Penderita Penyakit Kanker Paru-Paru

IIN SUNDARI

Program Studi Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas Padang,
Kampus UNAND Limau Manis Padang, Indonesia
iinsundari@ymail.com

Abstrak. Regresi Poisson termasuk dalam regresi nonlinier yang variabel responnya dimodelkan sebagai distribusi Poisson. Variabel respon dalam regresi Poisson berasal dari data cacahan yang kejadiannya jarang terjadi. Dalam model regresi Poisson terdapat asumsi yang harus dipenuhi yakni equidispersi yang berarti nilai variansi dari variabel respon Y yang diberikan oleh $X = x$ harus sama dengan nilai meannya yaitu $Var(Y|x) = E(Y|x) = \mu$. Untuk menguji kesesuaian model regresi Poisson digunakan statistik uji devians dengan rumus

$$G = -2\ln\left[\frac{L(y;\hat{\beta})}{L(y;\hat{\mu})}\right].$$

Penelitian ini membahas hubungan usia dan perilaku merokok terhadap jumlah kematian penderita penyakit paru-paru dengan menggunakan model regresi Poisson. Terdapat 9 kategori usia dan 4 kategori perilaku merokok. Diantara 13 parameter tersebut terdapat 10 parameter yang signifikan yakni usia 40-44, usia 45-49 tahun, 50-54 tahun, 55-59 tahun, 60-64 tahun, 65-69 tahun, 70-74 tahun, 75-79 tahun, perokok sedang, dan perokok berat.
***Kata kunci**: Data cacahan, Regresi Poisson, Devians.*

1 Pendahuluan

Analisis regresi adalah suatu metode yang digunakan untuk menganalisa hubungan antara variabel respon dan variabel prediktor. Pada umumnya analisis regresi digunakan untuk menganalisa variabel respon yang merupakan data kontinu dan mengikuti distribusi normal. Namun dalam beberapa aplikasinya, variabel respon yang akan dianalisisa dapat berupa data diskrit.

Salah satu contoh dimana variabel responnya diskrit adalah banyaknya kejadian yang jarang terjadi *(rare event)*. Misalkan banyaknya kecelakaan mobil setiap bulan, banyaknya hujan badai setiap tahun, banyaknya kebakaran hutan setiap tahun, dan banyaknya penderita kanker paru-paru yang meninggal setiap tahun. Salah satu model regresi yang dapat digunakan untuk menjelaskan hubungan antara variabel respon Y yang berupa data diskrit dengan variabel prediktor X berupa data diskrit, kontinu, kategorik atau campuran adalah model regresi Poisson.

Model regresi Poisson berasal dari distribusi Poisson dengan parameter intensitas μ yang bergantung pada variabel prediktor. Dalam model regresi Poisson terdapat beberapa asumsi yang harus dipenuhi yakni equidispersi yang berarti nilai variansi dari variabel respon Y yang diberikan oleh $X = x$ harus sama dengan nilai *mean*nya yaitu $Var(Y|x) = E(Y|x) = \mu$.

2 Regresi Poisson

Dalam analisis regresi, hubungan antara variabel respons dengan variabel-variabel prediktornya dinyatakan dalam suatu model yang disebut sebagai model regresi. Model tersebut menghubungkan variabel prediktor dan respons melalui parameter yang dinamakan sebagai parameter regresi dan dinotasikan dengan β. Salah satu model regresi yang dapat menggambarkan hubungan antara variabel respons Y yang berupa data diskrit dengan variabel prediktor X adalah regresi Poisson. Variabel respon dalam regresi Poisson berasal dari data cacahan yang diharapkan jarang terjadi.

Jika μ_i adalah rata-rata jumlah kejadian dalam periode t_i dan diasumsikan μ_i tidak berubah dari titik data ke titik data secara bebas maka μ_i dapat dimodelkan sebagai fungsi dari k variabel prediktor. Dalam *Generalized Linier Model* (GLM), terdapat sebuah fungsi g yang menghubungkan rata-rata dari variabel responnya dengan sebuah prediktor linier, yaitu:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

fungsi g disebut fungsi penghubung (*link function*).

Pada model regresi Poisson, fungsi penghubung yang digunakan adalah fungsi penghubung log karena fungsi log menjamin bahwa nilai variabel yang diharapkan dari variabel responnya akan bernilai nonnegatif. Berikut ini adalah fungsi penghubung yang digunakan untuk model regresi Poisson.

$$\begin{aligned} \ln E(y|x) = \ln(\mu_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \\ \mu_i &= \exp(x_i^{'}\hat{\beta}) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \end{aligned}$$

sehingga model regresi Poisson dapat ditulis sebagai berikut [5].

$$y_i = t_i \exp(x_i^{'}\hat{\beta}) + \varepsilon_i. \tag{1}$$

3 Penaksiran Parameter Regresi Poisson

Metode MLE *(Maximum Likelihood Estimator)* adalah salah satu metode penaksiran parameter yang dapat digunakan untuk menaksir parameter suatu model yang diketahui distribusinya. Sebagaimana diketahui bahwa taksiran parameter melalui metode MLE adalah melakukan turunan parsial fungsi kemungkinan terhadap parameter yang akan ditaksir. Fungsi kemungkinan untuk regresi Poisson adalah sebagai berikut.

$$\begin{aligned} L(y;\hat{\beta}) &= \prod_{i=1}^{n} P(y_i;\hat{\beta}) \\ &= \prod_{i=1}^{n} \frac{e^{-t_i[\mu(x_i;\hat{\beta})]}[t_i\mu(x_i;\hat{\beta})]^{y_i}}{y_i!} \end{aligned}$$

$$L(y;\hat{\beta}) = \frac{e^{-\Sigma_{i=1}^{n} t_i \mu(x_i;\hat{\beta})} \Pi_{i=1}^{n} [t_i \mu(x_i;\hat{\beta})]^{y_i}}{\Pi_{i=1}^{n} y_i!} \tag{2}$$

Metode yang digunakan untuk menentukan penaksir kemungkinan maksimum adalah metode *Newton-Raphson*. Metode ini digunakan untuk menyelesaikan persamaan berikut:

$$\frac{\partial \ln L(y;\hat{\beta})}{\partial \beta} = 0, \tag{3}$$

dimana

$$\ln L(y;\hat{\beta}) = \sum_{i=1}^{n} y_i \ln t_i \mu(x_i;\hat{\beta}) - \sum_{i=1}^{n} t_i \mu(x_i;\hat{\beta}) - \sum_{i=1}^{n} \ln(y_i)!. \tag{4}$$

Persamaan *likelihood* untuk mencari $\hat{\beta}$ adalah sebagai berikut [5]:

$$\begin{aligned} \sum_{i=1}^{n} \left[\frac{y_i}{\mu(x_i;\hat{\beta})} \right] \left[\frac{\partial\ \mu(x_i;\hat{\beta})}{\partial\ \hat{\beta}} \right] - \sum_{i=1}^{n} [t_i] \left[\frac{\partial\ \mu(x_i;\hat{\beta})}{\partial\ \hat{\beta}} \right] = 0 \\ \sum_{i=1}^{n} \left[\frac{y_i}{\mu(x_i;\hat{\beta})} - t_i \right] \left[\frac{\partial\ \mu(x_i;\hat{\beta})}{\partial\ \hat{\beta}} \right] = 0 \end{aligned} \tag{5}$$

4 Pengujian Regresi Poisson

Untuk menguji kesesuaian model regresi, digunakan *deviance statistic*. Berikut adalah hipotesis pengujian kesesuaian model regresi Poisson.

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (artinya tidak satupun peubah bebas berpengaruh terhadap Y)

H_1 : $\beta_j \neq 0, \quad j = 1, 2, 3, \cdots, k$ (artinya paling tidak satu peubah yang berpengaruh terhadap Y)

Statistik uji yang digunakan sebagai berikut:

$$G = -2\ln \left[\frac{L(y;\hat{\beta})}{L(y;\hat{\mu})} \right] \tag{6}$$

dengan $L(y;\hat{\beta})$ adalah fungsi kemungkinan maksimum untuk model lengkap dengan melibatkan variabel prediktor dan $L(y;\hat{\mu})$ adalah fungsi kemungkinan maksimum untuk model sederhana tanpa melibatkan variabel prediktor.

Kriteria untuk pengujian ini adalah tolak H_0 pada taraf signifikansi α, jika $G > \chi^2_{(n-k-1)(\alpha)}$, dengan n adalah banyak pengamatan dan k adalah banyak parameter.

Untuk mengetahui pengaruh yang diberikan setiap variabel prediktor tersebut, dilakukan pengujian parameter dengan menggunakan hipotesis sebagai berikut.

H_0 : $\beta_j = 0$ (pengaruh variabel ke-j tidak signifikan)
H_1 : $\beta_j \neq 0$ (pengaruh variabel ke-j signifikan)

Statistik uji yang digunakan

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}. \tag{7}$$

Kriteria pengujiannya adalah Tolak H_0 jika $|t_{hit}| > t_{(\alpha/2,v)}$ dimana α adalah tingkat signifikan dan v adalah derajat bebas [3].

5 Contoh Terapan

Sebagai contoh penerapan model regresi Poisson digunakan data suatu kasus penelitian yang diambil dari [1] tentang pengaruh usia dan perilaku merokok terhadap jumlah kematian penderita penyakit kanker paru-paru. Data tersebut terdiri dari 36 pengamatan, 1 populasi dan 3 peubah yang memuat 2 peubah bebas dan 1 peubah tak bebas.

Peubah bebas pertama adalah usia penderita kanker yang terdiri dari 9 kelompok umur. Kelompok umur pertama berusia 40-44 tahun, kelompok umur ke 2 berusia 45-49 tahun, kelompok umur ke 3 berusia 50-54 tahun, kelompok umur ke 4 berusia 55-59 tahun, kelompok umur ke 5 berusia 60-64 tahun, kelompok umur ke 6 berusia 65-69 tahun, kelompok umur ke 7 berusia 70-74 tahun, kelompok umur ke 8 berusia 75-79 tahun dan kelompok umur ke 9 berusia 80 tahun keatas. Peubah bebas kedua adalah perilaku merokok yang terdiri dari 4 tipe perokok yakni tidak merokok, perokok ringan, perokok sedang dan perokok berat. Peubah ketiga adalah peubah tak bebas yang menyatakan jumlah penderita kanker paru-paru yang meninggal setiap tahun.

Untuk menganalisa model regresi Poisson terlebih dahulu dilakukan pengujian menggunakan uji *Kolmogorov-Smirnov* menggunakan SPSS 16. Analisis output untuk uji *Kolmogorov-Smirnov* adalah sebagai berikut:

H_0 **:** Jumlah penderita kanker paru-paru yang meninggal dari populasi berdistribusi Poisson.
H_1 **:** Jumlah penderita kanker paru-paru yang meninggal dari populasi yang tidak berdistribusi Poisson.

Berdasarkan hasil output diperoleh nilai $D_{hitung} < D_{tabel}$, sehingga dapat disimpulkan bahwa H_0 diterima atau jumlah penderita penyakit kanker paru-paru yang meninggal berdistribusi Poisson.

Setelah dilakukan pengujian distribusi Poisson, langkah selanjutnya adalah mendefinisikan model dengan menggunakan 8 variabel dummy untuk usia dan 3 variabel dummy untuk perilaku merokok. Sehingga model regresi Poisson untuk kasus ini adalah

$$y_i = t_i exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_{11} x_{i11}) + \varepsilon_i.$$

Selanjutnya dibentuk model regresi Poissson dengan menaksir parameter-parameter $\beta_0, \beta_1, \beta_2, \beta_3, \cdots, \beta_{11}$ untuk mengetahui ada tidaknya pengaruh variabel prediktor terhadap variabel respon dengan menggunakan hipotesis sebagai berikut:

H_0 **:** $\beta_j = 0, \quad j = 1, 2, 3, \cdots, 11$
H_1 **:** $\beta_j \neq 0, \quad j = 1, 2, 3, \cdots, 11$

Nilai G_{hitung} pada analisis ini adalah $21,487$ sedangkan $\chi^2_{(0.05,22)} = 12,34$. Karena nilai $G_{hitung} > \chi^2_{(0.05,22)}$ maka tolak H_0 yang berarti paling tidak ada satu peubah yang berpengaruh terhadap jumlah kematian penderita kanker paru-paru. Dengan menggunakan bantuan SPSS 16 didapat nilai taksiran parameter model regresi Poisson. Berdasarkan output SPSS 16 tersebut ada 10 parameter model regresi Poisson yang signifikan dan 1 parameter yang tidak signifikan. Parameter yang tidak signifikan tersebut dikeluarkan dari model dan diuji kembali hingga diperoleh model terbaik. Berikut ini adalah taksiran parameter dan nilai $p - value$ model regresi Poisson:

Tabel 1. Taksiran Parameter Model Regresi Poisson

Parameter	Taksiran	Standar Error	P-Value
Konstanta(β_0)	$-0,852$	$0,0579$	$0,000$
Usia=40-44(β_1)	$-2,818$	$0,0789$	$0,000$
Usia=45-49(β_2)	$-2,268$	$0,0763$	$0,000$
Usia=50-54(β_3)	$-1,834$	$0,0729$	$0,000$
Usia=55-59(β_4)	$-1,443$	$0,0605$	$0,000$
Usia=60-64(β_5)	$-1,168$	$0,0575$	$0,000$
Usia=65-69(β_6)	$-0,819$	$0,0577$	$0,000$
Usia=70-74(β_7)	$-0,553$	$0,0596$	$0,000$
Usia=75-79(β_8)	$-0,288$	$0,0649$	$0,000$
Perokok Berat(β_9)	$0,214$	$0,0388$	$0,000$
Perokok Sedang(β_{10})	$0,412$	$0,0401$	$0,000$
Scale	1	.	.

Sehingga model regresi Poisson untuk kasus ini adalah

$$y_i = t_i \; exp(-0,852 - 2,818x_{i1} - 2,268x_{i2} - 1,834x_{i3} - 1,443x_{i4} - 1,168x_{i5} - 0,819x_{i6} - 0,553x_{i7} - 0,288x_{i8} + 0,214x_{i9} + 0,412x_{i10}).$$

dengan interpretasi

konstanta : log nilai harapan jumlah kematian penderita kanker paru-paru sebesar -0,852 ketika seseorang berusia 80 keatas dan tidak merokok.

usia 40-44 (x_{i1})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 2,818 lebih rendah untuk usia 40-44 tahun dengan asumsi peubah yang lainnya konstan.

usia 45-49 (x_{i2})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 2,268 lebih rendah untuk usia 45-49 tahun dengan asumsi peubah yang lainnya konstan.

usia 50-54 (x_{i3})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 1,834 lebih rendah untuk usia 50-54 tahun dengan asumsi peubah yang lainnya konstan.

usia 55-59 (x_{i4})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 1,443 lebih rendah untuk usia 55-59 tahun dengan asumsi peubah yang lainnya konstan.

usia 60-64 (x_{i5})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 1,168 lebih rendah untuk usia 60-64 tahun dengan asumsi peubah yang lainnya konstan.

usia 65-69 (x_{i6})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 0,819 lebih rendah untuk usia 65-69 tahun dengan asumsi peubah yang lainnya konstan.

usia 70-74 (x_{i7})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 0,553 lebih rendah untuk usia 70-74 tahun dengan asumsi peubah yang lainnya konstan.

usia 75-79 (x_{i8})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 0,288 lebih rendah untuk usia 75-79 tahun dengan asumsi peubah yang lainnya konstan.

Perokok berat (x_{i9})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 0,214 lebih tinggi untuk perokok berat dengan asumsi peubah yang lainnya konstan.

Perokok sedang (x_{i10})**:** perbedaan dalam log nilai harapan jumlah kematian diperkirakan 0,412 lebih tinggi untuk perokok sedang dengan asumsi peubah yang lainnya konstan.

6 Ucapan Terima kasih

Penulis mengucapkan terima kasih kepada Bapak Dodi Devianto, Bapak Yudiantri Asdi, Bapak Zulakmal, Ibu Hazmira Yozza dan Bapak Admi Nazra yang telah memberikan masukan dan saran sehingga makalah ini dapat diselesaikan dengan baik.

7 Daftar Pustaka

1. Anonymous. 2012. *Generalized Linier Models*, data.pricenton.edu/wws509/ datasets/ diakses tanggal 6 Agustus 2012
2. Cameron, A.C. dan P.K. Trivedi. 1998. *Regression analysis of count data*. Cambridge University Press, New York
3. Kleinbaum, D.G., L.L. Kupeer., A. Nizam., dan K.E. Muller. 1978. *Applied Regression Analysis and Other Multivariabel. Fourth Editon*. Thomson Brooks//Cole, Canada
4. McCullagh P., and J.A. Nelder. 1989. *Generalized Linier Models. Second Edition*. Chapman and Hall, London
5. Myers, R.H. 1990. *Classical and Modern Regression with Application. Second Edition*. PWS-KENT Publishing Company, Boston