INFERENSI, *Vol. 1(1), September. 2018, ISSN: 0216-308X*

Generalized Additive Logistic Pada Pemodelan Faktor-Faktor Yang Mempengaruhi Keuntungan PT. PDC

Kartika Fithriasari[1], Soehardjoepri[2], Nur Iriawan[1]
[1]Departemen Statistika, Fakultas Matematika, Komputasi dan Sains Data
[2]Departemen Aktuaria, Fakultas Matematika, Komputasi dan Sains Data
Institut Teknologi Sepuluh Nopember (ITS)
Jalan Arief Rahman Hakim, Surabaya 60111
E-mail: kartika_f@statistika.its.ac.id

***Abstrak—Generalized Additive Models* (GAM) merupakan kombinasi dari model additive dan *generalized linear models* (GLMs). GAM dengan variabel respon bertipe biner disebut model *generalized additive logistic*. Perbedaan hasil model regresi logistik pada GLMs dan GAM didapatkan pada pemodelan faktor-faktor yang mempengaruhi keuntungan PT.PDC. Dari studi kasus PT.PDC. terlihat bahwa GLMs hanya menangkap hubungan linier antara log-odds dan variabel prediktor, sedangkan GAM dapat menangkap hubungan kuadratik yang digambarkan dalam grafik prediksi parsial. Sehingga dapat disimpulkan bahwa GAM mampu memodelkan hubungan yang lebih kompleks dibanding GLMs.**

Kata Kunci—Cing Fong H*, Clayton, Copula, *Maintenance, Reliabilitas, Weibull

I. PENDAHULUAN

Generalized Additive Models (GAM) adalah sutu metode yang dikembangkan oleh [1]. GAM merupakan kombinasi dari model additive dan *generalized linear models* (GLMs). Model ini terdiri dari komponen random, komponen aditif dan fungsi link. GAM dan GLMs mempunyai tujuan analistis yang berbeda. GLMs digunakan untuk model parametrik [2], sedangkan GAM fokus pada eksplorasi data nonparametrik.

GAM mengganti fungsi yang didefinisikan pada GLMs dengan fungsi penghalus non-parametrik dalam menemukan hubungan yang ada. Beberapa penghalus berbeda tersedia, tetapi yang paling sering digunakan adalah spline atau loess. Penghalus memiliki parameter yang dapat digunakan untuk mengontrol kedekatan trend model dengan data. GAM adalah model aditif karena secara simultan GAM memodelkan efek yang berbeda dari setiap variabel bebas. Setiap efek dapat diestimasi dengan menggunakan penghalus atau fungsi matematika, yang mengarah pada GAM sebagai model semiparametrik.

II. GENERALIZED ADDITIVE MODELS (GAM)

Anggap Y adalah variabel acak respon dan $X_1, X_2, ..., X_p$ adalah himpunan variabel prediktor. Prosedur regresi dapat dipandang sebagai suatu metoda untuk estimasi bagaimana nilai Y tergantung pada nilai $X_1, X_2, ..., X_p$ [4]. Model regresi linier standard menganggap nilai harapan Y mempunyai bentuk linier

$$E(Y) = f(X_1,..., X_p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \quad (1)$$

Dalam regresi linier standard, parameter $\beta_0, \beta_1, ..., \beta_p$ diestimasi dengan metode *least square*. Model additive mengeneralisasi model linier dengan memodelkan nilai harapan Y sebagai bentuk

$$E(Y) = f(X_1,..., X_p) = s_0 + s_1(X_1) + ... + s_p(X_p) \quad (2)$$

dimana $s_i(X_i)$, untuk $i=1,..., p$ adalah fungsi penghalus yang diestimasi dengan cara nonparametrik. Jika penghalus nonparametrik $s_i(X_i)$, untuk $i=1,..., p$ adalah *smoothing spline*, maka s_i dapat diestimasi dengan motode *penalized least squares.*

GAM mengembangkan model linier standard dengan cara lain, yaitu adanya fungsi link antara $f(X_1, ..., X_p)$ dan nilai harapan Y. Dengan cara seperti itu, model lebih fleksible untuk diterapkan dalam mengatasi kasus dimana asumsi tidak terpenuhi [3]. GAM terdiri dari komponen random, komponen additive dan fungsi link yang menghubungkan kedua komponen. Random komponen, yaitu variabel respon Y, dianggap mempunyai distribusi keluarga eksponensial

$$f_Y(y;\theta;\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \quad (3)$$

dimana θ adalah parameter natural dan ϕ adalah parameter skala. Komponen additive pada GAM adalah kuantitas η dengan bentuk

$$\eta = s_0 + \sum_{i=1}^{p} s_i(X_i) \quad (4)$$

dimana variabel respon mempunyai fungsi kepadatan peluang (*probabiity density*) keluarga eksponensial. Hubungan antara rata-rata variabel respon μ dan η ditunjukkan dengan fungsi link $g(.)$, yang dapat ditulis sebagai

$$g(\mu) = \eta \quad (5)$$

INFERENSI, *Vol. 1(1), September. 2018, ISSN: 0216-308X*

III. GENERALIZED ADDITIVE LOGISTIC MODEL

Pada model regresi logistik, variabel respon *Y* mempunyai nilai 0 atau 1, yang disebut data biner. Anggap *Y* adalah variabel respon biner dan $X_1, X_2,..., X_p$ adalah variabel prediktor, sehingga model logistik linier menganggap bahwa log-odds adalah linier:

$$\log \frac{p(y_i \mid x_{i1}, x_{i2},..., x_{ip})}{1- p(y_i \mid x_{i1}, x_{i2},..., x_{ip})} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \qquad (6)$$

Dari model logistik linier dapat dikembangkan model *generalized additive logistic* sebagai berikut:

$$\log \frac{p(y_i \mid x_{i1}, x_{i2},..., x_{ip})}{1- p(y_i \mid x_{i1}, x_{i2},..., x_{ip})} = \beta_0 + s_1(x_{i1}) + s_2(x_{i2}) + ... + s_p(x_{ip}) \qquad (7)$$

Dimana fungsi $s_1, s_2,..., s_p$ adalah fungsi penghalus yang diestimasi secara nonparametrik. Fungsi s_j, untuk $j=1,..., p$ adalah fungsi yang tidak spesifik. Fungsi ini diestimasi dengan cara yang fleksible dengan menggunakan *scatterplot smoother*, sehingga dapat menangkap pengaruh nonlinieritas dari variabel prediktor.

IV. METODE PENGHALUS

Anggap ada titik-titik pada scatterplot (x_i,y_i), dimana *Y* adalah variabel respon dan *X* adalah variabel prediktor. Hubungan ketergantungan *Y* pada *X* ingin dibentuk oleh kurva penghalus *s*(*X*). Jika ingin dicari kurva yang meminimumkan kuadrat eror, hasilnya adalah kurva interpolasi yang tidak halus di semua titik. Penghalus cubic spline membebankan kehalusan pada fungsi *s(X)*. Ingin dicari fungsi *s(X)* yang meminimumkan

$$\sum [y_i - s(x_i)]^2 + \lambda \int s''(x)^2 dx \qquad (8)$$

$\int s''(x)^2 dx$ menghitung "wiggliness" dari fungsi *s(X)* sedangkan λ adalah parameter penghalus yang harus ditentukan dan nilainya tidak negatif . Jika *s(X)* linier, maka $\int s''(x)^2 dx = 0$. Untuk sembarang nilai λ, solusi dari (8) adalah cubic spline, yaitu sebuah *piecewise cubic polynomial* dengan penggabungan potongan (*pieces*) pada nilai *x* dalam dataset.

Anggap model additive yang sesuai

$$\hat{y}_i \approx \sum_j s_j(x_{ij}) \qquad (9)$$

Kriteria (8) dapat dispesifikan untuk masalah ini, dan prosedur iterasi tersedia untuk mengestimasi fungsi s_j. Penghalus *cubic spline* diterapkan pada $y_i - \sum_{j \neq k} \hat{s}_j(x_{ij})$ sebagai fungsi x_{ik}. Proses berlanjut sampai estimasi $\hat{s}_j$ stabil. Prosedur tersebut dikenal sebagai metode Backfitting.

V. STUDI KASUS

Data keuntungan tiap outlet PT. PDC. dicatat pada akhir tahun 2009. Jumlah pengamatan yang ada adalah 40 data. Pada tahun 2009 PT.PDC mempromosikan produk-produknya dengan membuka outlet-outlet di berbagai daerah di seluruh Indonesia. Pada akhir tahun manajemen mencatat besarnya luas outlet (dalam m^2) dan biaya akomodasi (juta rupiah) serta keuntungan perusahaan. Untuk variabel keuntungan perusahaan dicatat dalam dua kategori 1 yang berarti "untung" dan 0 yang berarti "tidak untung". PT.PDC ingin melihat pengaruh variabel prediktor: luas outlet (X_1) dan biaya akomodasi (X_2) terhadap keuntungan perusahaan (Y) . Data ini akan diolah dengan GLMs dan GAM. Dalam SAS untuk GLMs menggunakan PROC GENMOD, sedangkan untuk GAM menggunakan PROC GAM. Variabel respon pada kasus ini bertipe binary data, sehingga fungsi link yang digunakan adah transformasi logit.

Tujuan dari analisis ini adalah mengidentifikasi hubungan antara variabel prediktor (X_1, X_2) dan variabel respon Y dengan regresi logistik dalam konteks GLMs. Sehingga PROC GENMOD dalam SAS dapat digunakan untuk melihat hubungan variabel tersebut. Hasil PROC GENMOD dapat dilihat pada tabel 1.

Tabel 1 Analisis estimasi parameter untuk regresi logistik dalam GLMs

Analysis Of Parameter Estimates

Parameter	DF	Standard Estimate	Wald 95% Confidence Error	Limits	Chi-Square	Pr > ChiSq
Intercept	1	9.2536	3.9486	1.5145	16.9927	5.49
						0.0191
X1	1	-0.2376	0.1134	-0.4599	-0.0152	4.39
						0.0362
X2	1	-1.1257	0.5797	-2.2618	0.0105	3.77
						0.0521

Dari tabel 1, model yang didapat dengan menggunakan regresi logistik dalam GLMs adalah:

$$\log \frac{p(y_i \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})}{1- p(y_i \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})} = 9{,}2536 - 0{,}2376 x_{i1} - 1{,}1257 x_{i2}$$

Dari tabel 1 terlihat bahwa pengaruh X_1 cukup significant pada taraf 5%. Sedangkan X_2 tidak signifikan. Pada GLMs hubungan antara log odds dengan variabel prediktor diasumsikan linier, karena X_2 tidak signifikan, maka akan dilihat generalized additive logistic model untuk kasus tersebut. Untuk mendapatkan model GAM digunakan PROC GAM dalam SAS, hasilnya dapat dilihat pada tabel 2. Dan tabel 3.

Tabel 2 Estimasi parameter analisis model regresi dalam GAM

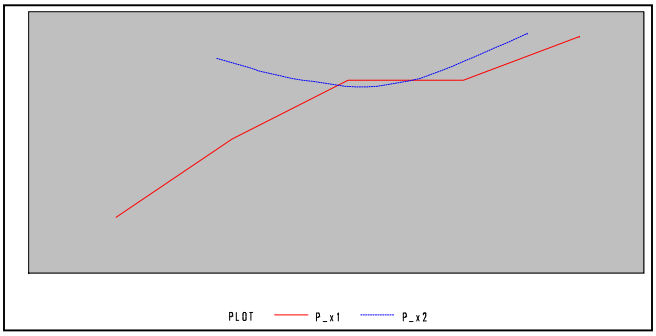
Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-8.44442	4.22376	-2.00	0.0539
Linear(x1)	0.24315	0.12224	1.99	0.0550
Linear(x2)	0.47629	0.86557	0.55	0.5858

INFERENSI, *Vol. 1(1), September. 2018, ISSN: 0216-308X*

Tabel 3 Analisis Smoothing Model Devians menggunakan prosedur GAM

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(x1)	3.00000	0.912680	1.1767	0.7586
Spline(x2)	3.00000	9.824273	12.6664	0.0054

Tabel 2 menunjukkan bahwa prediksi parsial X_1 hampir signifikan pada taraf 5% dengan pola linier, sedangan X_2 pada tabel 3 terlihat significant pada taraf 5%. Untuk selanjutnya dibuat grafik prediksi parsial dari tiap variabel prediktor, hasilnya dapat dilihat pada gambar 1.



Gambar 1 Prediksi parsial untuk setiap variabel prediktor

Pada gambar 1 dapat dilihat bahwa pola X_2 adalah kuadratik sedangkan pola X_1 mendekati linier, sehingga dicoba model kuadratik dengan menggunakan PROC GENMOD dan hasilnya dapat dilihat pada tabel 4.

Tabel 4 Analisis estimasi parameter dengan Proc GENMOD

Analysis Of Parameter Estimates

Parameter	DF	Standard Estimate	Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.4483	4.7515	-9.7611	8.8645	0.01	0.9248
X1	1	-0.3135	0.1257	-0.5598	-0.0672	6.22	0.0126
X2	1	12.8678	6.2699	0.5790	25.1567	4.21	0.0401
X2*x2	1	-3.6286	1.6790	-6.9193	-0.3379	4.67	0.0307

Tabel 4 menunjukkan semua parameter significant pada taraf 5% kecuali intersept, sehingga dicoba untuk mengolah kembali dengan model tanpa intersep.

Tabel 5 Analisis estimasi parameter dengan Proc GENMOD tanpa intersep

Analysis Of Parameter Estimates

Parameter	DF	Standard Estimate	Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
X1	1	-0.3164	0.1229	-0.5574	-0.0755	6.63	0.0101
X2	1	12.4553	4.4241	3.7843	21.1264	7.93	0.0049
X2*x2	1	-3.5204	1.2050	-5.8822	-1.1586	8.53	0.0035

Setelah intersep dihilangkan terlihat pada tabel 5 bahwa semua parameter sudah significant, sehingga model yang didapat adalah:

$$\log\frac{p(y_i \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})}{1-p(y_i \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})} = -0{,}3164x_{i1} + 12{,}4553x_{i2} - 3{,}5204x_{i2}^2$$

VI. KESIMPULAN

Dari analisis diatas, terlihat bahwa ketika memodelkan antara variabel luas outlet (X_1) dan biaya akomodasi (X_2) dengan variabel keuntungan (Y) pertama kali dengan menggunakan regresi logistik dalam GLMs, hasilnya adalah tidak ada hubungan antara variabel biaya akomodasi (X_2) dengan variabel keuntungan (Y). Selanjutnya dimodelkan dengan GAM, ternyata pola dari variabel biaya akomodasi (X_2) dapat ditangkap yaitu kuadratik, sehingga model yang menunjukan hubungan antara variabel luas outlet (X_1) dan biaya akomodasi (X_2) dengan variabel keuntungan (Y) dapat dibuat. GLMS hanya menangkap hubungan linier antara log-odds dan variabel prediktor, sedangkan GAM dapat menangkap hubungan kuadratik yang dgambarkan dalam grafik prediksi parsial. Sehingga dapat disimpulkan bahwa GAM mampu memodelkan hubungan yang lebih kompleks dibanding GLMs.

DAFTAR PUSTAKA

[1] T. Hastie dan R. Tibshirani, "Generalized Additive Models," *Statistical Science*, vol. I, no. 3, pp. 297-318, 1986.

[2] P. McCullagh dan J. A. Nelder, Generalized Linear Models, 2nd penyunt., London: Chapman & Hall, 1989.

[3] Y. Terzi dan M. A. Cengiz, "Using of generalized additive model for model selection in multiple poisson regression for air pollution data," *Scientific Research and Essay*, vol. IV, no. 9, pp. 867-871, September 2009.

[4] T. Hastie dan R. Tibshirani, Generalzed Additive Models, 1st penyunt., Chapman and Hall, 1990.