LAPORAN AKHIR
PENELITIAN DOSEN UTAMA

PENDEKATAN SEMIPARAMETRIK PADA REGRESI LOGISTIK UNTUK MENGURANGI BIAS KETIKA ADA MASALAH DATA *RARE EVENT*

TIM PENGUSUL:

SULIADI, S.SI., M.SI., PH.D (NIDN: 0416117202)

SITI SUNENDIARI, DRA., MS (NIDN: 0422106101)

DR. ACENG K. MUTAQIN, S.SI., MT., M.SI (NIDN: 0428117401)

NOER BUNGA AGRIANI S. PUTRI (NPM: 10060113035)

FEBRIAN TEGUH RAHARJO (NPM: 10060113029)

LEMBAGA PENELITIAN DAN PENGABDIAN PADA MASYARAKAT
UNIVERSITAS ISLAM BANDUNG
SEPTEMBER 2017

Halaman Pengesahan Penelitian Dosen Utama

Judul Penelitian : Pendekatan Semiparametrik Pada Regresi Logistik Untuk Mengurangi Bias Ketika Ada Masalah Data *Rare Event*

Ketua Peneliti

a. Nama Lengkap	: Suliadi, S.Si., M.Si., Ph.D
b. NIP/NIK	: D.97.0.267
c. NIDN	: 0416117202
d. Jabatan Fungsional	: Lektor
e. Fakultas/Program Studi	: MIPA/Statistika
f. Nomor HP	: 085846252822
g. Alamat E-mail	: suliadi@gmail.com

Anggota Peneliti

No.	Nama Lengkap	NIDN/NPM	Fakultas/Program Studi
1	Siti Sunendiari, Dra., MS	0422106101	MIPA/Statistika
2	Dr. Aceng K. Mutaqin, S.Si., MT., M.Si	0428117401	MIPA/Statistika
3	Noer Bunga Agriani S. Putri	10060113035	MIPA/Statistika
4	Febrian Teguh Raharjo	10060113029	MIPA/Statistika

Bandung, 15 September 2017

Mengetahui,
Dekan Fakultas MIPA
Universitas Islam Bandung

Dr. [illegible]iwanda, MS.
NIK: D.86.0.045

Ketua Peneliti,

Suliadi, S.Si., M.Si., Ph.D.
NIK: D.97.0.267

Mengetahui:
Ketua LPPM Universitas Islam Bandung,

Prof. Dr. Hj. Atie Rachmiatie, Dra., M.Si.
NIP: 195903301986012002

Lembar Persetujuan Reviewer

Reviewer 1

Dr. Rini Rinawati, M.Si

Reviewer 2

Abdul Kudus, S.Si., M.Si., Ph.D

RINGKASAN

Regresi logistik adalah metode yang paling sering dipergunakan untuk memodelkan respon biner, di mana respon hanya terdiri dari dua kemungkinan "sukses" dan "gagal". Metode yang biasa dipergunakan untuk menduga parameter regresi logisik adalah metode kemungkinan maksimum. Metode ini akan menghasilkan penduga yang tak bias dan ragam penduga akan minimum. Akan tetapi jika perbandingan antara banyaknya respon "sukses" dan banyaknya respon "gagal" terlalui tinggi, maka akan mengakibatkan bias terhadap parameter dugaan. Kasus ini disebut sebagai *rare event*. Beberapa metode telah diajukan untuk mereduksi bias, akan tetapi metode tersebut hanya efektif untuk data-data dengan ukuran sampel relative kecil. Ketika data lebih dari 200, maka metode tersebut tidak akan memberikan efek dalam mereduksi bias. Penelitian ini bertujuan untuk mendapatkan metode yang dapat mereduksi atau jika mungkin menghilangkan bias pada regresi logistik ketika ada masalah data *rare event*. Performa dari metode tersebut akan dievaluasi pada berbagai ukuran sampel dan tingkat kejarangan (level of rare) melalui simulasi

Dalam penelitian ini kami mengajukan pendekatan pemodelan semiparametrik sebagai pendekatan untuk mereduksi bias melalui regresi P-Spline. Kami mengevaluasi kinerja model semiparametrik melalui studi simulasi. Dari hasil simulasi kami peroleh bahwa model semiparamtrik dapat mereduksi bias, terutama bias pada koefisien intersep. Akan tetapi kemampuannya masih agak rendah, sehingga untuk tingkat kejarangan yang tinggi, bias masih agak besar. Meskipun demikian, bias dari model semiparametrik masih lebih kecil dibandingkan model parametrik.

PRAKATA

Pertama-tama kami ingin mengucapkan syukur kehadirat Allah SWT, karena atas rahmat-Nya kami bisa menyelesaikan laporan penelitian ini.

Tahapan penelitian telah dilaksanakan dengan pencapaian 100% dari target jadwal penelitian. Meskipun demikian diseminasi hasil penelitian belum 100% karena satu paper akan diseminarkan dalam seminar internasional yang akan dilaksanakan pada bulan September 2017. Selain itu satu artikel masih dalam bentuk draf, direncakan untuk dipubilkasikan dalam publikasi internasional.

Kami ingin mengucapkan terima kasih yang sebesar-besarnya kepada Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Islam Bandung atas dana yang disediakan untuk penelitian ini. Selain itu kami juga ingin mengucapkan terima kasih kepada ketua LPPM Unisba dan juga Dekan FMIPA Unisba atas bantuannya selama ini. Tak lupa ucapan terima kasih kami sampaikan kepada para staf LPPM Unisba atas semua bantuannya.

Bandung, Agustus 2017

Tertanda,

Tim Peneliti

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I. PENDAHULUAN

1.1 Latar Belakang

Dalam statistika, data biner atau data dikotomus merupakan data yang hanya mempunyai dua kemungkinan nilai, yang dibiasnya dinyatakan dengan "sukses" atau "gagal". Kejadian "sukses" mengacu pada suatu keadaan di mana obyek penelitian mengalami atau mempunyai karakteristik yang menjadi fokus perhatian penelitian dan kejadian "gagal" adalah sebaliknya. Kejadian sukses biasanya dikodekan dengan 1 (satu) dan kejadian gagal dikodekan dengan 0 (nol). Kasus-kasus data biner banyak terjadi hampir disemua bidang.

Pemodelan data biner yang paling umum dan paling banyak dipergunakan adalah dengan menggunakan model logit, yang disebut sebagai regresi logistik. Hal ini disebabkan karena dalam model logit, intepretasi koefisien regresi lebih mudah untuk diintepretasikan dibandingkan dengan model lain. Pendugaan model atau pendugaan koefisen regresi logistik biasanya menggunakan metode kemungkinan maksimum (*maximum likelihood*). Penduga kemungkinan maksimum dari koefisen regresi logistik mempunyai sifat-sifat yang baik, diantaranya adalah bersifat tak bias dan penduga memiliki ragam yang minimum. Akan tetapi ketidakbiasan penduga koefisien ini dapat tercapai jika proporsi respon sukses dengan gagal tidak terlalu besar (McCullagh dan Nelder, 1989). Jika ada perbedaan yang besar antara proporsi sukses dengan proporsi gagal, maka penduga bagi koefisien regresi akan bersifat bias, termasuk juga penduga peluang sukses maupun gagal juga akan bias (Guns dan Vanacker, 2012; King dan Zeng, 2001; Qiu, et. al., 2013). Kasus ini sering disebut dengan kejadian jarang (*rare events*) yang didefinisikan sebagai suatu kondisi data, di mana proporsi sukses (atau gagal) kurang dari 10% terhadap keseluruhan banyaknya data. Contoh-contoh dari kasus yang masuk kategori rare event diantaranya adalah kasus perang, kejadian tanah longsor, kasus kredit macet, kartu kredit macet dan sebagainya (Guns dan Vanacker, 2012; King dan Zeng, 2001; Qiu, et. al., 2013). Kasus-kasus lain yang termasuk dalam regresi logistik dengan kasus *rare event* antara lain kegagalan alat komunikasi (Weiss dan Hirsh, 2000), konflik internasional (King dan Zeng, 2001b), bahkan kasus tergelincirnya kerata api (Quigley et. al., 2007).

Perbankan di Indonesia juga mengalami permasalahan terkait dengan data *rare event*. Sebab Bank Indonesia mengeluarkan peraturan yang mengharuskan bank untuk mempunyai perangkat untuk memeringkatkan (rating) kredit, menghitung resiko kredit maupun resiko kartu kredit, yang dibuat oleh pihak luar maupun atau dibangun oleh kalangan internal itu sendiri (Peraturan BI no. 11 tahun 2009). Hal ini disebabkan regresi logistik merupakan alat utama dalam pemeringkatan maupun dalam menilai resiko pemberian suatu kredit atau kartu kredit kepada pihak lain (Rezac, 2011; Siddiqi, 2006), sedangkan data dalam termasuk dalam kasus *rare event* di mana kejadian debitur atau pemegang kartu kredit gagal bayar sangat sedikit dibandingkan dengan yang pembayarannya lancar. Oleh karena itu model regresi logistik yang dihasilkan akan bersifat bias dan bisa menimbulkan kerugian besar pada dunia perbankan di Indonesia.

Beberapa prosedur untuk mengkoreksi bias telah diajukan oleh beberapa penulis. McCullagh dan Nelder (1989) telah mengajukan koreksi bagi nilai dugaan koefisien regresi, sedangkan King dan Zeng (2001) mengajukan koreksi terhadap penduga bagi koefisien β_0 atau intersep dan koreksi terhadap peluang sukses. Pendekatan yang lain diajukan oleh Qiu, e.t al. (2013), di mana mereka mengajukan metode maksimum likelihood terboboti dengan terlebih dahulu merekonstruksi data serta melakukan koreksi sebagaimana disarankan oleh McCullagh dan Nelder (1989) dan King dan Zeng (2001). Akan tetapi koreksi terhadap bias, baik terhadap parameter regresi maupun terhadap peluang hanya akan berfungsi dengan baik jika ukuran sampelnya relatif kecil. Beberapa peneliti juga telah mengajukan metode berdasarkan rancangan sampling (Fithian & Hastie, 2014; Sei, 2014) yang disebut dengan *undersampling* dan *oversampling*. Akan tetapi metode ini punya kelemahan, yaitu masalah kehilangan informasi pada *undersampling* dan masalah *overfitting* pada *oversampling*.

Sejauh ini pemodelan data *rare event* tidak pernah menggunakan pendekatan model nonparametrik atau semiparametrik. Padahal pada kondisi yang ideal (tidak ada masalah *rare event*), model-model non dan semiparametrik mempunyai kemampuan untuk mereduksi bias (Eubank, 1999; Green & Silverman, 1994; Wu & Zhang, 2006; Suliadi, et al., 2010a,b; Suliadi, et al., 2013; Suliadi, 2014, Suliadi & Kudus, 2015). Oleh karena itu perlu dilakukan pengkajian kemampuan model regresi logistik semiparametrik dalam menghadapi permasalahan data *rare event* dalam kasus data biner.

1.2 Perumusan Masalah

Masalah utama dalam metode yang sudah ada selama ini adalah bahwa koreksi yang dilakukan punya efek yang cukup besar dalam mengkoreksi bias, hanya ketika ukuran sampelnya relatif kecil, yaitu kurang dari 200 pengamatan (King dan Zeng, 2001). Koreksi bias menjadi gagal ketika ukuran sampel besar. Terlebih untuk ukuran sampel sangat besar sampai lebih dari 10,000 pengamatan, koreksi bias menjadi hampir tidak ada artinya. Pada banyak kasus yang melibatkan data biner dengan permasalahan *rare event*, seringkali ukuran sampelnya adalah besar. Hal ini terjadi pada kasus-kasus kartu kredit, bencana alam, konflik internasional, dan banyak kasus lainnya. Sedangkan metode berdasarkan rancangan sampling punya kelemahan, yaitu adanya kehilangan informasi (pada kasus *undersampling*) atau *overfitting* (pada kasus *oversampling*). Sedangkan penggunaan model regresi logistik semiparametrik untuk mengatasi masalah data *rare event* belum pernah dikaji orang.

Oleh karena itu permasalahan utama yang akan diteliti di sini adalah bagaimana mereduksi bias pada regresi logistik ketika ada masalah *rare event*, terutama pada kasus ukuran sampel yang besar, dengan menggunakan model regresi logistik semiparametrik.

BAB II. TINJAUAN PUSTAKA

2.1 *State of The Art*

2.1.1 GLM dan Regresi Logistik

Generalized linear model (GLM) (McCullagh & Nelder, 1989) adalah suatu kelas model linier yang lebih umum dari pada regresi linier biasa. GLM menjadi kelas paling populer dalam memodelkan hubungan antara variabel respon dengan variabel bebas. Hal ini disebabkan GLM mampu menangani berbagai macam distribusi dalam satau kerangka kerja. Sehingga pendugaan dan juga teori-teori pengujian hipotesis untuk berbagai macam distribusi variabel respon bersifat umum, cukup satu tapi dapat dipakai untuk semua model. GLM dapat menangani semua distribusi yang masuk dalam keluarga eksponensial, diantaranya Bernaulli, Binomial, Poisson, nomal, eksponensial, dan lain-lain.

Ada tiga komponen dalam GLM, yaitu

1. Komponen acak (*random component*). Komponen ini merupakan spesifikasi dari distribusi varaiabel respon.
2. Komponen sistematik (*systematic component*). Komponen ini menyatakan bentuk fungsi linier dari variabel bebas $\mathbf{x} = (x_1, x_2, ..., x_p)^T$, yang biasanya dapat dinyatakan sebagai

$$\begin{aligned}\eta &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \\ &= x^T \beta\end{aligned}$$

   di mana $\beta=(\beta_1, \beta_2,..., \beta_p)^T$.
3. Fungsi pengubung (*link function*). Fungsi hubung ini menghubungkan antara komponen acak dengan komponen sistematik.

Misalkan $y_1, y_2, ...,y_n$ adalah sampel acak berukuran *n* dari keluarga distribusi eksponen, maka fungsi densitas peluangnya dapat dinyatakan sebagai

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \qquad (1)$$

di mana θ_i dan ϕ masing-masing adalah parameter kanonik dan parameter skala dari distribusi *f*. Parameter kanonik diasumsikan nilainya spefisik untuk sampel ke-i, sedangkan parameter skala diasumsikan berlaku umum untuk semua sampel dari populasi yang sama.

Karakterstik dari distribusi dari keluarga eksponen diberikan oleh nilai tengah dan ragamnya, yaitu

$$E(y_i) = \mu_i = b'(\theta_i)$$
$$Var(y_i) = b''(\theta_i)a(\phi).$$

Berdasarkan fungsi densitas peluang di atas (1), maka untuk sampel $y_1, y_2, \ldots, y_n$ fungsi likelihood (kemungkinan)-nya adalah

$$L = \prod_{i=1}^{n} f(y_i;\theta_i,\phi) = \prod_{i=1}^{n}\left[\exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}\right]$$

dan fungsi log-likelihoodnya adalah

$$l = \log(L) = \sum_{i=1}^{n}\log[f(y_i;\theta_i,\phi)] = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \qquad (2)$$

Misalkan $g(\mu_i)$ adalah fungsi penguhubung yang menghubungkan respon dengan kovariat (variabel bebas). Hubungan ini dapat dinyatakan sebagai

$g(\mu_i) = \eta_i$,

di mana η_i adalah komponen sistematiknya yang dapat dinyatakan sebagai

Suppose the relation between y_i and the covariate is through a link function in the form

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = x_i^T\beta \quad \text{atau}$$
$$\eta = X\beta \qquad (3)$$

dengan parameter regresinya adalah $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$., $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^T$.

Penduga kemungkinan maksimum bagi β diperoleh dengan memaksimumkan fungsi log-likelihoodnya *l*. Fungsi log-likelihoodnya akan maksimum (terhadap β) jika $\partial l/\partial\beta = 0$, yaitu

$$\frac{\partial l}{\partial \beta} = X^T\frac{\partial l}{\partial \eta} = 0$$

$$\frac{\partial l_i}{\partial \eta_i} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \left[\frac{y_i - \mu_i}{a(\phi)}\right]\frac{a(\phi)}{Var(y_i)}\frac{\partial \mu_i}{\partial \eta_i} = \left[\frac{y_i - \mu_i}{Var(y_i)}\right]\frac{\partial \mu_i}{\partial \eta_i}. \qquad (4)$$

Misalkan $W_1 = \text{Diag}\{[1/Var(y_i)][\partial\mu_i/\partial\eta_i]$, maka solusi dari persamaan di atas dapat diperoleh melalui persamaan penduga (*estimating equation*) :

$$U(\beta) = X^T W1(y-\mu) = 0, \qquad (5)$$

dengan $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$. Jika distribusi dari *y* bukan normal, maka persamaan (5) tidak mempunyai bentuk tertutup sehingga untuk mendapatkan solusi harus melalui iterasi.

Salah satu metode iterasi yang biasa dipakai adalah algoritma Fisher Scoring (*Fisher scoring algorithm)*, yaitu

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + I^{-1}U \qquad (6)$$

di mana $\hat{\beta}^{(m)}$ adalah nilai dugaan koefisien β pada iterasi ke-m. Iterasi dilakukan sampai tercapainya kriteria konvergensi tertentu. Matrik *I* pada persamaan di atas merupakan matrik informasi yang diberikan oleh

$$I = \mathrm{E}\left[\frac{\partial l}{\partial \beta}\frac{\partial i}{\partial \beta^T}\right] \tag{7}$$

atau

$$I = \mathrm{E}\left[-\frac{\partial^2 l}{\partial \beta \partial \beta^T}\right]. \tag{8}$$

Dengan menggunakan persamaan (5) di atas, maka matriks informasi dapat dinyatakan sebagai

$$I = \mathrm{E}\left[\frac{\partial l}{\partial \beta}\frac{\partial l}{\partial \beta^T}\right] = X^T W_1 \mathrm{E}[(y-\mu)(y-\mu)]^T W_1 X = X^T W X\,. \tag{9}$$

di mana

$$W = W_1 \mathrm{Diag}\{\mathrm{Var}(y_i)\}W_1 = \mathrm{Diag}\left\{\frac{1}{\mathrm{Var}(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right\}.$$

Jika kita menggunakan pendekatan persamaan (8) yang dikombinasikan dengan persamaan (5) untuk menghitung matrik *I*, maka hasil yang diperoleh juga akan sama dengan persamaan (9). Sedangkan penduga bagi ragam $\hat{\beta}$ adalah

$$\mathrm{Var}(\hat{\beta}) = I^{-1} = (X^T W X)^{-1}.$$

Dalam kelas GLM, selama distribusi dari respon termasuk kelurga eksponen, maka metode pendugaan dan juga pengujian hipotesisnya menggunakan metode yang sama.

Misalkan ada *n* subjek dan y_i adalah respon biner pada subjek ke-i dengan $y_i \in \{0,1\}$, $i = 1, 2, \ldots, n$, di mana $y_i = 1$ jika subjek ke-i memiliki karakteristik yang diminati (kejadian "sukses") dan $y_i = 0$ jika subjek ke-i tidak memiliki karakteristik yang diminati (kejadian "gagal"). Selain itu juga diamati vektor kovariat x_i yang mempengaruhi respon y dan x_i berdimensi k. Respon y_i berdistribusi Bernoulli dengan parameter π_i dan $\mathrm{E}(y_i) = \mathrm{P}(y_i=1) = \pi_i$. Misalkan variabel responnya bersifat biner, yaitu $y \in \{0,1\}$, sehingga respon y akan mengikuti distribusi Bernoulli dengan fungsi masa peluangnya diberikan oleh

$$f(y;\pi) = \pi^y (1-\pi)^{1-y}, \quad \text{untuk } y \in \{0,1\} \tag{10}$$

di mana $\pi = \mathrm{E}(y) = \mathrm{P}(y = 1)$. Persamaan (10) di atas dapat dituliskan sebagai

$$f(y;\pi) = \exp\left\{\log\left(\binom{1}{y}\right) + y\log(\pi) + (1-y)\log(1-\pi)\right\}$$
$$= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) - \log\left(\frac{1}{1-\pi}\right) + \log\binom{1}{y}\right\}$$
$$= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) - \log\left(1+\frac{\pi}{1-\pi}\right) + \log\binom{1}{y}\right\}.$$

Jika kita ambil

$$\theta = \log\left\{\frac{\pi}{1-\pi}\right\}$$
$$a(\phi) = 1$$
$$b(\theta) = \log\left(1+\frac{\pi}{1-\pi}\right) = \log(1+\exp(\theta))$$
$$c(y,\phi) = \log\binom{1}{y},$$

maka fungsi masa peluang distribusi Bernoulli akan memiliki bentuk fungsi masa peluang keluarga eksponen (1) di atas. Sehingga distribusi Bernoulli termasuk keluarga eksonen dan pendugaan parameter model dapat menggunakan konsep GLM seperti di atas.

Dengan menggunakan konsep GLM di atas, maka nilai harapan dan ragam dari respon adalah

$$\mathrm{E}(y) = b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \left[\frac{\pi}{1-\pi}\right] \times (1-\pi) = \pi,$$
$$\mathrm{Var}(Y) = b''(\theta)\cdot a(\phi) = \frac{\exp(\theta)}{[1+\exp(\theta)]^2} = \frac{\exp(\theta)}{1+\exp(\theta)} \times \frac{1}{1+\exp(\theta)} = \pi(1-\pi).$$

Ada tiga model yang biasanya digunakan untuk memodelkan data biner, yaitu: (i) model logit, (ii) model probit model, dan (iii) model complementary log-log (Fahrmeir & Tutz, 2001 [Chapter 2.]; Dobson, 2002 [Chapter 7]).

Model logit menggunakan fungsi penghubung logit, yang tidak lain merupakan fungsi penghubung kanonik (*canonical link function*) untuk distribusi Bernoulli dan distribusi Binomial, yaitu dalam bentuk:

$$g(\pi) = \mathrm{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta$$

di mana

$$\pi = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

Model probit menggunakan fungsi penghubung probit, yang mempunyai bentuk

$$g(\pi) = \Phi^{-1}(\pi) = \eta, \quad \text{di mana} \quad \pi = \Phi(\eta)$$

dengan $\Phi(\cdot)$ adalah fungsi distribusi kumulatif distribusi normal baku atau N(0,1).

Sedankan model ketiga, yaitu model complementary log-log menggunakan fungsi penghubung

$$g(\pi) = \log[-\log(1-\pi)] = \eta, \quad \text{dan } \pi = 1-\exp[-\exp(\eta)].$$

Regresi Logistik. Regresi logistik merupakan model regresi untuk data biner dengan mengambil fungsi logit sebagai fungsi penghubungnya. Sehingga keterkaitan antara respon y_i dengan kovariat (variabel bebas)-nya adalah $g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$, di mana fungsi penghubung $g(\pi_i) = \log(\pi_i / [1-\pi_i])$ yang akan menghasilkan

$$\pi_i = P(Y_i = 1) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}{1+\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}$$

Misalkan $y = (y_1, y_2, \ldots, y_n)^T$; $\pi = (\pi_1, \pi_2, \ldots, \pi_n)^T$; $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^T$; dan $X = (x_1, x_2, \ldots, x_n)^T$. Penduga maksimum likelihood untuk $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ adalah solusi dari persamaan penduga

$$U(\beta) = X^T W_1 (y - \pi) = 0$$

di mana $W_1 = \text{Diag}\{[1/\text{Var}(y_i)][\partial \pi_i / \partial \eta_i]\}$. Jika kita mengambil fungsi penghubung kanonik, maka $\partial \pi_i / \partial \eta_i = \text{Var}(y_i)$. Prosedur iterative dengan menggunakan Fisher *Scoring algorithm* untuk β adalah

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + I^{-1} U \tag{11}$$

di mana

$$I = E\left[\frac{\partial l}{\partial \beta}\frac{\partial l}{\partial \beta}\right] = X^T W_1 E[(y-\pi)(y-\pi)^T] W_1 X$$
$$= X^T W X$$

dengan $W = W_1 \text{Diag}\{\text{Var}(y_i)\} W_1 = \text{Diag}\left\{\frac{1}{\text{Var}(y_i)}\left(\frac{\partial \pi_1}{\partial \eta_i}\right)^2\right\}$.

2.1.2 *Rare Event* pada Regresi Logistik dan Koreksi Bias

Misalkan terjadi kasus *rare events* pada kasus Y=1, sehingga proporsi kejadian "sukses" adalah sangat sedikit dibandingkan kejadian Y=0 (kejadian "gagal"). Hal ini berakibat terjadinya bias, yaitu adanya *underestimate* pada dugaan P(Y=1|x) yang juga berimpilkasi pada *overestimate* pada dugaan P(Y=0|x). Kasus *rare event* terjadi jika persentase Y=1 terhadap total sampel adalah lebih kecil dari 10% (Qiu, e.t al., 2013). McCullagh dan Nelder (1989) menyatakan bahwa ketika banyaknya pengamatan Y=0 dan Y=1 tidak berimbang, maka penduga kemungkinan maksimum bagi parameter regresi logistik akan bias. Koreksi bias yang mereka sarankan adalah

$$bias(\hat{\beta}) = (X^T W X)^{-1} X^T W \psi$$

di mana

$$\psi_i = 0.5Q_{ii}((1+\omega_i)\hat{\pi}_i - \omega_i), i = 1,2,\ldots,n$$

Q_{ii} = diagonal ke-i matrik $[X(X^TWX)^{-1}X^T]$

$$W = diag[\hat{\pi}_i(1-\hat{\pi}_i)\omega_i]$$

dengan

$$\omega_i = \varpi_1 Y_i + \varpi_0(1-Y_i) \text{ dimana } \varpi_1 = \tau/\bar{y} \text{ dan } \varpi_0 = (1-\tau)/(1-\bar{y})$$

untuk τ adalah proporsi *rare event dari populasi*. Koreksi terhadap bias untuk $\hat{\beta}$ adalah

$$\tilde{\beta} = \hat{\beta} - bias(\hat{\beta}).$$

Selain koreksi terhadap bias $\hat{\beta}$, King dan Zeng (2001) juga menyarankan koreksi secara khusus terhadap $\hat{\beta}_0$ yaitu

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right].$$

Selain koreksi terhadap nilai dugaan parameter, King dan Zeng (2001) juga menyarankan koreksi dalam perhitungan P(Y=1), yaitu

$$P(Y_i = 1) \approx \tilde{\pi}_i + C_i; \quad C_i = (0.5-\tilde{\pi}_i)\tilde{\pi}_i(0.5-\tilde{\pi}_i)x_0^T V(\tilde{\beta})x_0$$

di mana $V(\tilde{\beta}) = \left[\frac{n}{n+k}\right]^2\left[\sum_{i=1}^{n}\tilde{\pi}_i(1-\tilde{\pi}_i)x_i x^T\right]^1$.

Koreksi bias di atas akan efektif untuk ukuran sampel kecil, yaitu kurang dari 200 pengamatan (King & Zeng, 2001). Sedangkan untuk ukuran sampel besar, koreksi bias efeknya sangat kecil.

2.1.3 Mengatasi Masalah *Rare Event* Melalui Rancangan Sampling

Data *rare event* sebenarnya bukan hanya menjadi masalah dalam kasus regresi logistik, akan tetapi juga dalam kasus analisis diskriminan dan penambangan data (*data mining*). Metode rekayasa rancangan sampling pada awalnya berkembang dalam kasus penambangan data dan analisis diskriminan (Weiss & Hirsh, 2000; Yap, et al., 2014). Prinsip dasar dari metode ini adalah dengan menyeimbangkan banyaknya (frekuensi) data "sukses" dengan banyaknya data "gagal". Hal ini berangkat dari pemikiran bahwa permasalahan pada data *rare event* adalah akibat tidak seimbangnya frekuensi "sukses" dengan frekuensi "gagal". Dua metode yang dikenal adalah

a. Metode *Undersampling*.

Misalkan data minoritas adalah kejadian "sukses", sehingga frekuensi sukses jauh lebih kecil dari pada frekuensi "gagal". Untuk menyeimbangkan frekuensi kejadian "sukses" dan "gagal", data dari kejadian "sukses" diambil semuanya sedangkan data dari kejadian "gagal" diambil sebagian secara acak, sedemikian sehingga persentase kejadian "sukses" dan kejadian "gagal" tidak lagi berbeda terlalu jauh.

Jadi dalam metode ini sebagian pengamatan dari kejadian "gagal" dibuang (tidak diikutsertakan) dalam analisis. Sehingga sangat mungkin banyak informasi yang tidak dimanfaatkan secara optimal. Sebagai contoh, misalkan ukuran sampel (banyaknya data/pengamatan) adalah n = 100,000, dengan kejadian sukses adalah 10% (10,000 pengamatan) dan 90% kejadian gagal (90,000 pengamatan). Misalkan diinginkan perbandingan "sukses" : "gagal" adalah 1:3, oleh karena itu dari kejadian "gagal" akan diambil secara acak sebanyak 30,000 pengamatan. Jadi akan ada 60,000 pengamatan yang akan dibuang.

b. Metode *Oversampling*.

Metode *oversampling* merupakan kebalikan *undersampling*. Dalam kasus kejadian "sukses" adalah minoritas, maka proses menyeimbangkan persentase kejadian "sukses" dengan persentase kejadian "gagal" adalah dengan menggandakan kejadian "sukses" sampai tercapainya proporsi kejadian "sukses" dan kejadian "gagal" yang diinginkan. Hal ini menyebabkan banyaknya data/pengamatan yang sama yang berulang beberapa kali, yang disebut juga dengan *overfitting*. sehingga ukuran data akan membengkak beberapa kali lipat. Sebagai contoh pada kasus *undersampling* di atas. Untuk menyeimbangkan data "sukses" dan "gagal", maka data yang sama pada kejadian "sukses" harus diulang sebanyak 3 kali (data yang sama direplikasi 2 kali) sehingga jumlah kejadian sukses menjadi 30,000 pengamatan dan total data adalah 120,000 pengamatan.

2.2 *Road Map*

Kami sudah melakukan studi pendahuluan terkait dengan kasus data *rare event* pada regresi logistik melalui penelitian yang dilakukan oleh Kudus, et. al. (2015) dan juga skripsi Wistara (2015) yang dibimbing oleh ketua tim peneliti ini. Dari kajian tersebut diperoleh bahwa perlu dicari metode untuk mereduksi atau jika mungkin menghilangkan bias terutama untuk data dengan ukuran sampel besar. Studi pendahuluan tersebut mengkonfirmasi adanya bias pada koefisien regresi yang semakin besar dengan semakin besarnya ketidakseimbangan n(Y=0) dan n(Y=1). Selain itu kami juga mendapati bahwa koreksi bias terhadap koefisien regresi dan juga P(Y=1) efeknya sangat kecil ketika diterapkan pada data dengan ukuran sampel besar.

Berdasarkan hasil penelitan kami sebelumnya (Suliadi, et al., 2016), kami mendapatkan bahwa bias yang terjadi bukan hanya terjadi pada koefien intersep (b_0) saja, akan tetapi semua koefisien variabel bebas bersifat bias, di mana bias akan semakin besar dengan semakin besarnya perbedaan persentase kejadian "sukses" dengan kejadian "gagal". Kami juga sudah melakukan penelitian terkait dengan model regresi logistik semiparametrik (Suliadi, et al., 2010a,b; Suliadi, et al., 2013; Suliadi, 2014, Suliadi & Kudus, 2015) dan mendapati bahwa model-model nonparametrik dan semiparametrik sangat baik dalam mengendalikan bias akibat ketidakcocokan

model. Sedangkan penelitan pengendalian bias dengan menggunakan model semiparametrik sejauh ini belum pernah ditemukan oleh tim pengusul kami.

Payung besar dari penelitian ini adalah pemodelan *credit scoring* dan pemodelan resiko kredit dengan tujuan jangka panjangnya adalah menemukan alat dengan kinerja yang baik dalam mengidentifikasi calon debitur atau calon pemegang kartu kredit yang baik (kemungkinan macet/tidak). Ada beberapa permasalahan dalam memodelkan *credit scoring* dan resiko kredit di perbankan, baik pada kasus kredit maupun kartu kredit, di mana regresi logistik merupakan alat utama dalam pembuatan modelnya. Permasalahan yang masih menjadi topik dalam statistika terkait dengan data biner adalah *rare event*, *rare event* dan masalah multikolinier, *rare event* dan pengamatan berpengaruh, *rare event* dan seleksi variabel, serta *rare event* dan data berkorelasi. Road map penelitian kami dalam topik *credit scoring* dan resiko kredit terkait dengan data *rare event* pada regresi logistik dapat dilihat pada Gambar 1.

Gambar 1. *Road Map* Penelitian

BAB III. TUJUAN DAN MANFAAT PENELITIAN

3.1 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mencari suatu metode yang dapat mengatasi bias pada regresi logistik ketika ada masalah *rare event* pada data, terutama pada data dengan ukuran sampel besar, dengan menggunakan model regresi logistik semiparametrik.

3.2 Manfaat Penelitian

Dalam statistika, bias adalah permasalahan serius dalam pendugaan parameter. Karena jika nilai dugaan suatu parameter bersifat bias, maka nilai dugaan parameter yang diperoleh akan jauh dari nilai yang sesungguhnya. Oleh karena itu sangat penting untuk bisa mendapatkan penduga bagi suatu parameter yang bersifat tidak bias.

Lembaga perbankan dan juga lembaga keuangan di Indonesia sangat berkepentingan dengan permasalahan regresi logistik degan kasus data *rare event*, sebab mereka harus mempunyai model yang mampu mendeteksi calon debitur yang baik berdasarkan regesi logistik. Sementara pada era infomasi seperti sekarang ini, banyak kasus di mana ukuran sampel dari data lebih dari 200 bahkan ratusan ribu pengamatan, sementara metode reduksi bias akan gagal jika diterapkan pada kasus tersebut.

Oleh karena itu sangatlah penting untuk bisa menemukan metode yang dapat mengurangi atau jika mungkin menghilangkan bias pada regresi logistik ketika ada masalah *rare event* serta ukuran sampelnya besar.

3.3 Target Luaran

Penelitian ini diharapkan dapat menemukan suatu metode yang dapat mengurangi atau jika mungkin menghilangkan bias pada regresi logistik ketika ada masalah *rare event* serta ukuran sampelnya besar. Pada jangka panjang, penelitian ini diharapan dapat menghasilkan metode analisis yang handal dalam menangani permasalahan data *rare event*. Hasil yang diperoleh nantinya akan dipublikasikan melalui seminar dan publikasi melalui jurnal internasional.

BAB IV. METODE PENELITIAN

4.1 Tahapan Penelitian

Penelitian ini juga terkait dengan road map riset unggulan Unisba dalam bidang rekayasa industri khususnya industri perbankan, yakni untuk mendapatkan tools yang sangat baik dalam memprediksi peluang kredit macet baik bagi calon nasabah maupun nasabah yang sudah *on going* serta memodelkan resiko kredit. Dalam tahun pertama metode dasar yang akan dipakai adalah regresi nonparametrik P-Spline biasa (*ordinary P-Spline*) sedangkan pada tahun kedua akan dikembangkan ke bentuk lain dari P-Spline yaitu *varying coefficient* P-Spline.

Tahapan yang akan dilalui dalam penelitian ini adalah sebagai berikut:

1. Mencari metode mereduksi bias dalam analisis regresi logistik pada kasus data rare event:
   (a). Melakukan kajian/studi pustaka mengenai masalah fenomena kejadian jarang serta dampaknya terhadap validitas kesimpulan hasil analisis regresi logistik.
   (b). Melakukan kajian/studi pustaka metode yang sudah ada dalam menangani masalah rare event dalam analisis regresi logistik maupun metode yang ada dalam analisis diskriminan.
   (c). Melakukan kajian/studi pustaka terkait dengan regresi semiparametrik untuk data biner, secara khusus metode P-Spline
   (d). Mencari model regresi logistik yang mampu menangani data *rare event* dengan menggabungkan hasil yang diperoleh pada langkah (*a*) - (*c*) di atas dan mencari metode estimasi yang efesien dan efektif, serta menulis program komputer untuk mengestimasi model regresi. Prosedur pada tahap c ini adalah:
      (i) Membuat model-model regresi semiparametrik;
      (ii) Menyusun algoritma estimasi model;
      (iii) Menuliskan program komputer sebagai implementasi dari langkah (ii);
      (iv) Evaluasi terhadap langkah (i) - (iii) di atas.
2. Kajian simulasi terhadap metode baru dalam menangani bias pada kasus rare event pada analisis regresi logistik. Kajian ini dimaksudkan untuk melihat karakteristik penduga bagi model regresi maupun koefisien regres, dalam hal: Bias, varian penduga dan distribusi dari penduga. Langkah-langkah pada tahapan ini adalah:
   (a). Pembuatan program simulasi dengan menggunakan SAS IML, Makro Minitab, atau bahasa R.
   (b). Membangkitkan data respon biner: Y = 0, 1 dengan model Logit(μ) = $\eta = 1.0 + 1.0X1 - 1.0X2 + \sin(4\pi Z)$, dengan X1 ~ normal(0,1) dan X2 ~ Bernoulli(0.5). Variabel X21 merepresentasikan variabel numerik (kontinyu) sedangkan variabel X2

merepresentasikan variabel kategorik. Variabel Z adalah variabel nonparametrik yang dibangkitkan dari U(0, π). Data yang akan dibangkitkan sebanyak 10,000,000 pengamatan dan dianggap populasi pasangan (Y, X1, X2, Z).

(c). Melakukan sampling dari populasi data (dari langkah 2.b di atas) untuk berbagai ukuran sampel dan tingkat kejarangan.

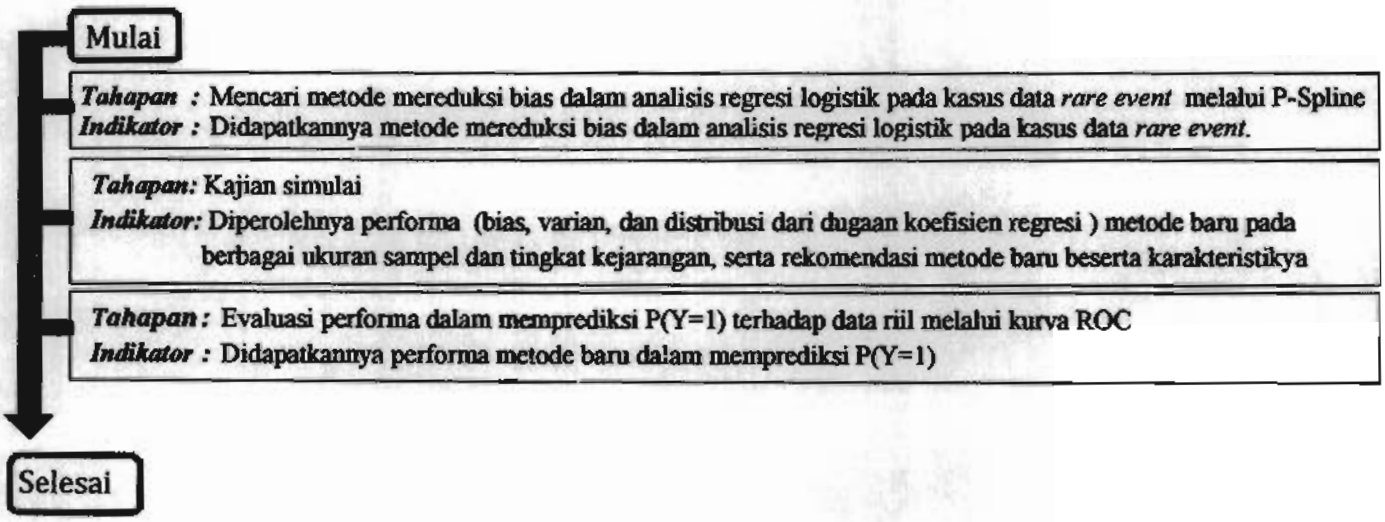
(d). Menerapkan model pada langkah 1.d terhadap data yang diperoleh pada langkah 2.c.

3. Melakukan evaluasi terhadap kinerja metode/algoritma yang diajukan dalam hal ini kemampuan mereduksi bias berdasarkan langkah 2.(c) di atas.

4.2 Luaran dan Indikator Capaian

Luaran yang diharapkan dari penelitian ini mencakup buku ajar, publikasi dalam seminar nasional, publikasi dalam seminar internasional dan jurnal internasional.

Penelitian ini akan dimonitor dan dievaluasi berdasarkan indikator capaian yang disajikan pada Gambar 2.



Gambar 2. Tahapan Penelitian dan Indikator Capaian

BAB V. HASIL YANG DICAPAI

5.1 Pendahuluan

Ada dua pendekatan yang akan diterapkan untuk mengurangi bias pada kasus regresi logistik, yaitu (i) melalui pendekatan semiparametrik dengan P-Spline dan (2) kombinasi bootstrap undersampling dengan model semiparamatrik.

5.2 Mengatasi Masalah Rare Event melalaui Regresi Nonparametrik : P-Spline

Dalam regresi parametrik, bentuk hubungan antara variabel bebas dengan tak bebas telah diketahui kecuali parameter model. Jika model tidak tepat, maka hasilnya adalah nilai dugaan bagi parameter model akan bias. Dalam regresi nonparmetrik, bentuk hubungan variabel bebas dan tak bebas tidak lagi diperlukan. Asumsinya adalah bahwa bentuk hubungannya adalah suatu fungsi, baik linier maupun non linier, yang bentuknya sembarang fungsi. Prinsip dasar dari regresi nonparaterik adalah "biarkan data bicara", dalam arti bahwa bentuk fungsi hubungan akan ditentukan oleh data. Hal ini menyebabkan regresi nonparametrik sangat fleksibel dan secara asimptotik menghasilkan penduga fungsi regresi yang tidak bias ((Eubank, 1999; Green & Silverman, 1994; Suliadi, et al., 2010a,b; Suliadi, et al., 2013; Suliadi, 2014, Suliadi & Kudus, 2015)

Dalam banyak kasus, memodelkan hubungan variabel bebas dan variabel tak bebas secara nonparametrik seringkali tidak relevan karena bentuk hubungan beberapa variabel bebas lainnya sudah diketahui. Dalam kasus ini maka regresi semiparametrik adalah model yang cocok. Dalam regresi semiparametrik, sebagian variabel bebas mempengaruhi tak variabel bebas dalam bentuk fungsi tertentu (diketahui) sedangkan sebagian variabel bebas yang lain mempengaruhi variabel tak bebas secara nonparametrik dalam bentuk sembarang fungsi.

Ada beberapa kelas regresi nonparametrik, diantaranya adalah *local polynomial Kernel* (LPK), regresi spline, penalized spline (P-Spline) dan smoothing spline. Dalam penelitian ini yang dipergunakan adalah P-Spline, karena lebih cocok untuk data besar dan secara komputasi lebih sederhana. Untuk P-spline ini kami merujuk pada Ruppert, et al. (2003), Green & Silverman (1994), dan Wu & Zhang (2006).

5.2.1 P-Spline respon kontinyu

Misalkan n sampel pasangan pengamatan (x_i,y_i) mempunyai model hubungan

$$y_i = f(x_i) + e_i. \qquad (12)$$

Ide dari P-Spline adalah dari ekspansi Taylor. Fungsi f di atas dapat didekati melalui polinomial berorde k. Untuk meningkatkan fleksibilitas dalam memodelkan fungsi non-linier, fungsi f dimodelkan dalam beberapa selang/interval. Jadi daerah fungsi x dibagi menjadi beberapa

selang: $t_1 < t_2 < \ldots < t_K$. Nilai $t_1, t_2, \ldots, t_K$ disebut sebagai knot. P-Spline dibentuk dengan menggunakan *truncated power basis* berderajat k dengan K knot $t_1 < t_2 < \ldots < t_K$:

$$1, x, \ldots, x^k, (x-t_1)^k_+, (x-t_2)^k_+, \ldots, (x-t_K)^k_+ \tag{13}$$

dan fungsi f di atas dinyatakan dalam bentuk

$$\begin{aligned} g(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \beta_{k+1}(x-t_1)^k_+ + \beta_{k+2}(x-t_2)^k_+ + \cdots + \beta_{k+K}(x-t_K)^k_+ \\ &= \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{K} \beta_{k+r}(x-t_r)^k_+ \end{aligned} \tag{14}$$

dengan $w_+ = \text{maksimum}(0,w)$. Misalkan

$$\beta = (\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}, \ldots, \beta_{k+K})^T \;;\; x_i = (1, x_i, \ldots, x_i^k, (x_i-t_1)^k_+, (x_i-t_2)^k_+, \ldots, (x_i-t_K)^k_+)^T$$

$$X = (x_1, x_2, \ldots, x_n)^T;\; \eta = (f(x_1), f(x_2), \ldots, f(x_n))^T;\; y = (y_1, y_2, \ldots, y_n)$$

sehingga $g(x_i) = x^T\beta$.

Solusi bagi β melalui metode kemungkinan maksimum adalah

$$b = (X^TX)^{-1}X^Ty \tag{14}$$

Metode ini disebut sebagai regresi spline. Regresi spline sangat tergantung kepada jumlah knot dan lokasi knot. Oleh karena itu metode untuk menentukan jumlah dan lokasi knot sangat penting daalm regresi spline. Knot terlalu banyak akan menghasilkan fungsi yang terlalu kasar dan ada kemungkinan overfitting. Untuk mengatasi hal tersebut, P-Spline mengatasi hal tersebut dengan memberikan penalty tingkat kekasaran fungsi (*roughness penalty*), sehingga fungsi yang diperoleh tidak terlalu kasar akan tetapi optimal. Oleh karena itu perbedaan P-Spline dengan regresi spline terletak pada *roughness penalty* pada fungsi tujuan (*objective function*). Jadi P-Spline mempunyai model dan basis yang sama dengan regresi spline, tetapi fungsi tujuannya adalah

$$\sum_{i=1}^{n} y_i - x(t_i)^T\beta + \lambda\beta^TG\beta = (y - X\beta)^T(y - X\beta) + \lambda\beta^TG\beta \tag{15}$$

dengan

$$G = \begin{bmatrix} 0_{(k+1)\times(k+1)} & 0_{(k+1)\times K} \\ 0_{K\times(k+1)} & I_K \end{bmatrix}.$$

Fungsi tujuan pada persamaan (15) merupakan fungsi tujuan metode kuadrat terkecil dengan tambahan penalty pada koefisien regresinya. Fungsi tujuan pada (15) disebut sebagai *penalized least square* (PLS), dengan solusi persamaan (15) bagi β adalah

$$\hat{\beta} = (X^TX + \lambda G)^{-1}X^Ty.$$

Nilai dugaan y pada titik t adalah

$$\hat{f}(t) = x(t)^T\hat{\beta}.$$

Bentuk $\beta^TG\beta$ mengukur derajat kekasaran fungsi f dan λ disebut sebagai parameter pemulus (*smoothing parameter*) yang mengontrol tingkat kesesuaian model dengan data (*goodness of fit*)

yang direpresentasikan oleh jumlah kuadrat galat oleh $(y-X\beta)^T(y-X\beta)$ dengan derajat kekasaran kurva *f*.

5.2.2 Regresi logistik semiparametrik berdasarkan P-Spline

Misalkan dari sample berukuran *n* masing-masing diamati respon biner $y_i \in \{0,1\}$ serta variabel bebas $v_{1i}, v_{2i}, \ldots, v_{pi}$ dan r_i. Variabel $v_{1i}, v_{2i}, \ldots, v_{pi}$ mempengaruhi respon y_i secara parametrik sedangkan variabel r_i mempengaruhi y_i secara nonparametrik melalui P-Spline berderajat *k* dengan knot $t_1, t_2, \ldots, t_K$. Bentuk hubungan variabel-variabel tersebut adalah

$$\eta_i = \beta_0 + \underbrace{\delta_1 v_{1i} + \cdots + \delta_p v_{pi}}_{\text{Komp.Parametrik}} + \underbrace{\alpha_1 r_i + \alpha_2 r_i^2 + \cdots + \alpha_k r_i^k + u_1 (r_i - t_1)_+^k + u_2 (r_i - t_2)_+^k + \cdots + u_K (r_i - t_K)_+^k}_{\text{Komp.Nonparametrik: P-Spline berderajat } k} \quad (16)$$
$$= x_i^T \beta + z_i^T u$$

di mana

$$x_i = (1 \;\; v_{1i}, \ldots, v_{1i}, r_i, r_i^2, \ldots, r_i^k)^T; \;\; z_i = [(r_i - t_1)_+^k, (r_i - t_2)_+^k, \ldots, (r_i - t_K)_+^k)^T;$$
$$\beta = (\beta_0, \delta_1, \ldots, \delta_p, \alpha_1, \ldots, \alpha_k)^T; u = (u_1, u_2, \ldots, u_K)^T.$$

Distribusi dari y_i adalah Bernoulli(μ_i) dengan $E(y_i) = \mu_i$ dan $Var(y_i) = \mu_i(1-\mu_i)$ yang merupakan komponen acak. Hubungan antara komponen acak dan komponen sistematik menggunakan fungsi hubung logit

$$logit(\mu_i) = \frac{\mu_i}{1-\mu_i} = \eta_i$$

atau regresi logistik. Misalkan $\theta = (\beta^T, u^T)^T$ dan $X_i = (x_i^T, z_i^T)^T$, sehingga model (16) dapat dituliskan sebagai

$$\eta_i = X_i^T \theta.$$

Dengan menggunakan konsep yang sama pada respon kontinyu di atas, maka fungsi penalized log-likelihoodnya adalah

$$\Pi = l(\theta) - (1/2)\lambda \theta^T G \theta,$$

di mana $l(\theta)$ adalah fungsi log-likelihood untuk θ. Penduga bagi θ diperoleh sebagai solusi dari memaksimumkan *penalized likelihood function*

$$\Pi = L(\theta) - (1/2)\lambda \theta^T G \theta$$

di mana $l(\theta) = \log L(\theta)$ dengan

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid x_i, r_i, \theta).$$

dan

$$G = \begin{pmatrix} 0 & 0 \\ 0 & I_K \end{pmatrix}.$$

Fungsi Π akan maksimum jika $U = \partial\Pi/\partial\theta=0$ dimana

$$U = \frac{\partial L(\theta)}{\partial \theta} - (1/2)\lambda \frac{\partial}{\partial \theta}(\theta^T G \theta)$$
$$= X^T W_1 (y - \mu) - \lambda G \theta$$

Bentuk persamaan penduga di atas tidak lain adalah persamaan penduga pada sub bab 2.1.1 di atas. Solusi dari persamaan di atas tidak dapat diperoleh dalam bentuk tertutup (*closed form*) dan biasanya mengunakan prosedur iterative misalkan Fisher *Scoring Algorithm*:

$$\hat{\theta}^{r+1} = \hat{\theta}^r + I^{-1}U$$

di mana $I = E(-\partial^2\Pi / \partial\theta\partial\theta^T))$ yang diberikan pada sub bab 2.1.1.

5.2.3 Penentuan Knot dan Parameter Pemulus

Dalam P-Spline ada empat komponen yang harus ditentukan untuk melakukan pendugaan parameter model regresi, yaitu (i) derajat (degree) dari P-Spline atau *k*, (ii) jumlah knot atau *K*, (iii) lokasi knot, dan (iv) parameter pemulus, λ. Keempat hal tersebut akan sangat menentukan apakah fungsi yang diperoleh akan optimal ataukah terlalu mulus ataukah terlalu kasar. Jika model terlalu mulus (mendekati linier), maka bias akan besar tapi ragam jadi kecil. Sedangkan jika model terlalu kasar (mendekati interpolasi setiap titik), maka bias jadi kecil tapi ragam jadi besar. Jadi pemilihan knot dan parameter pemulus harus memperhatikan keempat hal tersebut di atas.

Menurut Rupper et al. (2003) dan Wu & Zhang (2006) spline berderajat $k = 2$ atau paling tinggi k = 3 sudah mencukupi untuk hampir semua kasus. Nilai k yang lebih besar dari tiga (k > 3) tidak akan memberikan keuntungan yang berarti karena hasil fungs tidak ada perbedaan yang berarti. Oleh karena itu dalam penelitian ini kami menggunakan P-Spline berderajat 3 (P-Spline degree of 3) atau kubik P-Spline.

a. Penentuan jumlah knot K dan lokasi knot

Metode pertama untuk memilih knot adalah dengan menyediakan knot yang cukup banyak, selanjutnya dengan tehnik seperti pada pemilihan variabel (mis: backward, forward atau stepwise) diperoleh knot yang diinginkan. Dari knot yang diperoleh, dilakukan regresi spline yang biasa (Friedman & Silverman, 1989; Friedman, 1991; Stone, et al., 1997). Metode ini disebut regresi spline. Alternatif lainnya adalah mengambil semua pengamatan r_i yang berbeda sebagai knot, yang akan membawa kepada metode regresi smoothing spline (Green & Silverman, 1994). Sedangkan P-Spline membolehkan knot yang banyak, untuk mengatasi

overfitting, maka harus diberikan penalty terhadap koefisien regresi melalui parameter pemulus, λ.

Ruppert (2002) memberikan beberapa acuan untuk menentukan besarnya nilai K, yaitu

- Ada suatu nilai K sebagai kecukupan minimum, misalnya K_{min}. Jika banyaknya knot kurang dari K_{min} akan menghasilkan penduga dengan bias dan jumlah kuadrat sisaan yang lebih tinggi. Jika banyaknya kno lebih dari K_{min} akan memberikan hasil dugaan yang lebih baik.
- Berbagai Knot yang lebih banyak dari K_{min} akan memberikan hasil yang tidak terlalu jauh berbeda.
- Untuk fungsi tidak bersifat oscilasi, K=min(n/4, 40) akan memberikan hasil yang baik.

Jika banyaknya knot sudah diketahui, maka yang perlu diperhatikan adalah menentukan lokasi knot. Ada dua pendekatan yang umum digunakan untuk menetukan lokasi knot secara otomatis, yaitu

- Eilers & Marx (1996) menganjurkan untuk menggunakan metode interval yang sama dari ruang r (variabel bebas-non parametrik komponen). Misalkan r_{min} dan r_{max} masing-masing adalah data minimum dan maksimum dari variabel bebas r dan banyaknya knot adalah K. Maka dalam interval $[r_{min}, r_{mak}]$ akan dibagi dalam (K+1) selang yang sama. Sehingga knot ke-i adalah

$$t_i = r_{min} + \left[i \times \left(\frac{r_{max} - r_{min}}{K+1} \right) \right], i = 1,2,...,K.$$

- Sedangkan Ruppert (2002) dan Yu & Ruppert (2002) menyarankan untuk menggunakan sampel kuantil yang sama (*equally-spaced sample quantile*), sehingga knot didasarkan jarak kuantil (banyaknya pengamatan dalam interval dua knot berturutan). Misalkan K adalah banyaknya knot, maka knot ke-i ditentukan melaui rumus:
  - Jika lok=$i \times$ (n+1)/(K+1) adalah bulat, maka

$$t_i = r_{[lok]}, i = 1,2,...,K.$$

  - Jika lok= $i \times$ (n+1)/(K+1) adalah tidak bulat, maka lok= $\lfloor i \times (n+1)/(K+1) \rfloor$ dan

$$t_i = \left(\frac{r_{[lok]} + r_{[lok+1]}}{2} \right), \quad i = 1,2,...,K.$$

b. Penentuan parameter pemulus

Dalam P-Spline, parameter pemulus λ sangat menentukan kemulusan kurva dan bias dari dugaan kurva. Jika λ terlalu kecil, maka kurvaakan menjadi sangat kasar akan tetapi bias akan kecil. Sebaliknya jika λ terlalu besar maka kurva akan terlalu mulus tetapi bias akan besar. Untuk data

kontinyu, ada beberapa metode yang dapat dipergunakan untuk memilih parameter pemulus ini, diantaranya CV, GCV, AIC, Mallow-CP. Penentuan λ untuk data non-Gaussian lebih rumit dibandingkan data kontinyu. Metode pemilihan ini pada dasarnya adalah pengembangan dari kasus data kontinyu.

Ruppert, et al. (2003) GCV (generalized cross validation) dan AIC (Akaike Information Criterion) berdasarkan statistik devian. Misalkan A adalah matrik pemulus atau matrik dan D(y, $\hat{y}$) adalah devian dari model berdasarkan data sampel berukuran n. Maka GCV dan AIC berdasarkan devian didefinisikan sebagai

$$GCV_R(\lambda) = \frac{n^{-1}D(y,\hat{y})}{[1-n^{-1}tr(A)]^2},$$

$$AIC_R(\lambda) = n^{-1}[D(y,\hat{y}) + 2tr(A)\phi].$$

Nilai λ yang optimal adalah yang meminimumkan nilai GCV atau AIC.

Xiang & Wahba (1996) memberikan pendekatan untuk CV dan GCV. Misalkan respon y berasal dari keluarga eksponen dengan parameter kanonik $\theta = \eta(x)$. Sehingga $E(y_i) = b'(\eta(x_i))$ dan $Var(y_i) = b''(\eta(x_i)).a(\psi)$. Misalkan $W = \text{Diag}(b''(\eta(x_1),\ldots,b''(x_n)) = \text{Diag}(w_1,\ldots,w_n)$ dan H adalah invers dari matrik Hessian dengan diagonal elemennya adalah h_{ii}. Maka ACV (approximate of cross validation) dan GACV (generalized approximate cross validation) didefinisikan sebagai:

$$ACV_{XW}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y\eta(x_i) + b(\eta(x_i))] + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{ii}y_i[y_i - \mu(x_i)]}{1 - h_{ii}b''(\eta(x_i))}\right]$$

dan

$$GACV_{XW}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y\eta(x_i) + b(\eta(x_i))] + \left[\frac{tr(H)}{n}\right]\frac{\sum_{i=1}^{n} y_i[y_i - \mu(x_i)]}{n - tr(W^{1/2}HW^{1/2})}.$$

Green & Silvermann (1994) juga mengajukan pendekatan dari GCV yang didefinisikan sebagai

$$GCV_{GS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i - x_i^T\hat{f}}{(1-n^{-1}tr(A))^2}\right)^2$$

di mana $z_i = (y-\hat{\mu})g'(\mu_i) + x_i^T\hat{f}$ adalah working response dan A adalah hat matrik.

Pendekatan lain adalah melalui pendekatan dengan adanya hubungan antara P-Spline dengan Mixed Model (Ruppert, et al., 2003; Wu & Zhang, 2006). Dengan adanya hubungan ini, maka pemilihan parameter pemulus bisa menggunakan pendekatan mixed model. Hubungan ini disebabkan model P-Spline (16) dapat dinyatakan sebagai mixed model, yaitu

$$\eta_i = x_i^T \beta + z_i^T u = \tilde{y} = x_i^T \beta + z_i^T u$$

tidak lain adalah mixed model

$$\tilde{y} = X\beta + Zu + \varepsilon$$
$$u \sim N(0,[\sigma^2 / \lambda]I), \varepsilon \sim (0,\sigma^2 I)$$

dengan demikian jika diperoleh Var(u) = Ω^2, maka kita bisa dapatkan $\lambda = \sigma^2 / \Omega^2$.

5.3 Studi Simulasi

Simulasi dilakukan untuk mengevaluasi kinerja dari model semiparametrik dalam mengatasi masalah bias ketika ada masalah *rare event* pada data. Hal ini didasari pada kenyataan bahwa model nonparamaterik dan model semiparamterik bekerja berdasarkan pada data, dalam arti mengikuti pola data dan biarkan data bicara. Sehingga dapat mengatasi bias pada model parametrik ketika model yang digunakan tidak tepat. Karakteristik tersebut berlaku untuk model-model dalam kelas *generalized linear model* (GLM).

Dalam simulasi ini kami menggunakan model

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sin(4\pi Z)$$

dengan $\beta_0 = 1, \beta_1 = 1$, dan $\beta_2 = -1$. Fungsi hubung yang digunakan adalah

$$\text{logit}(\mu) = \log\left[\frac{\mu}{1-\mu}\right] = \eta.$$

Variabel respon Y dibangkitkan dari distribusi Bernoulli(μ) dimana variabel bebas X1 dibangkitkan dari distribusi Normal(0,1) sebagai representasi dari variabel kontinyu sedangkan variabel X2 dibangkitkan dari distribusi Bernoulli(0.5) sebagai representasi variabel diskrit dalam bentuk variabel dummy. Variabel bebas nonparametriknya adalah Z yang dibangkitkan dari distribusi U(0, π). Kami membangkitkan 10,000,000 pengamatan sebagai perwakilan populasi. Selanjutnya pengamatan dengan respon Y=0 kami pisahkan dengan pengamatan dengan respon Y=1, yang masing-masing dikelompokkan dalam group G0 dan G1.

Misalkan diinginkan set data dengan ukuran sampel *n*, dengan persentase Y=1 (*rare event*) sama dengan *p*. Maka set data tersebut diperoleh dengan cara sebagai berikut:

- Ambil secara acak dari G1 sebanyak $n1 = n*p$;
- Ambil secara acak dari G0 sebanyak $n0 = n*(1-p)$;
- Gabungkan kedua set data tersebut sebagai set data yang diinginkan.

Untuk setiap set data, kami melakukan pendugaan model semiparametrik seperti pada persamaan (16) dengan orde k=2 (kuadratik) dan banyaknya knot adalah 35 dengan posisi knot adalah didasarkan kuantil yang sama (*equally spaced quantile*).

Ukuran sampel yang kami gunakan adalah n = *200, 500,* dan *1000;* dengan p = #(Y=1)/n adalah 5%, 10%, 20% dan 50%. Untuk setiap kombinasi n dan p, kami mengulang sebanyak 350 kali. Untuk setiap set data, kami lakukan pendugaan dua kali, yaitu menggunakan model semiparametrik

$$\hat{\eta} = b_0 + b_1 X_1 + b_2 X_2 + f(Z)$$

dan menggunakan model parametrik

$$\hat{\eta} = b_0 + b_1 X_1 + b_2 X_2 + b_3 Z .$$

Selanjutnya kami mengevaluasi perilaku bias dan ragam penduganya.

Tabel 1. Besarnya Bias Dugaan Koefisien Regresi Model semiparametrik (Semi) dan Model Parametrik (Param) untuk Beberapa Ukuran Sampel dan Tingkat Kejarangan.

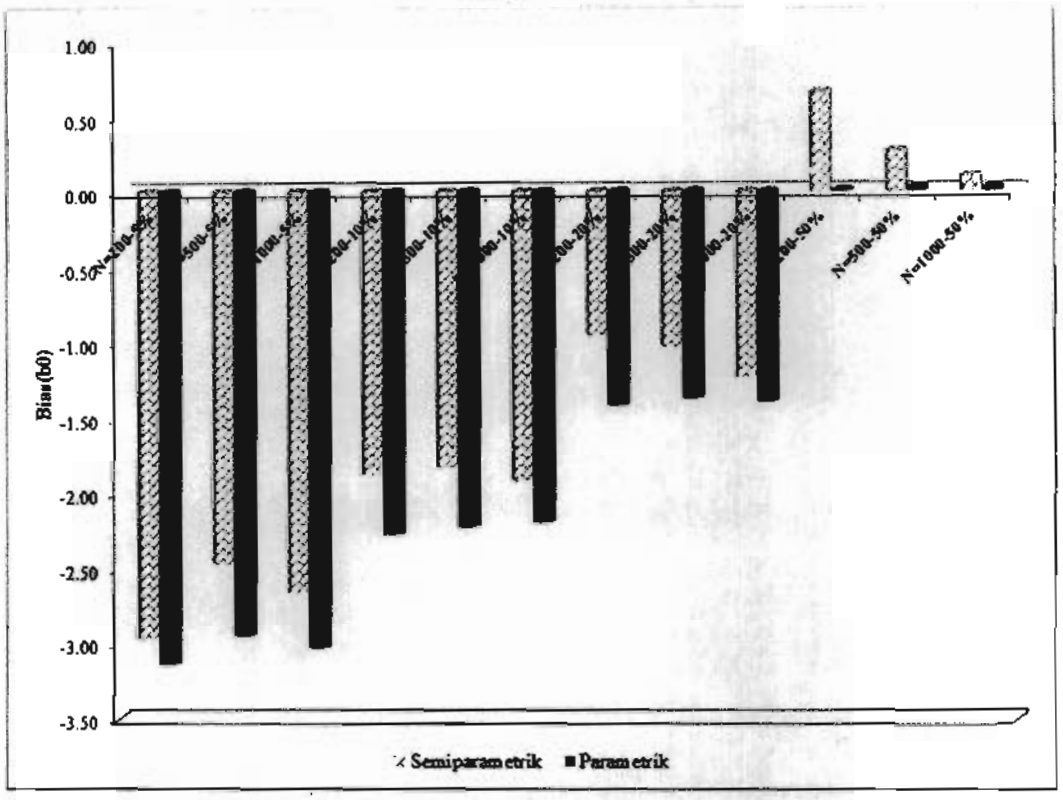
Ukuran sampel	Persen Y=1	β_0		β_1		β_2	
		Semi	Param	Semi	Param	Semi	Param
200	5	-2.963	-3.138	0.039	-0.002	0.028	0.063
500	5	-2.467	-2.949	-0.057	-0.087	-0.012	0.015
1000	5	-2.652	-3.027	-0.028	-0.065	0.003	0.049
200	10	-1.874	-2.271	0.013	-0.031	0.024	0.067
500	10	-1.828	-2.224	-0.011	-0.058	0.015	0.049
1000	10	-1.915	-2.191	0.013	-0.037	0.015	0.061
200	20	-0.948	-1.417	0.007	-0.035	0.044	0.088
500	20	-1.026	-1.370	0.005	-0.050	0.022	0.072
1000	20	-1.227	-1.392	-0.020	-0.085	0.034	0.104
200	50	0.663	0.009	0.008	-0.040	0.022	0.068
500	50	0.271	0.024	-0.003	-0.062	-0.027	0.039
1000	50	0.103	0.023	-0.009	-0.081	-0.004	0.064

Catatan: ±30% pada pendugaan model semiparametrik tidak konvergen.

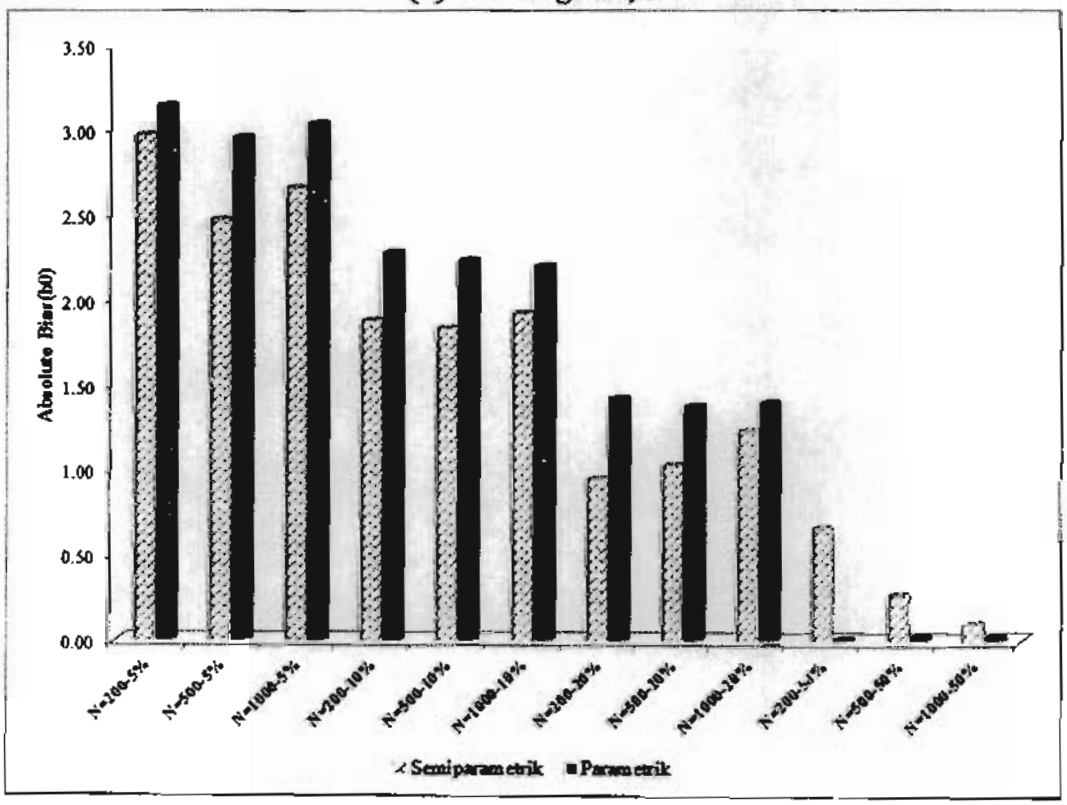
a. Perilaku Bias

Besarnya bias dugaan koefesien regresi disajikan pada Tabel 1. Dari Tabel 1 ini terlihat bahwa kedua pendekatan model menghasilkan nilai dugaan untuk β_0 yang bias. Sedangkan untuk dugaan koefisien variabel bebas X1 dan X2 (β_1 dan β_2) bersifat bias tetapi sangat kecil, bahkan bisa dianggap tidak bias. Hal ini sesuai dengan referensi yang ada di Bab II, yaitu

bahwa ketika ada masalah *rare event*, koefisien variabel bebas bersifat tidak bias sedangkan nilai dugaan koefisien intersep (β_1) bersifat bias.



(a) Bias dugaan β_0



(b) Absolute bias dugaan β_0

Gambar 3. Besarnya Bias dan Absolut Bias Dugaan β_0 untuk Pendugaan dengan Model Semiparametrik dan Model Parametrik

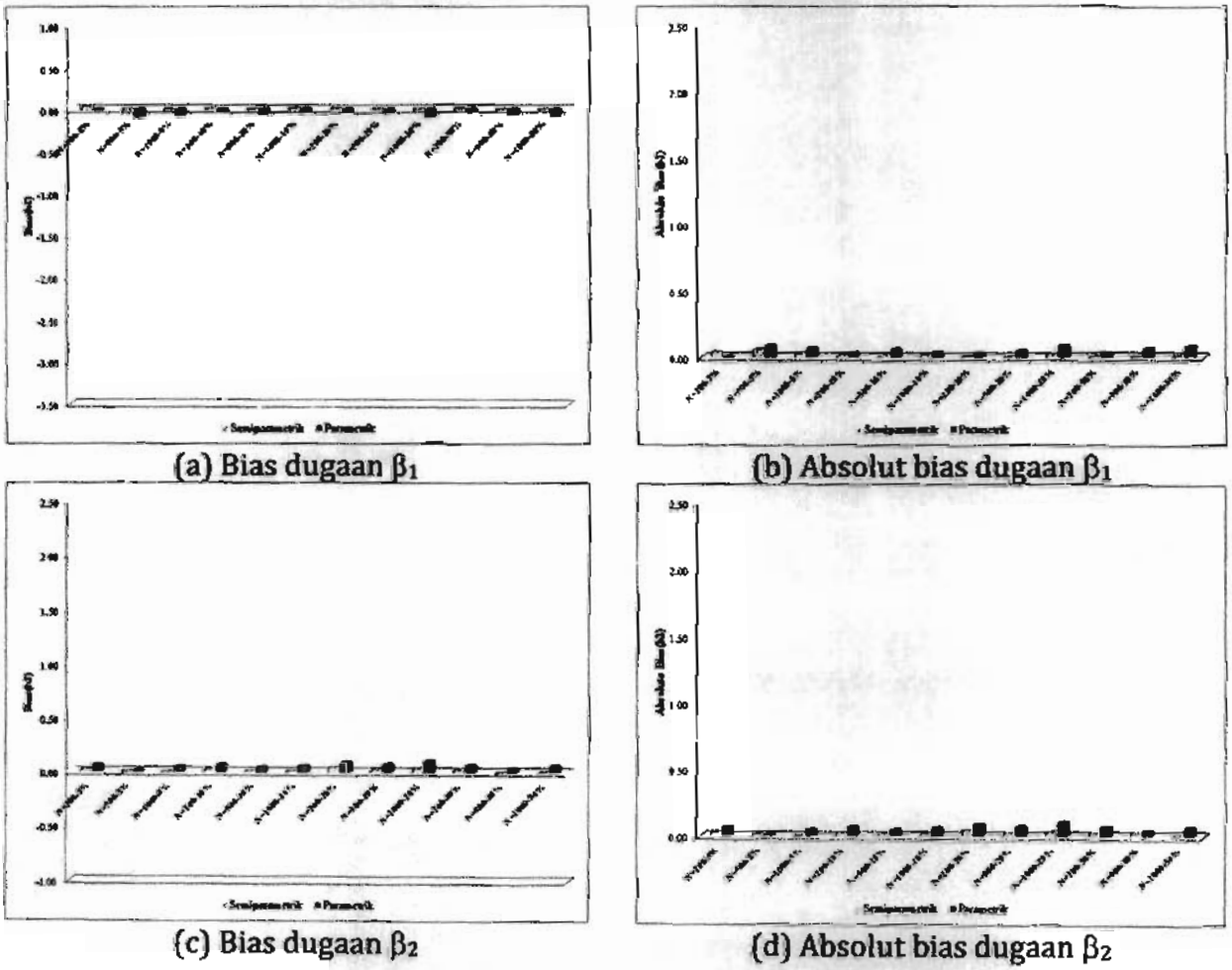
Besarnya bias dan arah bias dugaan β_0 dapat dilihat pada Gambar 3. Dari gambar tersebut terlihat bahwa semakin tinggi tingkat kejarangan (*level of rareness*) (semakin kecilnya persentase Y=1), maka bias akan semakin besar dan arah bias adalah negatif. Hal ini juga sesuai dengan referensi yang ada. Dari gambar tersebut terlihat bahwa bias dugaan β_0 untuk persentase Y=1 adalah 5%, 10%, 20% terjadi bias negatif dan besarnya bias semakin menurun. Hal ini berlaku untuk kedua model, semiparametrik dan parametrik. Meskipun demikian ada satu kecenderungan, yaitu ketika ada masalah *rare event*, besarnya bias model semiparametrik lebih kecil dibandingkan bias untuk model parametrik. Hal ini berlaku untuk semua ukuran sampel dan berbagai tingkat kejarangan. Hasil ini menunjukkan bahwa model semiparametrik dapat mereduksi bias (dalam hal ini untuk koefisien intersep) karena menghasilkan dugaan dengan bias yang lebih kecil dibandingkan dengan model parametrik. Dari Tabel 1 dan Gambar 3 terlihat bahwa meskipun medel semiparametrik menghasilkan bias yang lebih kecil dibandingkan dengan model parametrik tetapi perbedaannya tidak terlalu besar. Dari sini terlihat bahwa, ketika ada masalah *rare event*, kemampuan model semiparametrik untuk mereduksi bias masih rendah.

Hasil di atas sepertinya agak kontradiktif dengan karakteristik model-model non dan semiparametrik yang bersifat *follow the data* dan membiarkan data bicara. Fenomena ini dapat dijelaskan dari sisi model dan struktur data. Model non dan semiparametrik dapat mengatasi bias ketika model parametrik yang digunakan tidak tepat atau tidak cocok dengan data, melalui mekanisme bentuk fungsi yang bebas yang tidak terikat dalam bentuk persamaan tertentu. Permasalahan *rare event* berbeda dengan permasalahan ketidaktepatan atau ketidakcocokan model yang digunakan. Permasalahan ada pada struktur data respon yang tidak seimbang antara banyaknya pengamatan dengan Y=0 (kelas mayoritas) dan Y=1 (kelas minoritas), yang mengakibatkan P(Y=0) cenderung akan menarik P(Y=1). Dengan kata lain dugaan P(Y=0) akan bias ke atas dan P(Y=1) akan bias ke bawah (King & Zheng, 2001; Qiu, et. al, 2013). Karena nilai variabel bebas bervariasi, maka mekanisme bias P(Y=0) dan P(Y=1) adalah melalui koefisen intersep. Hal ini berlaku baik untk model parametrik maupun model non dan semiparametrik. Sehingga fleksibilitas model semiparametrik masih tidak mempu untuk mereduksi bias dalam jumlah besar.

Satu hal yang cukup menarik adalah bahwa pada ukuran sampel besar (200 atau lebih), efek ukuran sampel terhadap bias adalah kecil. Pada berbagai ukuran sampel yang

dicobakan, n = 200, 500 dan 1000, tidak ada perbedaan yang berarti terhadap besarnya bias. Yang paling berperan dalam besarnya bias koefisien intersep adalah tingkat kejarangan dari Y=1 atau tingkat ketidakseimbangan antara banyaknya Y=0 dengan banyaknya Y=1.

Persentase Y=1 sebesar 50% adalah merupakan kondisi ideal, yang seharusnya menghasilkan dugaan koefisien intersep yang tidak bias. Hasil dari model parametrik sesuai dengan yang diharapkan, di mana bias yang ada sangat kecil yang bisa dianggap tidak bias. Sedangkan bias koefisien intersep untuk model semiparametrik, meskipun kecil, tetapi lebih besar dibandingkan bias dari model parametrik, akan tetapi besarnya bias semakin menurun dengan semakin besarnya ukuran sampel.



(a) Bias dugaan β_1 (b) Absolut bias dugaan β_1

(c) Bias dugaan β_2 (d) Absolut bias dugaan β_2

Gambar 4. Besar Bias dan Absolute Bias Dugaan β_1 dan β_2 untuk Berbagai Ukursan Sampel dan Tingkat Kejarangan

Besarnya bias dan absolut bias dugaan koefisien regresi variabel bebas X1 dan X2 disajikan pada Gambar 4. Dari Tabel 1 dan Gambar 4 ini terlihat bahwa dugaan koefisien variabel bebas biasnya kecil sekali, yang dapat dianggap bersifat tidak bias. Hal ini berlaku baik untuk model semiparametrik dan model parametrik, berbagai tingkat kejarangan dan

berbagai ukuran sampel yang dicobakan. Meskipun demikian, jika diamati lebih seksama tampak ada pola bahwa bias untuk model semiparametrik lebih kecil dibandingkan dengan model parametrik.

Tabel 2. Ragam Dugaan Koefisien Regresi Model semiparametrik (Semi) dan Model Parametrik (Param) untuk Beberapa Ukuran Sampel dan Tingkat Kejarangan.

Ukuran sampel	Persen Y=1	β_0		β_1		β_2	
		Semi	Param	Semi	Param	Semi	Param
200	5	1.315	0.539	0.214	0.195	0.645	0.613
500	5	0.420	0.181	0.072	0.067	0.201	0.191
1000	5	0.268	0.107	0.030	0.028	0.120	0.116
200	10	0.835	0.204	0.090	0.080	0.216	0.190
500	10	0.303	0.097	0.034	0.034	0.114	0.111
1000	10	0.106	0.044	0.021	0.019	0.059	0.059
200	20	0.481	0.180	0.056	0.052	0.146	0.145
500	20	0.361	0.079	0.020	0.019	0.085	0.082
1000	20	0.090	0.030	0.009	0.007	0.045	0.043
200	50	0.524	0.109	0.032	0.029	0.124	0.118
500	50	0.239	0.048	0.010	0.010	0.062	0.061
1000	50	0.120	0.026	0.006	0.007	0.023	0.025

Catatan: ± 30% pada pendugaan model semiparametrik tidak konvergen.

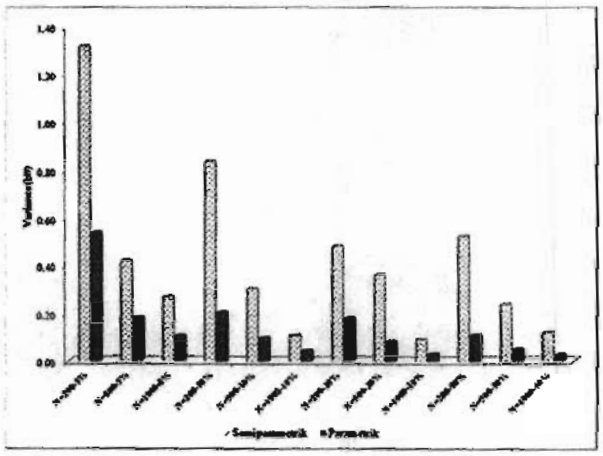
b. Perilaku Ragam Penduga

Ragam penduga koefisien regresi untuk model semiparametrik dan model parametrik disajikan pada Tabel 2 dan Gambar 5. Ada beberapa pola yang menarik yang diperoleh dari tabel dan gambar tersebut. Pola-pola tersebut diantaranya merupakan pola yang sesuai dengan aspek teoritis dalam pendugaan parameter.

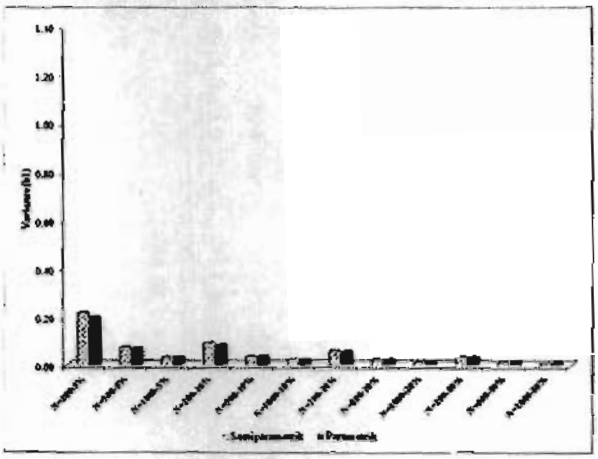
Koefisien b_0 memiliki ragam terbesar yang diikuti oleh ragam b_2 kemudian ragam b_1. Koefisien b_1 memiliki ragam terkecil karena b_1 merupakan koefisien dari variabel yang nilainya bervariasi (*varying covariate*) sehingga menghasilkan penduga yang lebih konsisten dibandingkan dengan variabel X_2 yang merupakan variabel. Variabel X2 adalah variabel biner dengan hanya memiliki dua kemungkinan nilai, sedangkan intersep sebenarnya memiliki variabel bebas dengan dengan hanya satu nilai, yaitu "1". Hal ini menyebabkan ragam b_0 > ragam b_2 > ragam b_1, sebab penduga koefisien regresi sangat dipengaruhi oleh tipe kovariatnya (Suliadi, et al, 2013).

Dilihat dari ukuran sampel, terlihat bahwa dengan semakin besarnya ukuran sampel maka ragam juga semakin besar. Hal ini sesuai dengan kaidah statistika, bahwa jika ukuran sampel semakin besar, maka ragam penduga akan semakin kecil ragamnya.

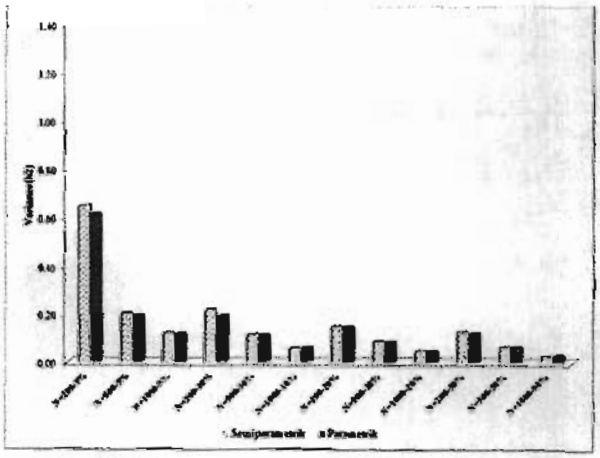
Dilihat dari model yang digunakan, terlihat tidak ada perbedaan ragam penduga β_1 dan β_1 yang mencolok antara model semiparametrik dan model parametrik. Sedangkan untuk penduga β_0, ragam untuk model semiparametrik lebih besar dibandingkan dengan yang diperoleh dari model parametrik. Hal ini adalah realistik dikaitkan dengan bias b_0, di mana bias b_0 dari model semiparametrik lebih kecil dari pada bias b_0 dari model parametrik, sehingga ragam b_0 model semiparametrik akan lebih besar dibandingkan ragam b_0 dari model parametrik.



(a) Ragam b_0



(b) Ragam b_1



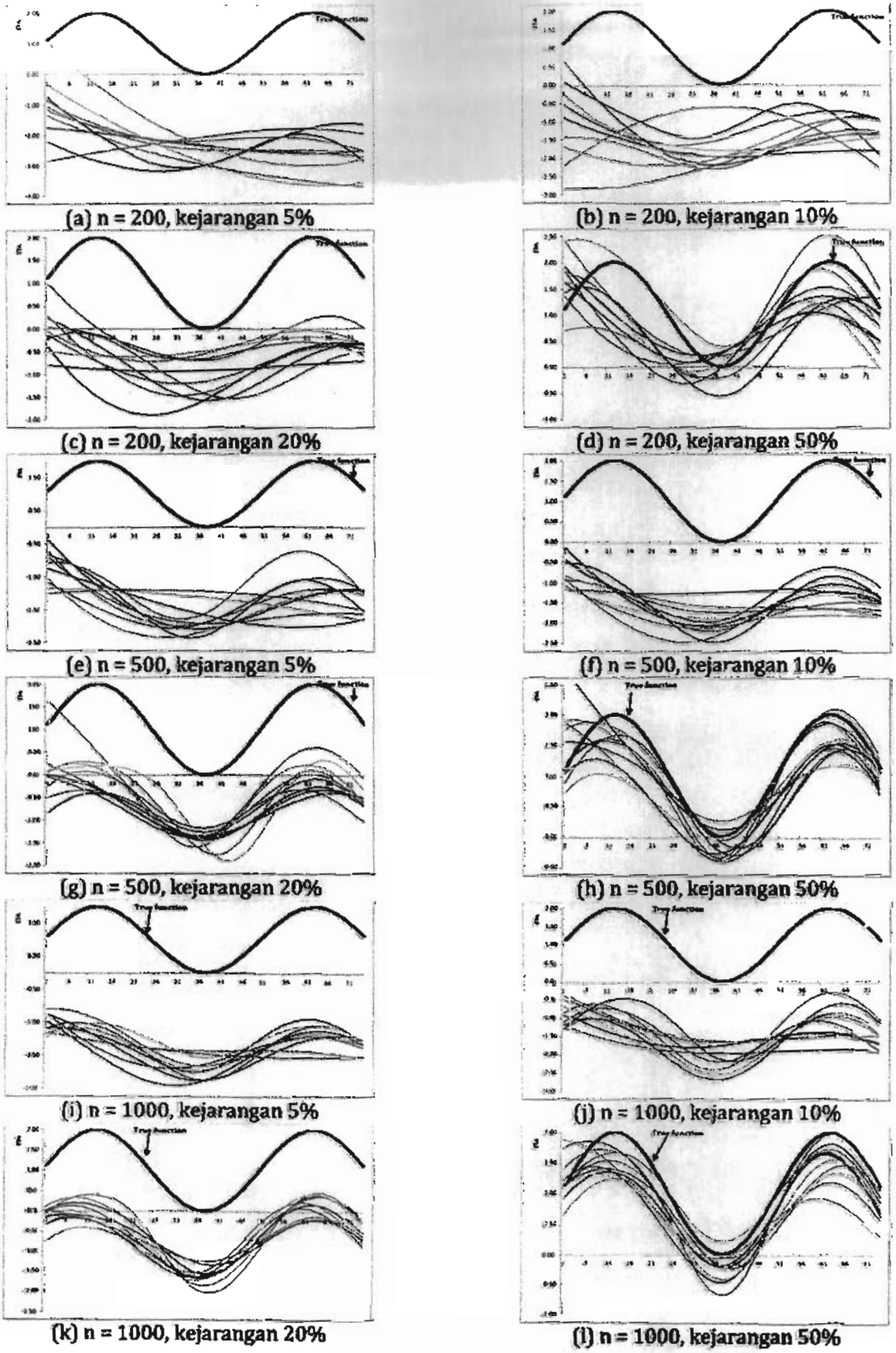
(c) Ragam b_2

Gambar 5. Penduga Koefisien Regresi Model Semiparametrik dan Parametrik untuk Berbagai Ukuran Sampel dan Tingkat Kejarangan

b. Perilaku Komponen Nonparametrik

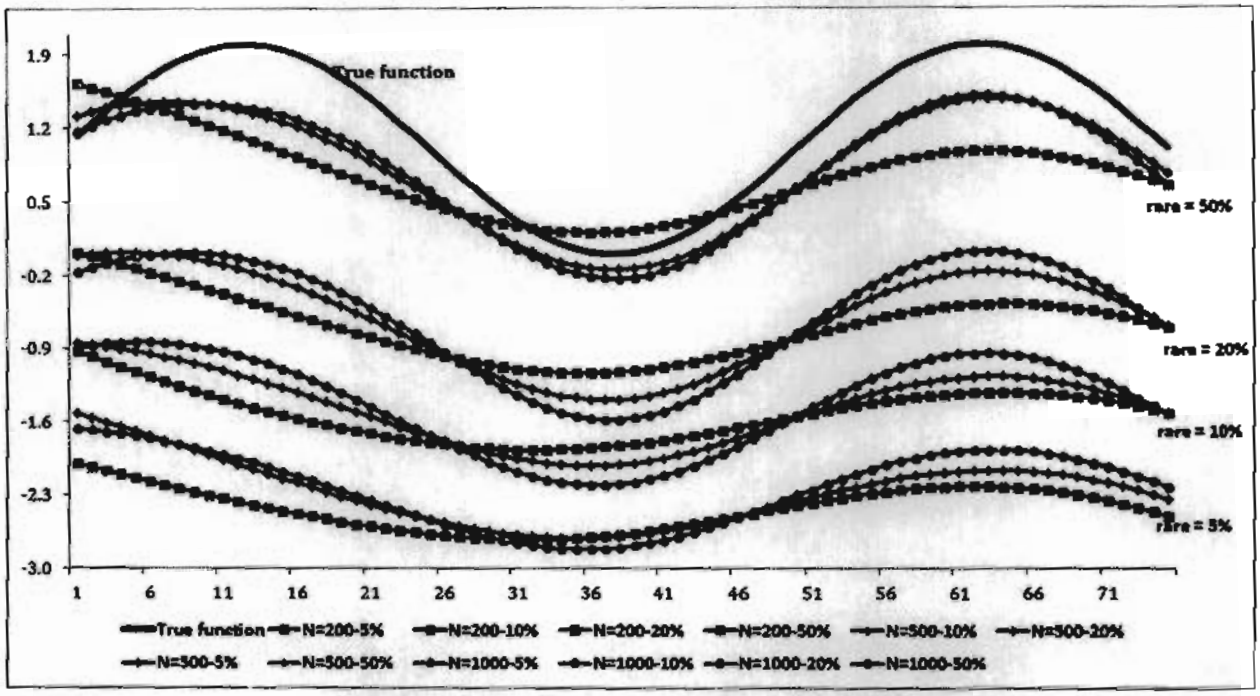
Kompenen nonparametrik yang akan di evalusi merupakan bagian dari persamaan (16), yaitu

$$f(x) = \beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 (z - t_1)_+^2 + \beta_4 (z - t_2)_+^2 + \cdots + \beta_{37} (x - t_K)_+^2.$$

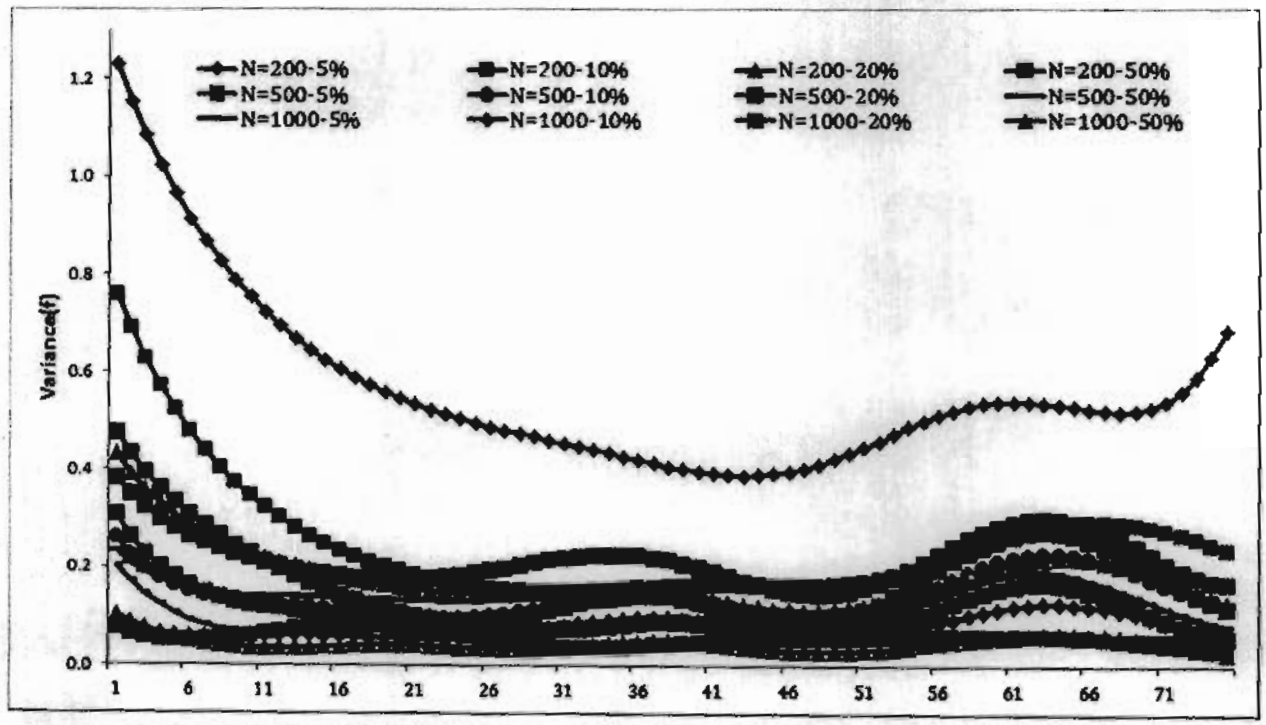


Gambar 6. Kurva Dugaan Komponen Nonparametrik 10 Ulangan Pertama untuk Beberapa Ukuran Sampel dan Tingkat Kejarangan

Kami mengambil 75 titik amatan dari variabel Z untuk dievaluasi terutama perilaku bias dan ragamnya. Besarnya bias dan absolut bias dari komponen nonparametrik dapat dilihat pada Lampiran XX. Gambar 6 menyajikan dugaan kurva untuk 10 amatan pertama. Sedangkan kurva dugaan untuk setiap kombinasi ukuran sampel dan tingkat kejarangan dapat dilihat pada Gambar 7.



Gambar 7. Bias Komponen Nonparametrik

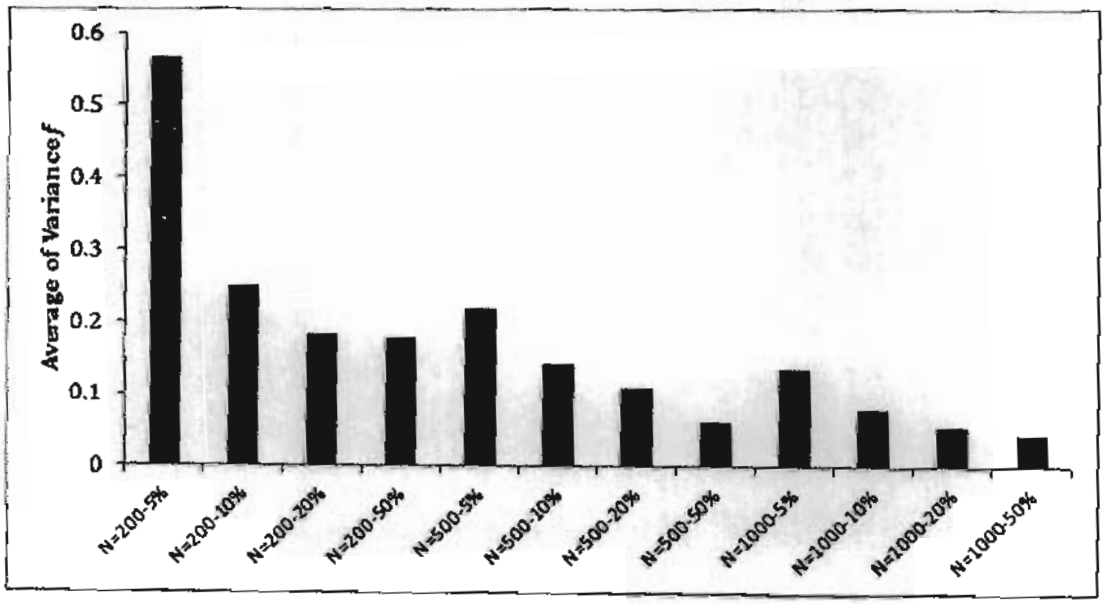


Gambar 8. Ragam Nilai Dugaan Kurva Nonparametrik untuk 75 Titik

Perlu dicatat di sini bahwa komponen nonparametrik di sini mengandung unsur intersep atau b_0. Penyumbang bias terbesar adalah adanya *rare event*, semakin tinggi tingkat kejarangan (persentase Y=1 semakin kecil) maka bias akan semakin besar. Sedangkan pengaruh ukuran sampel terhitung kecil. Tidak ada perbedaan bias yang berarti untuk ketiga ukuran sampel.

Tabel 3. Rata-rata Ragam dari Dugaan Kurva Nonparametrik dari 75 Titik

Ukuran Sampel	Persentase Y=1	Rata-rata Ragam
200	5	0.566
200	10	0.250
200	20	0.182
200	50	0.178
500	5	0.218
500	10	0.142
500	20	0.109
500	50	0.061
1000	5	0.136
1000	10	0.079
1000	20	0.056
1000	50	0.045



Gambar 9. Rata-rata Ragam 75 Titik pada Kurva Nonparametrik

Ragam dari nilai dugaan 75 titik pada kurva nonparametrik dapat dilihat pada Gambar 8, sedangkan rata-ratanya dapat dilihat pada Tabel 3 dan Gambar 9. Dari gambar dan tabel tersebut terlihat ada pola ragam dugaan terkait dengan tingkat kejarangan dan ukuran

sampel. Terlihat bahwa semakin rendah tingkat kejarangan maka ragam dugaan komponen nonparametrik akan semakin rendah dan sebaliknya semakin tinggi tingkat kejarangan ragam penduga komponen nonparametrik juga semakin tinggi.

Faktor lain yang mempengaruhi ragam komponen ini adalah ukuran sampel. Tampak terlihat bahwa semakin besar ukuran sampel, maka ragam akan semakin kecil. Hal ini konsisten dengan teori statistik, bahwa ragam suatu peduga dipengaruhi oleh ukuran sampel.

BAB VI. KESIMPULAN DAN SARAN

6.1 Kesimpulan

Model semiparametrik untuk memodelkan regresi logistik ketika ada masalah *rare event* menghasilkan penduga yang tidak bias untuk koefisien variabel bebas. Model ini juga dapat mereduksi bias koefisien intersep, akan tetapi kemampuan reduksinya masih rendah karena bias nilai dugaan untuk koefisien intersep masih cukup besar meskipun tidak sebesar model parametrik. Bias koefisien intersep yang lebih kecil pada model semiparametrik dibandingkan pada model parametrik berimplikasi ragam dugaannya yang menjadi lebih besar.

6.2 Saran

Pendugaan model semiparamerik menggunakan SAS Macro Glimmix yang menggunakan pendekatan *generalized linear mixed model* untuk menduga model regresi P-Spline dalam kelas GLM. Dari pengalam yang diperoleh dalam simulasi pada penelitian ini, ada sekitar 30% replikasi yang tidak konvergen. Kami menyarankan untuk menggunakan program yang dilandaskan pada model P-Spline sendiri dalam pendugaan parameter model meskipun diperlukan waktu yang agak lama untuk menduga satu model.

DAFTAR PUSTAKA

Dobson, A. J. (2002). An Introduction to Generalized Linear Models. Chapman and Hall, New York, USA, 2nd edition.

Eiler, P.H.C. and Marx, B.D. (1996). Flexible Smooting With B-Splines and Penalities. Statistical Science, 11, 89-121.

Eubank, R. L. (1999). Nonparametric Regression and Smoothing Spline. Marcel Dekker, Inc, New York, USA, 2nd edition.

Fahrmeir, L. and Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models. Springer-Verlag, New York, USA, 2nd edition.

Fithian, W. and Hastie, T. (2014). Local Case-Control Sampling: Efficient Subsampling In Imbalanced Data Sets. The Annals of Statistics, 42, 1693–1724.

Friedman, J. H. and Silverman, B.W. (1989). Flexibel Parsimounious Smoothing and Additive Modelling. Technometrics, 31, 3-39.

Friedman, J.H. (1991). Multivariate Adaptive Regression Spline. The Annal of Statistics, 19, 1-141.

Green, P. J. and Silverman, B. W. (1994). Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach. Chapman & Hall/CRC, New York, USA.

Guns, M. and Vanacker, V. (2012). Logistic Regression Applied to Natural Hazards: Rare Event Logistic Regression with Replication. Nat. Hazards Earth Syst. Sci, 12, 1937-1947.

King, G. and Zheng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9,137-163.

King, G. and Zheng, L. (2001b). Improving Forecast of State Failure. World Politics, 54 (4), 623-658.

Kudus, A., Suliadi, Wachidah, L., Ishmatullah, H. dan Nurwahidah, A.I. (2015). Pengaruh Multikolinier dan Kejadian Jarang (Rare Events) Terhadap Kinerja Model Regresi Logistik. Laporan Penelitian. Bandung: Lembaga Penelitian dan Pengabdian Kepada Masyarakat, Universitas Islam Bandung.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman & Hall. London, UK.

Peraturan Bank Indonesia No. 11/25/PBI/2009 tentang Perubahan atas Peraturan Bank Indonesia No. 5/8/PBI/2003 tentang Penerapan Manajemen Risiko Bagi Bank Umum.

Qiu, Z., Li, H., Su, H., Ou, G., and Wang, T. (2013). Logistic Regression Bias Correction for Large Scale Data with Rare Events. In Advanced Data Mining and Applications. Springer-Verlag, Berlin, Germany. pp 133-143.

Quigley, J., Bedford, T., and Walls, L. (2007). Estimating Rate of Ocurrence of Rare Events with Empirival Bayes: A Railway Application. Reliability Engineering & System Safety, 92, 619-627.

Rezac, M. 2011. How to Measure the Quality of Credit Scoring Models. Journal of Economic and Finance, 5, 486-507.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). Semiparametric Regression. Cambridge University Press, Cambridge, UK.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. Journal of Computational and Graphical Statistics, 11, 735-757.

Sei, T. (2014). Infinitely imbalanced binomial regression and deformed exponential families. Journal of Statistical Planning and Inference, 149, 116–124.

Siddiqi, N. 2006. Credit Risk Scorecard. Developing and Implementing Intellegent Credit Scoring. John Willey & Sons, New Jersey, USA.

Stone, C.J, Hansen, M., Kooperberg, C., and Truong, Y. K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling, The Annal of Statistics, 25, 1371-1470.

Suliadi, Ibrahim, N. A., Daud, I. and Krishnarajah, I. S., (2010). Nonparametric Regression for Longitudinal Binary Data Based on GEE-Smoothing Spline. Journal of Applied Statistics and Probability, 5, 77-93

Suliadi, Ibrahim, N. A., Daud, I. and Krishnarajah, I. S., (2010b). GEE-Smoothing Spline for Semiparametric Estimation of Longitudinal Binary Data. International Journal of Applied Mathematics and Statistics, 18, 82-95.

Suliadi, Ibrahim, N. A. and Daud, I. (2013) Semiparametric Estimation with Profile Algorithm for Longitudinal Binary Data. Communication in Statistics-Simulations and Computations, 42, 138-152.

Suliadi and Kudus, A. (2015). Performance of Covariate-Based Partitioning Goodness of Fit Test for Semiparametric Logistic GEE Regression. International Journal of Applied Mathematics and Statistics, 53, 44-54.

Suliadi, Kudus, A., Yanti, T.S., Permadi, A. D., dan Rohayati, A. (2016). Metode Mereduksi Bias Pada Regresi Logistik Ketika Ada Masalah Data Rare Event. Laporan Penelitian Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Islam Bandung.

Suliadi. (2014). Testing of Goodness of Fit in Semiparametric Logistic Regression Models of Correlated Binary Data. International Journal of Applied Mathematics and Statistics, 52, 141-151.

Weiss, G.M. and Hirsh, H. (2000). Learning to Predict Extremely Rare Events. In AAAI Workshop on Learning from Imbalance Data Sets, pp 64-68.

Wistara, R.R.A. (2015). Regresi Logistik pada Data Rare Event. Skripsi tidak dipublikasikan. Bandung: Program Studi Statistika, Fakultas Matenatika dan Ilmu Pengeahuan Alam, Universitas Islam Bandung.

Wu, H. and Zhang, J. T. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis. John Wiley & Sons, New Jersey, USA.

Xiang, D. and Wahba, G. (1996). A Generalized Approximate Cross Validation For Smoothing Splines With Non-Gaussian Data. Statistica Sinica, 6,675-692.

Yap, B. W., Abd Rani, K., Abd Rahman, H. A, Fong, S., Khairudin, Z., Abdullah, N. N. (2014). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Volume 285 of the series Lecture Notes in Electrical Engineering pp 13-22, Springer Science+Bussiness Media, Singapore.

Yu, Y., and Ruppert, D. (2002), Penalized Spline Estimation for Partial Linear Single-Index Models, Journal of the American Statistical Association, 97, 1042-1054.

Lampiran 1. Acceptance Letter dan Artikel Ilmiah

ICoMSE 2017

ACCEPTANCE LETTER

July 19, 2017

Author (s) : Suliadi Suliadi
Paper Title : Performance Of Semiparametric Modeling In Reducing Bias On Logistic Regression With Rare Event Data
ICOMSE ID : 12017071

Dear Author (s),

This is to inform you that your paper has been accepted for presentation at the 1st International Conference on Mathematics, Science, and Education (ICoMSE 2017), which will take place in Malang, Indonesia on August 29-30, 2017.

The abstract has been peer reviewed by the editorial committee of the conference. If you have not sent a full paper yet, and wish to have your paper included in the conference proceedings, please submit a full paper through:

http://icomse.fmipa.um.ac.id/index.php/paper-submission

no later than September 16, 2017. Ensure that your paper following the ICoMSE 2017 template full paper.

We suggest you to make an early registration and get an advantage of the reduced fee. Early registration will be effective till August 14, 2017. Please make your payment through the following accounts:

Bank name	: BNI
Swift code	: BNINIDJAXXX
Brach name	: Cab. Malang
Account number	: 0387338108 – IDR / USD
Account holder	: Ibu Nursasi Handayani

To secure your paper to be included in the program, please send **the copy of the transfer receipt** (either scanned or photographed) and only for international/local student have to send **the copy of the student identitly card** no later than August 26, 2017 to website:

http://icomse.fmipa.um.ac.id/index.php/payment/

Thank you for participating in ICoMSE 2017. For all information regarding the conference, please kindly visit the conference website: *http://icomse.fmipa.um.ac.id.* We hope that you are able to attend the conference, and look forward to seeing you at the conference.

Best Regards,

Hadi Suwono
Chairman of the ICoMSE 2017 Organizing Committee.

Performance of Semiparametric Modeling in Reducing Bias on Logistic Regression with Rare Event Data

Suliadi[1,a)], Siti Sunendiari[1,b)] and Aceng K. Mutaqin[1,c)]

[1)]Dept. of Statistics - Bandung Islamic University
Jl. Ranggamalela No. 1 Bandung
West Java - Indonesia

[a)]Corresponding author: suliadi@gmail.com
[b)]diarisunen22@gmail.com
[c)]aceng.k.mutaqin@gmail.com

Abstract. Logistic regression is commonly used to model binary data in which the response for each observation has only two possibilities, "success" and "fail" . It uses maximum likelihood estimation (MLE) to estimate the regression parameter. This method has been known that has good properties under ideal conditions. But the MLE gives biased estimate for regression coefficients if the number of "success" and "fail" are greatly unbalanced. In this paper we propose to use modeling the data semiparametrically, since it is well known that non and semiparametric models have good capability in reducing bias. We evaluate the performance of semiparametric model in reducing bias on logistic regression through simulation study. We obtained that semiparametric model has capability in reducing bias on logistic regression when there is rare event problem on the data, specifically on coefficient of intercept, b_0. However this capability is not large and followed by increasing the variance of b_0.

INTRODUCTION

Binary response data (Y = 0 or 1) is commonly modeled using logistic regression with maximum likelihood estimation (MLE) used as the method to estimate the regression parameter. Property of the parameters estimate of this method has been known well. The estimate of this method is unbiased and has minimum variance. However, this unbiased property holds if the proportion of success (Y=1) and proportion fail (Y=0) has no large different (McCullagh & Nelder, 1989, Chap 4 & 15). If the proportions of those two categories largely different, then MLE results bias estimate specifically the coefficient of intercept. This implies the estimate of P(Y=0) and P(Y=1) are also biased (King & Zeng, 2001; Qiu, et al, 2013). The unbalanced of the proportion of Y=0 and Y=1 is known as rare event problem. Bias of the estimate is considered severe if the proportion of the minority category or class is 10% or less. King & Zeng (2001,2001b), Qiu, et al (2013), and others have given several cases in rare event: war, landslides, fraudulent of credit cards, international conflict, oil spill etc. Other cases are the failure of communication tools (Weiss and Hirsh, 2000), even the case of the derailing of fires (Quigley et al., 2007)

Several procedures have been proposed to correct this bias. McCullagh& Nelder (1989) proposed to correct the estimate of regression coefficients, whilst King & Zeng (2001) proposed to correct the intercept coefficient. Qiu, et al (2013) used different approach, by using weighted maximum likelihood that employed to the reconstructed data and combine with correction of McCullagh & Nelder's and King & Zeng's methods. However those methods are appropriate if the sample size is small, i.e $n \leq 200$. Suliadi,et al (2016) proposed different approach to overcome the bias problem by combining undersampling method and bootstrap method.

In this paper we apply semiparametric model to overcome bias problem in logistic regression with rare event data and the objective of this paper is to evaluate capability of this semiparametric model in reducing bias. This

approach is motivated by known fact that non and semiparamtric models have good capability to overcome bias problem (Wassernan, 2004; Suliadi, et al, 2013; Suliadi, 2014; Suliadi and Kudus, 2015).

LOGISTIC REGRESSION

Suppose there are n subject and y_i be a binary response for the *i-th* subject where the $y_i \in \{0,1\}$, for i = 1, 2, ..., n. Response $y_i = 1$ if the *i-th* subject has a specific characteristics (success event) and $y_i = 0$ if the *i-th* subject has no that characteristics (fail event). For each subject is also measured k-dimensional vector of covariates x_i that affects the response. The response y_i has follow Bernoulli distribution with parameter μ_i and $E(y_i) = P(y_i=1) = \mu_i$. The probability mass function (pmf) of Y_i is $f(y_i \mid \mu_i) = \mu^{y_i}(1-\mu_i)^{1-y_i}$. Suppose the response y_i is related the covariates through a link function

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}.$$

Logistic regression takes canonical parameter $g(\mu) = \eta = \log(\mu/[1-\mu])$ as the link function. It will take

$$\mu_i = P(Y_i = 1) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}{1+\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}.$$

Let $y = (y_1, y_2, \ldots, y_n)^T; \mu = (\mu_1, \mu_2, \ldots, \mu_n)^T; \eta = (\eta_1, \eta_2, \ldots, \eta_n)^T; X = (x_1, x_2, \ldots, x_n)^T$. The ML estima-tor for $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is the solution of $U(\beta) = X^T W_1 (y - \pi) = 0$. Where W1=Diag{[1/var(yi)]} x $[\partial\mu_i/\partial\eta_i]$. For canonical link function it has $\partial\mu_i/\partial\eta_i = \mathrm{Var}(y_i)$. The iterative procedure by using Fisher Scoring algorithm for β is given by

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{m} + I^{-1}U,$$

where

$$I = E\left[\frac{\partial l}{\partial \beta}\frac{\partial l}{\partial \beta}\right] = X^T E_1 E(y-\pi)(y-\pi)^T) W_1 X = X^T W X \text{ and } W = W_1 \mathrm{Diag}\{\mathrm{Var}(y_i)\} W_1 = \mathrm{Diag}\left\{\frac{1}{\mathrm{Var}(y_i)}\left(\frac{\partial \pi_1}{\partial \eta_i}\right)^2\right\}.$$

SEMIPARAMETRIC APPROACH

P-Spline Regression for Continuous

In this paper we use P-spline since it is appropriate for large data and computationally simple. For this P-Spline we refer to Ruppert, et al. (2003), Green & Silverman (1994), dan Wu & Zhang (2006).

Suppose for the n pairs of observation (x_i, y_i) have relation $y_i = f(x_i) + e_i$. The idea of P-Spline is the Taylor's expansion. The function f can be approximated by polynomial of order k. To increase the flexibility, P-Spline divides the range of covariate x in several ranges/intervals and the function f modeled in this each interval. Suppose the minimum and maximum of x are a and b respectively. Hence we have interval $a<t_1< t_2< \ldots< t_K<b$. The values of $t_1, t_2, \ldots, t_K$ are called as knot. P-Spline is constructed by using *truncated power basis* of degree k with K knot $t_1< t_2< \ldots< t_K$:

$$1, x, \ldots, x^k, (x-t_1)^k_+, (x-t_2)^k_+, \ldots, (x-t_K)^k_+$$

and the function f is approximated by

$$g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \beta_{k+1}(x-t_1)^k_+ + \beta_{k+2}(x-t_2)^k_+ + \cdots + \beta_{k+K}(x-t_K)^k_+$$

$$= \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{K} \beta_{k+r}(x-t_r)^k_+$$

with $w_+ = \mathrm{maximum}(0,w)$.

Let

$$\beta = (\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}, \ldots, \beta_{k+K})^T \;;\; x_i = (1, x_i, \ldots, x_i^k, (x_i - t_1)^k_+, (x_i - t_2)^k_+, \ldots, (x_i - t_K)^k_+)^T$$

$$X = (x_1, x_2, \ldots, x_n)^T;\; \eta = (f(x_1), f(x_2), \ldots, f(x_n))^T;\; y = (y_1, y_2, \ldots, y_n)$$

hence we have $g(x_i) = x^T \beta$. The objective function of P-Spline is to minimize

$$\sum_{i=1}^{n} y_i - x(t_i)^T \beta + \lambda \beta^T G \beta = (y - X\beta)^T (y - X\beta) + \lambda \beta^T G \beta$$

where

$$G = \begin{bmatrix} 0_{(k+1)\times(k+1)} & 0_{(k+1)\times K} \\ 0_{K\times(k+1)} & I_K \end{bmatrix}.$$

This is called as penalized least square. The solution for β is

$$\hat{\beta} = (X^T X + \lambda G)^{-1} X^T y.$$

Semiparametric Logistic Regression Based on P-Spline

Suppose we have n samples and each observed binary respon $y_i \in \{0,1\}$ and covariates $v_{1i}, v_{2i}, \ldots, v_{pi}$ dan r_i. The varriables $v_{1i}, v_{2i}, \ldots, v_{pi}$ affect the y_i parametrically whilstvariabel r_i affects y_i nonparametrically through P-Spline of degree k with knot $t_1, t_2, \ldots, t_K$. The relation among variables is on form

$$\begin{aligned} \eta_i &= \beta_0 + \underbrace{\delta_1 v_{1i} + \cdots + \delta_p v_{pi}}_{\text{Parametric part}} \\ &\quad \underbrace{+ \alpha_1 r_i + \alpha_2 r_i^2 + \cdots + \alpha_k r_i^k + u_1 (r_i - t_1)_+^k + u_2 (r_i - t_2)_+^k + \cdots + u_K (r_i - t_K)_+^k}_{\text{Nonparametric part: P-Spline of degree } k} \\ &= x_i^T \beta + z_i^T u \end{aligned} \tag{1}$$

where

$$x_i = (1 \; v_{1i}, \ldots, v_{1i}, \; r_i, \; r_i^2, \ldots, r_i^k)^T; \; z_i = [(r_i - t_1)_+^k, (r_i - t_2)_+^k, \ldots, (r_i - t_K)_+^k)^T;$$

$$\beta = (\beta_0, \delta_1, \ldots, \delta_p, \alpha_1, \ldots, \alpha_k)^T; u = (u_1, u_2, \ldots, u_K)^T.$$

The relation between the response and the covariates is through logit link function

$$logit(\mu_i) = \frac{\mu_i}{1 - \mu_i} = \eta_i.$$

Suppose $\theta = (\beta^T, u^T)^T$ and $X_i = (x_i^T, z_i^T)^T$, thus model (1) can be written as

$$\eta_i = X_i^T \theta.$$

For non-continuous data, we use the same idea as penalized least square, known as penalized maximum likelihood (PML) with the function of PML is $\Pi = l(\theta) - (1/2)\lambda \theta^T G \theta$, where $l(\theta)$ is the likelihood function for θ and

$$G = \begin{pmatrix} 0 & 0 \\ 0 & I_K \end{pmatrix}.$$

Solution for θ is obtained by maximizing the penalized maximum likelihood function Π, and obtained as the solution of $U = \partial\Pi/\partial\theta = 0$. The form of U is

$$\begin{aligned} U &= \frac{\partial L(\theta)}{\partial \theta} - (1/2)\lambda \frac{\partial}{\partial \theta}(\theta^T G \theta) \\ &= X^T W_1 (y - \mu) - \lambda G \theta. \end{aligned}$$

Solution for θ has no close form, hence we may iterate using Fisher's scoring algorithm with form

$$\hat{\theta}^{r+1} = \hat{\theta}^r + I^{-1} U.$$

In smoothing spline, all distinct values of nonparametric covariate are taken as knot. Meanwhile P-Spline allows many knots, but not all distinct values. To overcome the over fitting, P-Spline introducing smoothing parameter λ as the penalty of smoothness. Thus in P-Spline, there are three values that should be determined before running iteration in Fisher's scoring algorithm, that are smoothing parameter λ, degree of polynomial k, and the number of knot K. Ruppert (2002) gave a guidance to determine the number of knot K as min(n/4, 40). The position of knot may be determined using the suggestion of Ruppert (2002) and Yu & Ruppert (2002). They suggested using

equally-spaced sample quantile as the knot. Suppose K is the number of knot, then the *i-th* knot is defined as follows. Let $l = i \times (n+1)/(K+1)$. If l is integer, then the *i-th* knot is $t_i = r_{[l]}$, otherwise $l = \lfloor i \times (n+1)/(K+1) \rfloor$ and

$$t_i = \left(\frac{r_{[l]} + r_{[l+1]}}{2} \right), \quad i = 1, 2, \ldots, K.$$

The smoothing parameter can be chosen using CV, GCV, AIC or BIC method. But these need complicated programming and time consuming on iteration. We may use another approach by using the connection of P-Spline and linear mixed model (Ruppert, et al., 2003; Wu & Zhang, 2006), specifically approaching P-Spline logistic regression with generalized linear mixed model for logistic regression (GLMM). Thus we do not need to write program just use software that provide GLMM analysis, for example macro "glimmix" in SAS.

PERFOMANCE OF SEMIPARAMETRIC MODEL IN REDUCING BIAS

Simulation Study

We run simulation to evaluate the performance of semiparametric model to reduce bias in case of rare event problem of logistic regression. The response y generated from Bernoulli(μ), logit(μ) = $\eta = 1 + X_1 - X2 + \sin(4\pi Z)$. The covariate X1 was generated from Normal(0,1) and X2 generated from Bernoulli(0.5), whilst Z generated from Uniform(0, π). We generated 10,000,000 observations as population representative. We separate observation with Y=0 and Y=1 as G0 and G1 group respectively. We set three sample sizes n = 200, 500, and 1,000. We take Y=0 as the majority class/category and Y=1 as the minority class. Level or rareness is presented by the proportion of the number of observation with Y=1 to total observation or sample size, p = #(Y=1)/n. We take four levels of rareness p = 5%, 10%, 20% and 50%. For combination of sample size *n* and the level of rareness *p*, we choose randomly n_0 = (1-p) × n observations from G0 and n_1 = × n observations from G1 and united as the data set. This data set is modelled parametrically: $\eta = \beta_0 + \beta_1 X_1 + \beta_1 X2 + \beta_1 Z$ and semiparametrically: $\eta = \beta_0 + \beta_1 X_1 + \beta_1 X2\, f(Z)$. For each scenario we run 400 replications.

TABLE 1. Bias of the estimate of parametric component for parametric and semiparametric models

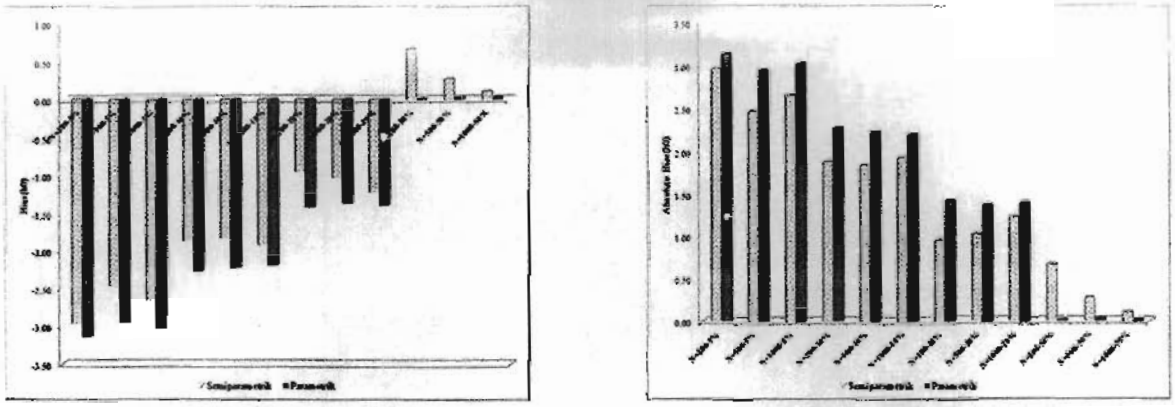
Sample size	Percentage Y=1	β_0		β_1		β_2	
		Semi	Param	Semi	Param	Semi	Param
200	5	-2.963	-3.138	0.039	-0.002	0.028	0.063
500	5	-2.467	-2.949	-0.057	-0.087	-0.012	0.015
1000	5	-2.652	-3.027	-0.028	-0.065	0.003	0.049
200	10	-1.874	-2.271	0.013	-0.031	0.024	0.067
500	10	-1.828	-2.224	-0.011	-0.058	0.015	0.049
1000	10	-1.915	-2.191	0.013	-0.037	0.015	0.061
200	20	-0.948	-1.417	0.007	-0.035	0.044	0.088
500	20	-1.026	-1.370	0.005	-0.050	0.022	0.072
1000	20	-1.227	-1.392	-0.020	-0.085	0.034	0.104
200	50	0.663	0.009	0.008	-0.040	0.022	0.068
500	50	0.271	0.024	-0.003	-0.062	-0.027	0.039
1000	50	0.103	0.023	-0.009	-0.081	-0.004	0.064

Note: ± 30% in estimation of the semiparametric models do not converge.

Parametric Component: Bias of The Estimates

Table 1 gives bias of the estimates for both parametric and semiparametrci model. It is seen that these models result unbiased estimate of the coefficient of covariates, i.e b_1 and b_2. However both models give biased estimate of the coefficient of intercept, b_0 when there is rare event problem, i.e percentage of Y=1 < 50%. The bias is negative that means the estimate of P(Y=1) will be under estimate. This result is consistent with references.

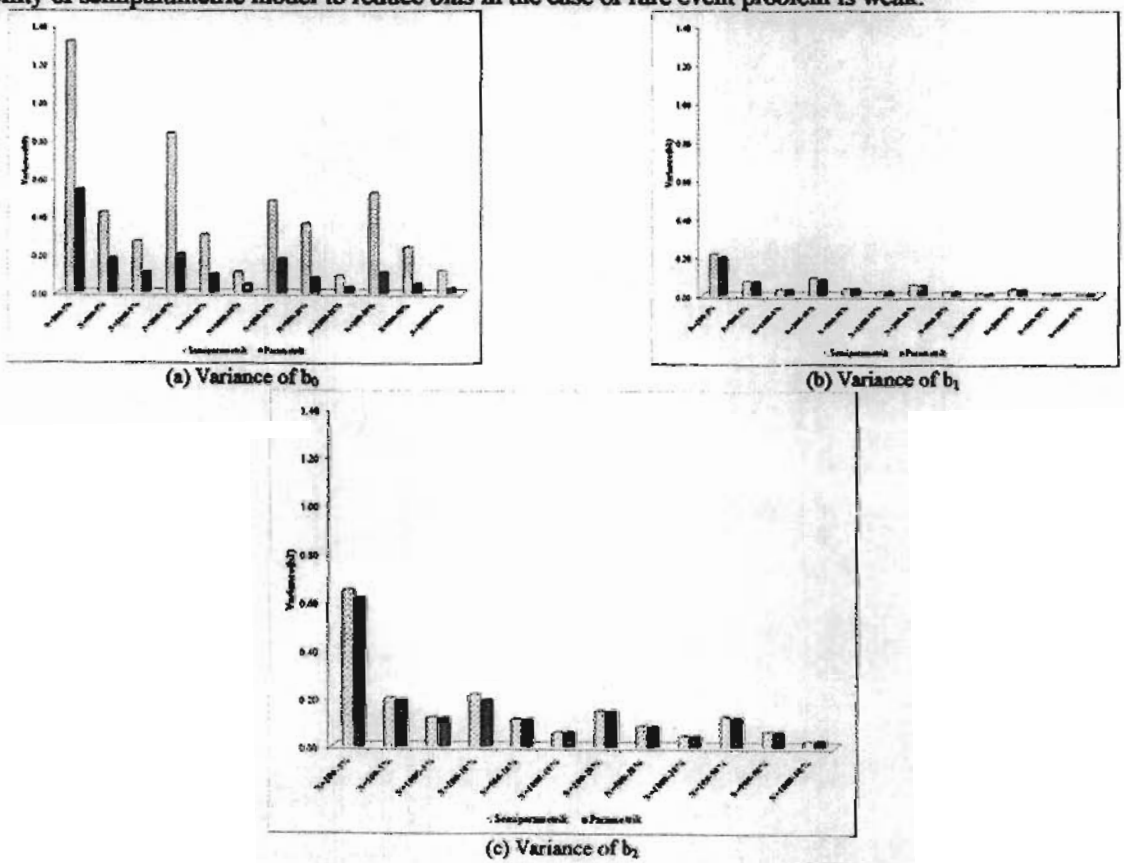
Figure 1 shows bias and absolute bias of b_0. From this figure it can be seen that the most influence to bias is level of rareness. The bias increases if the level of rareness increases (proportion of minority class decreases). The effect of sample size to bias estimates is relatively small than the effect of level of rareness. These characteristics are valid for both parametric and semiparameric models.



(a) (b)

FIGURE 1. Bias (a) and absolute bias (b) of b_0

From Fig. 2 we can see that bias of b_0 obtained from semiparametric model is smaller than that obtained from parametric model. This means that model semiparametric can reduce the bias of the coefficient of intercept when there rare event problem in the data. However, the different of bias b_0 from both models is small that indicates the capability of semiparametric model to reduce bias in the case of rare event problem is weak.



(a) Variance of b_0 (b) Variance of b_1

(c) Variance of b_2

FIGURE 2. Variance of the coefficients regression estimate

Parametric Component: Variance of The Estimates

Figure 2 gives variance of the estimate for both parametric and semiparametric model. As expected, the pattern

of variance is influenced by the sample size and the level of rareness. If the sample size increases then the variance decreases, and if the level of rareness increases then the variance also increases. For the coefficients of covariate, b_1 and b_2, variances from model parametric and semiarametric are comparable. But variances of b_0 from both models have different pattern. Variance of b0 from semiparametric model is larger than from the parametric one. It can be understood, since the bias of b_0 from semiparametric model is smaller than is from parametric one. As well known that mean square error equals to square of bias plus variance, thus the smaller bias is compensated to larger variance.

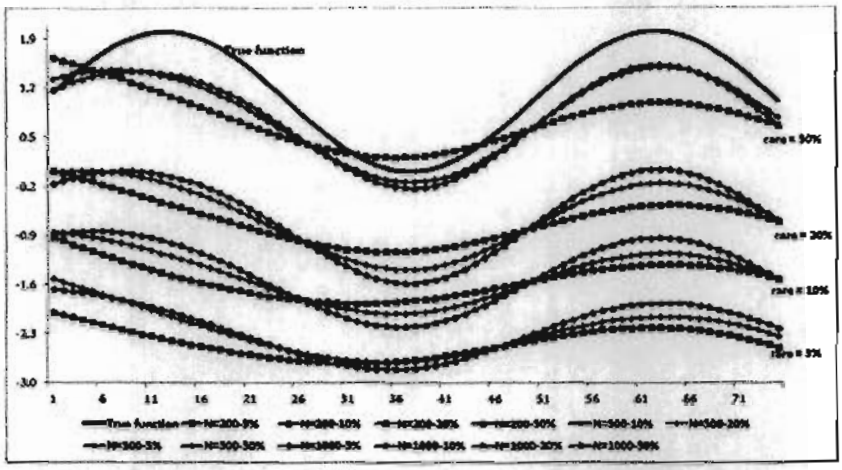


FIGURE 3. The average of 75 points estimates of the nonparametric function *f*

Semiparametric Component: Bias and Variance of the Function Estimate

In order to evaluate bias and variance, for each data set and scenario we took 75 points and then estimated the function *f* at those points. The average of the function estimate for the semiparametric model is depicted in Fig. 3. This figure shows that if there is rare event problem the semiparametric model still gives biased estimate of the nonparametric function *f*. The bias increases as the level of rareness increases. The bias is large enough, implies the weakness of this approach in reducing bias.

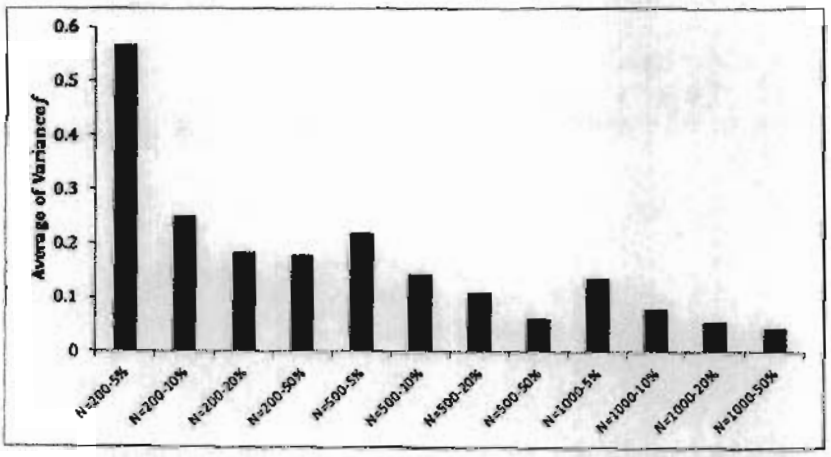


FIGURE 4. The average variance of 75 points estimates of the nonparametric function *f*

The average of variance for of all 75 points estimate for each scenario is shown in Fig. 4. This figure shows that the variance of the curve estimate is influenced by the sample size and the level of rareness. Variance decreases as the sample increases and if the level of rareness increases then the variance is also increases.

DISCUSSION

This study gives new information on the capability of non and semiparametric models to reduce bias. It is well known that non and semiparametric models have good capability in reducing bias. However this results show that semiparametric model is almost fail to reduce bias.

If the parametric model is incorrect then the estimate will be biased. This bias can be reduced (almost removed) if we use nonparametric model or semiparametric model, since the incorrectness of the model will be absorbed by the nonparametric function, i.e f. Meanwhile, the problem of rare event is not in the model, but the structure of binary response, i.e the unbalanced of proportion Y=0 and Y=1. This reason explains why semiparametric model fails to reduce bias of the regression estimate, specifically coefficient of intercept, when problem of rare event exists.

Computationally, approaching GLM P-Spline with GLMM should be careful, since it may not give a convergent result. It does not mean that this approach cannot be used, but it may give no result.

ACKNOWLEDGMENTS

This research is supported by LPPM of Bandung Islamic Univeristy in 2017. We also like to thank to reviewer that accept this article to be presented in ICOMSE 2017.

REFERENCES

1. P.J Green and B.W. Silverman, *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach* (Chapman & Hall/CRC, New York, USA, 1994).
2. G. King and L. Zheng, Political Analysis, 9,137-163 (2001).
3. G. King and L. Zheng, World Politics, 54 (4), 623-658. (2001b).
4. P. McCullagh and J. A. Nelder, (Generalized Linear Models. Chapman & Hall. London, UK, 1989).
5. Z. Qiu, H. Li, H. Su, G. Ou, and T. Wang, "Logistic Regression Bias Correction for Large Scale Data with Rare Events" In *Advanced Data Mining and Applications*. (Springer-Verlag, Berlin, Germany,2013). pp 133-143.
6. J. Quigley, T Bedford, and L. Walls, Reliability Engineering & System Safety, 92, 619-627 (2007).
7. D. Ruppert, M. P. Wand, and R. J. Carroll, (Semiparametric Regression. Cambridge University Press, Cambridge, UK, 2003).
8. D. Ruppert, Journal of Computational and Graphical Statistics, 11, 735-757 (2002).
9. Suliadi, N. A. Ibrahim and I. Daud, Communication in Statistics-Simulations and Computations, 42, 138-152 (2013).
10. Suliadi and A. Kudus, International Journal of Applied Mathematics and Statistics, 53, 44-54 (2015).
11. Suliadi, A. Kudus, T. S. Yanti, A. D. Permadi, and A. Rohayati, "Metode Mereduksi Bias Pada Regresi Logistik Ketika Ada Masalah Data Rare Event," Technical report, Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Islam Bandung. (2016).
12. Suliadi, International Journal of Applied Mathematics and Statistics, 52, 141-151 (2014)..
13. L. Wasserman, *All of Statistics. A Concise Course in Statistical Inference*, (Springer, New York, USA, 2004).
14. G. M. Weiss and H. Hirsh, "Learning to Predict Extremely Rare Events," In *AAAI Workshop on Learning from Imbalance Data Sets*, (2000), pp 64-68.
15. H. Wu and J. T. Zhang, *Nonparametric Regression Methods for Longitudinal Data Analysis* (John Wiley & Sons, New Jersey, USA, 2006).
16. Y. Yu and D. Ruppert, Journal of the American Statistical Association, 97, 1042-1054 (2002).

Lampiran 2. Log Book Penelitian.

Tanggal	Kegiatan	Keterangan
17/1/2017	Browsing & Review:	Multicategorical spline model for item response theory
		Suplement#bootstrapping for penalized spline regression
		Bootstrap confidence bands in nonparametric regression
20/1/2017	Browsing & Review:	Proc glimmix_ from penalized splines to mixed models
		Semiparametric regression book rupert et al##sasmacros
		Understanding splines in the effect statement
24/1/2017	Browsing & Review:	Smoothing with mixed model software
		Lme4- mixed-effects modeling with r
		Spline regression in the presence of categorical predictors
27/1/2017	Browsing & Review:	On bootstrap confidence intervalas in nonparametric regression
		Semiparametric book ruppert etal
		Growing up fast- sas 9.2 enhancements to the glimmix procedure
31/1/2017	Browsing & Review:	A note on penalized spline smoothing with correlated errors
		Smoothing parameter and model selection for general smooth models
		Functional generalized additive models
2/2/2017	Browsing & Review:	Theoretical and practical aspects ofpenalized spline smoothing##dissertation
		Nonparametric regression and the parametric bootstrap for local dependence assessment
		Thesis##penalized spline regression and its applications
6/2/2017	Browsing & Review:	Bootstrap confdence intervals in nonparametric regression with built-in bias correction
		On semiparametric regression with o'sullivan penalized splines
8/2/2017	Browsing & Review:	Penalized regression, mixed effects models and appropriate modelling
		Selecting the number of knots for penalized splines
13/2/2017	Browsing & Review:	Generalized_linear_mixed_models_curren_draft
		B(asic)-spline basics
		Splines, knots, and penalties
15/2/2017	Browsing & Review:	Simultaneous bootstrap confidence bands in nonparametric regression
		Semiparametric multinomial logit models for analysing consumer choice behaviour
		Simultaneous confidence bands for penalized spline estimators
20/2/2017	Browsing & Review:	Tvem (time-varying effect modeling) sas macro users guide
		A primer on regression splines
		Weighted tvem sas macro users' guide version 2.6
22/2/2017	Browsing & Review:	Introduction to glimmix #slide
		Fast adaptive penalized splines
		Flexible smoothing with b-splines and penalties
2/3/2017	Browsing & Review:	A crash course on p-splines_course_handout
		Bootstrapping in nonparametric regression# local adaptive smoothing and confidence bands
		A rasch model for partial credit scoring
6/3/2017	Browsing & Review:	Semiparametric methods for the generalized linear model##chen_jinsong_d_2010
		Bayesian smoothing and regression splines for measurement error problems
		Introduction to b-spline curves

9/3/2017	Browsing & Review:	Confidence bands in nonparametric regression
		The vgam package for categorical dataanalysis
		Varying index coefficient models
13/3/2017	Browsing & Review:	Bivariate penalized splines for regression
		A simple bootstrap method for constructing nonparametric confidence bands for functions
		Bootstrap simultaneous error bars for nonparametric regression
		Simultaneous bootstrap confidence bands in regression
		Generalized linear mixed models
16/3/2017	Browsing & Review:	A very short note on b-splines
		Bootstrap selection of bandwidth and confidence bands for nonparametric regression
		Selecting the number of knots for penalized splines
		Bootstrapping for penalized spline regression
20/3/2017	Browsing & Review:	Semiparametric regression ruppert etal errata
		Monotone spline transformations for dimension reduction
		Variable selection using p-splines
		Spline and penalized regression
		Dissertation#nonparametric regression models and bootstrap inference
23/3/2017	Browsing & Review:	A short introduction to splines in least squares regression analysis
		On knot placement for penalized spline regression
		Nonparametric and semiparametric methodsin r
		Smoothing terms in gam models
		Convergence rates for uniform confidence intervals based on local polynomial regression estimators
27/3/2017	Browsing & Review:	Bias-reduced and separation-proof conditional logistic regression with small sample
		A statistical method for studying correlated rare events and their risk factors.
		Experimental perspectives on learning from imbalanced data
		Credit Scoring Using Global and Local Statistical Models
3/4/2017	Browsing & Review:	Analyzing Rare Events with Logistic Regression
		Firth logistic regression SAS & R.docx
		Estimation of Rare Events from Biased Sampling
5/4/2017	Browsing & Review:	Refined Estimation of a Light Tail
		A Solution to Separation in Binary Response Models
		Separation-Resistant and Bias-Reduced Logistic
7/4/2017	Browsing & Review:	Bias Adjustment in Logistic Regression Models
		Classification - the Ubiquitous Challenge ## Page 436
		Application of Support Vector Machines in a Life Assurance Environment
10/4/2017	Browsing & Review:	Estimation of Rare Events from Biasod Sampling
		Credit Scoring Using Semiparametric Methods
		FIRTH#Bias Reduction of Maximum Likelihood Estimates
12/4/2017	Browsing & Review:	Fully Parametric and Semi-Parametric Regression Models for Common Events withCovariate Me
		Michael_R._Kosorok_Introduction_to_Empirical_Processes_and_Semiparametric_Inference
		A solution to the problem of separation in logistic regression

12/4/2017	Browsing & Review:	Model-Based Classification of Clustered Binary Data with Non-ignorable Missing Values
		Robust weighted kernel logistic regression in imbalanced and rare events data.pd
		Probabilistic Kernel Regression Models
14/4/2017	Browsing & Review:	THESIS ## A comparison of different methods for modelling rare events data
		Predicting Bankruptcy with Semi-Parametric Single-Index Model
		Penalized Regression, Mixed Effects Models and Appropriate Modelling
17/4/2017	Browsing & Review:	Smoothing with SAS Proc Mixed
		Informative Patterns for Credit Scoring
		Predicting Bankruptcy with Semi Parametric Single Index Model
19/4/2017	Browsing & Review:	How Can Non-invariant Statistics Work in Our Benefit in the Semi-parametric Estimation of
		Kernel methods and the exponential family##Journal Version
		Simon_Wood_Generalized_Additive_Models_An_Introduction_with_R
20-31/ 4/2017	Programming	Fail/slow to Converge
1-20/ 5/2017	Programming	Fail/slow to Converge
20-30/ 5/2017	Review	Review on Spline & GLMM
4-15/6/2017	Browsing internet	Mencari program GLMM
16-30/6/2017	Programming	Modifikasi macro SAS "glimmix"
1-15/7/2017	Programming	Membuat macro SAS sebagai komplemen macro glimmix
16-30/7/2017	Simulasi	Melakukan simulasi
1-7/8/2017	Analisis data	Melakukan analisis data hasil simulasi
8-25/8/2017	Pelaporan	Penulisan laporan akhir dan artikel ilmiah

GENERATING DATA

```
LIBNAME IN 'D:\TEMP\TEST';
PROC IML;

NSAMPLE=10000000;
DATA=J(NSAMPLE,4,.);
PI=3.14159265358979;

DO N=1 TO NSAMPLE;
DATA[N,1]=RAND('NORMAL',0,1);
DATA[N,2]=RAND('BERNOULLI',0.5);
DATA[N,3]=PI*RAND('UNIFORM');
ETA=1+SIN(DATA[N,3]*4)+DATA[N,1]-DATA[N,2];
MU=EXP(ETA)/(1+EXP(ETA));
DATA[N,4]=RAND('BERNOULLI',MU);
END;
NAME={'X1' 'X2' 'NON' 'Y'};
CREATE IN.DATASEMPAR FROM DATA[COLNAME=NAME];
APPEND FROM DATA;SAVE;
CLOSE IN.DATASEMPAR;
QUIT;
PROC FREQ DATA=IN.DATASEMPAR;
TABLE Y;
PROC GENMOD DATA=IN.DATASEMPAR DESC;
MODEL Y=X1 X2 NON/DIST=BIN LINK=LOGIT;
   RUN;
PROC SORT DATA=IN.DATASEMPAR;
BY Y;RUN;
```

```
MACRO RUNNING GLIMMIX MACRO
*************************************************************************

OPTIONS PAGESIZE=32000 LINESIZE=64 NODATE PAGENO=1;

LIBNAME IN 'D:\TEMP\TEST';
/**CATATAN:
N(Y = 0) = 4079594;
N(Y = 1) = 5920406;
TOTAL DATASEMPAR=10.000.000
**/

/**/
%MACRO RUNGLIMMIX(REPLICATION,SAMPLEN,PERCENT);
%DO ITERASI=1 %TO &REPLICATION;

PROC IML; /** PREPARING DATA **/
/****  GENERATING DATA
NSAMPLE=1000;
DATA=J(NSAMPLE,4,.);
PI=3.14159265358979;

DO N=1 TO NSAMPLE;
DATA[N,1]=RAND('NORMAL',0,1);
DATA[N,2]=RAND('BERNOULLI',0.5);
DATA[N,3]=PI*RAND('UNIFORM');
ETA=1+SIN(DATA[N,3]*4)+DATA[N,1]-DATA[N,2];
MU=EXP(ETA)/(1+EXP(ETA));
DATA[N,4]=RAND('BERNOULLI',MU);
```

```
END;
NAME={'X1' 'X2' 'NON' 'Y'};
CREATE IN.DATASEMPAR FROM DATA[COLNAME=NAME];
APPEND FROM DATA;SAVE;
CLOSE IN.DATASEMPAR;
***/
/****  GENERATING DATA  ***/


/*** MACRO GLIMMIX  **/
/*** MACRO GLIMMIX  **/
/*** MACRO GLIMMIX  **/


PI=3.14159265358979;
K=35;
POSITION_Y0=1:4079594;
POSITION_Y1=4079595:10000000;
N=&SAMPLEN;
PERCENTAGE=&PERCENT/100;
N0=FLOOR((1-PERCENTAGE)*N);
N1=N-N0;

Y0=J(1,N0,.);
Y1=J(1,N1,.);

DO I=1 TO N0;
Y0[I]=CEIL(RAND('UNIFORM')*4079594);
END;

DO I=1 TO N1;
Y1[I]=4079594+CEIL(RAND('UNIFORM')*5920406);
END;

POSITION=Y0||Y1;

USE IN.DATASEMPAR;
READ POINT POSITION  VAR{X1 X2 NON Y};
CLOSE IN.DATASEMPAR;

NON2=NON##2;
UNIQUENON=UNIQUE(NON);
N_UNIQUE=NCOL(UNIQUENON)*NROW(UNIQUENON);
KNOT=J(1,K,.);
Z2=J(NROW(NON),K,.);
DO I=1 TO K;
     QUANTILE=((I+1)/(K+2))*(N_UNIQUE+1);
     KNOT[I]=(UNIQUENON[FLOOR(QUANTILE)]+UNIQUENON[CEIL(QUANTILE)])/2;
     Z2[,I]=((NON-KNOT[I])#((NON-KNOT[I])>0))##2;
END;
/**NAME_B=ROWCATC((J(1,NX,'B'))`||(CHAR(1:NX))`);**/
NAMEZ  =ROWCATC((J(1,K,'Z2_'))`||(CHAR(1:K))`);
NAME={'Y' 'X1' 'X2' 'NON' 'NON2'}||T(NAMEZ);
NAME=T(NAME);
CREATE IN.DATANAMEVAR FROM NAME[COLNAME={DVAR}];
APPEND FROM NAME;SAVE;CLOSE IN.DATANAMEVAR;


NAME=T(NAME);
DATA=Y||X1||X2||NON||NON2||Z2;

CREATE IN.COBA1 FROM DATA[COLNAME=NAME];
APPEND FROM DATA;SAVE;CLOSE IN.COBA1;
KNOT=T(KNOT);
```

```
CREATE IN.DKNOT FROM KNOT[COLNAME={'KNOT'}];
APPEND FROM KNOT;SAVE;CLOSE IN.DKNOT;

/*
DX=X1||X2;
DZ=J(NROW(X1),1,1)||NON||NON2||Z2;
CREATE IN.DATAX FROM DX[COLNAME={'X1' 'X2'}];
APPEND FROM DX;SAVE;CLOSE IN.DATAX;

NAME=T({'X1' 'X2'});
CREATE IN.NAMEX FROM NAME[COLNAME={'VARX'}];
APPEND FROM NAME;SAVE;CLOSE IN.NAMEX;


NAME={'CONSTANT' 'NON' 'NON2'}||NAMEZ`;
CREATE IN.DATAZ FROM DZ[COLNAME=NAME];
APPEND FROM DZ;SAVE;CLOSE IN.DATAZ;

NAME=T(NAME);
CREATE IN.NAMEZ FROM NAME[COLNAME={'VARZ'}];
APPEND FROM NAME;SAVE;CLOSE IN.NAMEZ;

*/

/*CREATE IN.DATAW FROM DTEMP[COLNAME=COLN];
APPEND FROM DTEMP; SAVE; CLOSE IN.DATAW;
*/
QUIT;
%INCLUDE 'D:\TEMP\TEST\GLIMMIX.SAS';
%GLIMMIX(DATA=IN.COBA1,PROCOPT=METHOD=REML,
        STMTS=%STR(
                    MODEL Y=X1 X2 NON NON2/ SOLUTION;
                    RANDOM
Z2_1 Z2_2 Z2_3 Z2_4 Z2_5 Z2_6 Z2_7 Z2_8 Z2_9 Z2_10
Z2_11 Z2_12 Z2_13 Z2_14 Z2_15 Z2_16 Z2_17 Z2_18 Z2_19 Z2_20
Z2_21 Z2_22 Z2_23 Z2_24 Z2_25 Z2_26 Z2_27 Z2_28 Z2_29 Z2_30
Z2_31 Z2_32 Z2_33 Z2_34 Z2_35

                     / TYPE=TOEP(1) S;
                         ODS OUTPUT SOLUTIONR=IN.BETAACAK;
                         ODS OUTPUT SOLUTIONF=IN.BETAFIXED;
                         ),
                    ERROR=BINOMIAL,
                    LINK=LOGIT,
                    OUT=FITTED);
DM 'LOG;CLEAR;OUTPUT;CLEAR';

PROC IML;
USE IN.POINT1000RI;
READ POINT 1 INTO POINT;
CLOSE IN.POINT1000RI;

POINT=T(POINT);

USE IN.DKNOT;
READ ALL INTO KNOT;
CLOSE IN.DKNOT;

Z=J(NROW(POINT),35,.);
DO I=1 TO 35;
     Z[,I]=((POINT-KNOT[I])##2)#(POINT>KNOT[I]);
END;
Z100=J(NROW(POINT),1,1)||POINT||POINT##2||Z;
```

```
USE IN.BETAFIXED;
READ ALL VAR{ESTIMATE} INTO ESTIMATE_F;
CLOSE IN.BETAFIXED;

USE IN.BETAACAK;
READ ALL VAR{ESTIMATE} INTO ESTIMATE_R;
CLOSE IN.BETAACAK;

BETAR=ESTIMATE_F[{1 4 5}]//ESTIMATE_R;
NONPAR=Z100*BETAR;


NSAMPLE=&SAMPLEN;
PERCENTAGE=&PERCENT;
ESTIMATE_F=T(ESTIMATE_F)||NSAMPLE||PERCENTAGE;
EDIT IN.BETAF_ALL;
APPEND FROM ESTIMATE_F;
CLOSE IN.BETAF_ALL;

NONPAR=T(NONPAR)||NSAMPLE||PERCENTAGE;


EDIT IN.POINT100;
APPEND FROM NONPAR;
CLOSE IN.POINT100;

QUIT;
%END;
%MEND RUNGLIMMIX;
*************************************************************
RUNNING PROC LOGISTIC
*************************************************************

OPTIONS PAGESIZE=32000 LINESIZE=64 NODATE PAGENO=1;


LIBNAME IN 'D:\TEMP\TEST';
/**CATATAN:
N(Y = 0) = 4079594;
N(Y = 1) = 5920406;
TOTAL DATASEMPAR=10.000.000
**/

/**/
%MACRO RUNLOGISTIC(REPLICATION,SAMPLEN,PERCENT);
%DO ITERASI=1 %TO &REPLICATION;

PROC IML; /** PREPARING DATA **/
/****  GENERATING DATA
NSAMPLE=1000;
DATA=J(NSAMPLE,4,.);
PI=3.14159265358979;

DO N=1 TO NSAMPLE;
DATA[N,1]=RAND('NORMAL',0,1);
DATA[N,2]=RAND('BERNOULLI',0.5);
DATA[N,3]=PI*RAND('UNIFORM');
ETA=1+SIN(DATA[N,3]*4)+DATA[N,1]-DATA[N,2];
MU=EXP(ETA)/(1+EXP(ETA));
DATA[N,4]=RAND('BERNOULLI',MU);
END;
NAME={'X1' 'X2' 'NON' 'Y'};
CREATE IN.DATASEMPAR FROM DATA[COLNAME=NAME];
```

```
APPEND FROM DATA;SAVE;
CLOSE IN.DATASEMPAR;
***/
/****  GENERATING DATA  ***/


/*** MACRO GLIMMIX  **/
/*** MACRO GLIMMIX  **/
/*** MACRO GLIMMIX  **/


PI=3.14159265358979;
K=35;
POSITION_Y0=1:4079594;
POSITION_Y1=4079595:10000000;
N=&SAMPLEN;
PERCENTAGE=&PERCENT/100;
N0=FLOOR((1-PERCENTAGE)*N);
N1=N-N0;

Y0=J(1,N0,.);
Y1=J(1,N1,.);

DO I=1 TO N0;
Y0[I]=CEIL(RAND('UNIFORM')*4079594);
END;

DO I=1 TO N1;
Y1[I]=4079594+CEIL(RAND('UNIFORM')*5920406);
END;

POSITION=Y0||Y1;

USE IN.DATASEMPAR;
READ POINT POSITION  VAR{X1 X2 NON Y};
CLOSE IN.DATASEMPAR;

NON2=NON##2;
UNIQUENON=UNIQUE(NON);
N_UNIQUE=NCOL(UNIQUENON)*NROW(UNIQUENON);
KNOT=J(1,K,.);
Z2=J(NROW(NON),K,.);
DO I=1 TO K;
     QUANTILE=((I+1)/(K+2))*(N_UNIQUE+1);
     KNOT[I]=(UNIQUENON[FLOOR(QUANTILE)]+UNIQUENON[CEIL(QUANTILE)])/2;
     Z2[,I]=((NON-KNOT[I])#((NON-KNOT[I])>0))##2;
END;
/**NAME_B=ROWCATC((J(1,NX,'B'))`||(CHAR(1:NX))`);**/
NAMEZ  =ROWCATC((J(1,K,'Z2_'))`||(CHAR(1:K))`);
NAME={'Y' 'X1' 'X2' 'NON' 'NON2'}||T(NAMEZ);
NAME=T(NAME);
CREATE IN.DATANAMEVAR FROM NAME[COLNAME={DVAR}];
APPEND FROM NAME;SAVE;CLOSE IN.DATANAMEVAR;


NAME=T(NAME);
DATA=Y||X1||X2||NON||NON2||Z2;

CREATE IN.COBA1 FROM DATA[COLNAME=NAME];
APPEND FROM DATA;SAVE;CLOSE IN.COBA1;
KNOT=T(KNOT);

CREATE IN.DKNOT FROM KNOT[COLNAME={'KNOT'}];
APPEND FROM KNOT;SAVE;CLOSE IN.DKNOT;
```

```
QUIT;

PROC GENMOD DATA=IN.COBA1 DESC;
MODEL Y = X1 X2 NON / DIST = BIN
LINK = LOGIT;
ODS OUTPUT PARAMETERESTIMATES=IN.BETASEMTOPARAM;
ODS OUTPUT CONVERGENCESTATUS=IN.STATUSSEMTOPARAM;

RUN;
/**/
DM 'LOG;CLEAR;OUTPUT;CLEAR';

PROC IML;
USE IN.BETASEMTOPARAM;
READ ALL VAR{ESTIMATE} INTO ESTIMATE_F;
CLOSE IN.BETASEMTOPARAM;

USE IN.STATUSSEMTOPARAM;
READ ALL VAR{STATUS};
CLOSE IN.STATUSSEMTOPARAM;

NSAMPLE=&SAMPLEN;
PERCENTAGE=&PERCENT;
ESTIMATE_F=T(ESTIMATE_F)||NSAMPLE||PERCENTAGE||STATUS;
EDIT IN.SEMTOPARAM;
APPEND FROM ESTIMATE_F;
CLOSE IN.SEMTOPARAM;

QUIT;
%END;
%MEND RUNLOGISTIC;

*****************************************************************
MACRO COMBINE
*****************************************************************

/**------------- CATATAN:
N(Y = 0) = 4079594;
N(Y = 1) = 5920406;
TOTAL DATASEMPAR=10.000.000
----------------------**/

%MACRO COMBINE(RSTART,REND,SAMPLEN,PERCENT);

%DO ITERASI=&RSTART %TO &REND;

proc iml; /** PREPARING DATA **/


pi=3.14159265358979;
K=35;  /* NUMBER OF KNOT */
POSITION_Y0=1:4079594;        /* DEPEND ON DATA*/
POSITION_Y1=4079595:10000000;/* DEPEND ON DATA*/
N=&SAMPLEN;
PERCENTAGE=&PERCENT/100;
N0=FLOOR((1-PERCENTAGE)*N);
N1=N-N0;

Y0=j(1,n0,.);
Y1=j(1,n1,.);

do i=1 to n0; /* DEPEND ON DATA*/
y0[i]=ceil(rand('uniform')*4079594);
```

```
end;

do i=1 to n1;/* DEPEND ON DATA*/
y1[i]=4079594+ceil(rand('uniform')*5920406);
end;

POSITION=Y0||Y1;

USE IN.DATASEMPAR;/* DEPEND ON DATA*/
READ POINT POSITION  VAR{X1 X2 NON Y};
CLOSE IN.DATASEMPAR;

NON2=NON##2;
UNIQUENON=UNIQUE(NON);
N_UNIQUE=NCOL(UNIQUENON)*NROW(UNIQUENON);
KNOT=J(1,K,.);
Z2=J(NROW(NON),K,.);

DO I=1 TO K;  /* CREATING BASIS */
     QUANTILE=((I+1)/(K+2))*(N_UNIQUE+1);
     KNOT[I]=(UNIQUENON[FLOOR(QUANTILE)]+UNIQUENON[CEIL(QUANTILE)])/2;
     Z2[,I]=((NON-KNOT[I])#((NON-KNOT[I])>0))##2;
END;

/**name_b=rowcatc({}(1,nx,'B'))`||(char(1:nx))`);**/

NAMEZ  =ROWCATC({J(1,K,'Z2_'))`||(CHAR(1:K))`);
NAME={'Y' 'X1' 'X2' 'NON' 'NON2'}||T(NAMEZ);
NAME=T(NAME);

/* SAVING THE NAME OF VARIABLES */
CREATE IN.DATANAMEVAR FROM NAME[COLNAME={DVAR}];
APPEND FROM NAME;SAVE;CLOSE IN.DATANAMEVAR;


NAME=T(NAME);
DATA=Y||X1||X2||NON||NON2||Z2;

CREATE IN.COBA1 FROM DATA[COLNAME=NAME];  /* SAVING DATA SET BEING RUNNING */
APPEND FROM DATA;SAVE;CLOSE IN.COBA1;
KNOT=T(KNOT);

CREATE IN.DKNOT FROM KNOT[COLNAME={'KNOT'}];  /* SAVING KNOT  */
APPEND FROM KNOT;SAVE;CLOSE IN.DKNOT;

quit;
%include 'd:\Temp\Test\glimmix.sas';
%glimmix(data=in.coba1,procopt=method=reml,
        stmts=%str(
                    model y=x1 x2 non non2/ solution;
                    random
Z2_1 Z2_2 Z2_3 Z2_4 Z2_5 Z2_6 Z2_7 Z2_8 Z2_9 Z2_10
Z2_11 Z2_12 Z2_13 Z2_14 Z2_15 Z2_16 Z2_17 Z2_18 Z2_19 Z2_20
Z2_21 Z2_22 Z2_23 Z2_24 Z2_25 Z2_26 Z2_27 Z2_28 Z2_29 Z2_30
Z2_31 Z2_32 Z2_33 Z2_34 Z2_35

                     / type=toep(1) s;
                         ods output solutionR=BetaAcak;
                         ods output solutionF=BetaFixed;
                         ),
                    error=binomial,
                    link=logit,
                    out=fitted);
```

```
PROC GENMOD DATA=IN.COBA1 DESC;  /* PROC GENMOD FOR PARAMETRIC APPROACH */
MODEL Y = X1 X2 NON / DIST = BIN
LINK = LOGIT;
ODS OUTPUT PARAMETERESTIMATES=BETASEMTOPARAM;
ODS OUTPUT CONVERGENCESTATUS=STATUSSEMTOPARAM;
RUN;
/*dm 'log;clear;output;clear';
*/


PROC IML;
/* --------------------------------
FOR SEMIPARAMETRIC APPROAH
---------------------------------*/


USE IN.POINT100ORI;     /* TAKING 100 POINT DATA FOR POINT ESTIMATES*/
READ POINT 1 INTO POINT;
CLOSE IN.POINT100ORI;

POINT=T(POINT);

USE IN.DKNOT;
READ ALL INTO KNOT;
CLOSE IN.DKNOT;

Z=J(NROW(POINT),35,.);
DO I=1 TO 35;           /* CREATING BASIS */
     Z[,I]=((POINT-KNOT[I])##2)#(POINT>KNOT[I]);
END;
Z100=J(NROW(POINT),1,1)||POINT||POINT##2||Z;


USE BETAFIXED;
READ ALL VAR{ESTIMATE} INTO ESTIMATE_F;
CLOSE BETAFIXED;

USE BETAACAK;
READ ALL VAR{ESTIMATE} INTO ESTIMATE_R;
CLOSE BETAACAK;

BETAR=ESTIMATE_F[{1 4 5}]//ESTIMATE_R;
NONPAR=Z100*BETAR;


NSAMPLE=&SAMPLEN;
PERCENTAGE=&PERCENT;
NOITER=&ITERASI;

ESTIMATE_F=T(ESTIMATE_F)||NSAMPLE||PERCENTAGE||NOITER;
EDIT IN.FSEMTOSEM;
APPEND FROM ESTIMATE_F;
CLOSE IN.FSEMTOSEM;

ESTIMATE_R=T(ESTIMATE_R)||NSAMPLE||PERCENTAGE||NOITER;
EDIT IN.RSEMTOSEM;
APPEND FROM ESTIMATE_R;
CLOSE IN.RSEMTOSEM;

NONPAR=T(NONPAR)||NSAMPLE||PERCENTAGE||NOITER;
EDIT IN.POINT100STS;
APPEND FROM NONPAR;
CLOSE IN.POINT100STS;
```

```
/* ------------------------------
FOR SEMIPARAMETRIC APPROAH
------------------------------*/
USE BETASEMTOPARAM;
READ ALL VAR{ESTIMATE} INTO ESTIMATE_F;
CLOSE BETASEMTOPARAM;

USE statussemtoparam;
READ ALL VAR{STATUS};
CLOSE statussemtoparam;

NSAMPLE=&SAMPLEN;
PERCENTAGE=&PERCENT;
ESTIMATE_F=T(ESTIMATE_F)||NSAMPLE||PERCENTAGE||STATUS||NOITER;
EDIT IN.SEMTOPARAM;
APPEND FROM ESTIMATE_F;
CLOSE IN.SEMTOPARAM;

QUIT;

%END;
%MEND COMBINE;
```

CATATAN:
Macro "glimmix.sas" tidak ditampilkan disini. Bisa dilihat pada:
www.umich.edu/~kwelch/genmod/glimmix.sas