Statistika, Vol. 8 No. 1, 19 – 29
Mei 2008

Pemilihan Model Terbaik pada Mars Respon Kontinu

BAMBANG WIDJANARKO OTOK

Tenaga Pengajar di Jurusan Statistika, ITS, Surabaya
e-mail: bambang_wo@statistika.its.ac.id; otok_bw@yahoo.com

ABSTRAK

Multivariate adaptive regression spline (MARS) adalah salah satu model regresi nonparametrik, yaitu model yang mengasumsikan bentuk hubungan fungsional antara variabel respon dan prediktor tidak diketahui. MARS adalah kombinasi yang komplek antara metode spline dengan rekursif partisi untuk menghasilkan estimasi fungsi regresi yang kontinu.Hasil penelitian menunjukkan bahwa estimasi parameter model MARS untuk variabel respon kontinu dilakukan dengan penalized least square (PLS). Pemilihan model MARS terbaik dilakukan dengan prosedur forward dan backward stepwise didasarkan pada nilai GCV. Prosedur forward adalah tahapan untuk mendapatkan fungsi basis maksimum yang mencakup pengaruh efek utama, interaksi, dan knot. Sedangkan prosedur backward adalah tahapan untuk mengeliminasi fungsi basis yang kontribusinya tidak signifikan. Hasil kajian juga menunjukkan bahwa GCV dengan potongan regresi linear dapat terbukti bekerja dengan baik dalam menentukan pemilihan model terbaik pada MARS respon kontinu.

Kata-kata Kunci: MARS, Penalized Least Square, GCV.

1. Pendahuluan

Dalam berbagai disiplin ilmu, permasalahan penelitian umumnya adalah mendekati sebuah fungsi beberapa variabel prediktor pada berbagai titik dalam ruang variabel respon. Masalah ini terjadi dalam bidang statistika (regresi nonparametrik multivariat), dan bidang ilmu komputasi dan rekayasa (*neural network*). Hal ini diakibatkan kemajuan yang cukup pesat pada teknik-teknik modern untuk pengumpulan data, yang telah menghasilkan data dalam jumlah besar baik dalam ukuran maupun dimensi.

Untuk menyelesaikan permasalahan tersebut diperlukan suatu model pada variabel respon (Y) dalam satu atau lebih variabel prediktor $(x_1,...,x_m)$ yang memberikan hubungan pada data. Secara umum, analisis yang digunakan untuk mengeksplorasi hubungan antara (Y) dan $(x_1,...,x_m)$ dikenal dengan regresi.

Regresi linear adalah salah satu pendekatan parametrik yang paling populer dalam pemodelan data. Dalam praktek, pendekatan parametrik seringkali tidak fleksibel dalam memodelkan pola nonlinear yang tersembunyi pada data berdimensi tinggi. Hal ini memotivasi penggunaan regresi nonparametrik, yang mengutamakan fleksibilitas dengan mengasumsikan smooth (mulus) dalam arti fungsi regresi yang tidak diketahui adalah kontinu dan diferensiabel.

Multivariate adaptive regression spline (MARS) adalah salah satu model regresi nonparametrik, yaitu model yang mengasumsikan bentuk hubungan fungsional antara variabel respon dan prediktor tidak diketahui, dan mempunyai bentuk fungsional yang fleksibel. MARS adalah metode yang diperkenalkan oleh Friedman (1991). Metode ini adalah implementasi teknik-teknik untuk menyelesaikan masalah regresi, dengan tujuan untuk memprediksi variabel respon yang bernilai kontinu berdasar beberapa variabel prediktor.

Model MARS disusun pada pengaturan beberapa koefisien fungsi basis dimana secara keseluruhan dikendalikan pada data regresi. Hastie *et al.* (2002) menyatakan bahwa model MARS berguna untuk mengatasi permasalahan data dimensi tinggi yang dikenal dengan curse of dimensionality dan menghasilkan prediksi variabel respon yang akurat, serta untuk

mengatasi kelemahan regresi partisi rekursif (RPR) yaitu menghasilkan model yang kontinu pada knot, yang didasarkan pada nilai *generalized cross validation* (GCV) minimum.

Oleh karena itu penelitian mengkaji secara teoritis pemilihan model terbaik pada MARS dengan kriteria GCV. Selain itu, dilakukan simulasi dan kajian studi kasus untuk implementasi dari kajian teori.

2. Tinjauan Pustaka

Friedman (1991) menyatakan bahwa model MARS adalah kombinasi yang komplek antara metode spline dengan rekursif partisi untuk menghasilkan estimasi fungsi regresi yang kontinu. Spline adalah salah satu jenis potongan polinomial, yaitu polinomial yang memiliki sifat tersegmen. Sifat tersegmen ini memberikan fleksibilitas lebih daripada polinomial biasa, sehingga memungkinkan untuk menyesuaikan diri secara lebih efektif terhadap karakteristik lokal pada suatu fungsi atau data. Wahba (1990) menunjukkan bahwa spline memiliki sifat-sifat statistik yang berguna untuk menganalisis hubungan dalam regresi. Spline dalam regresi nonparametrik terus berkembang sampai pada model adaptive (Bilier dan Fahrmeir, 2001) dan multivariat respon (Holmes dan Mallick, 2003). Selain itu, He dan Shi (1998) mengembangkan pendekatan monotonicity untuk mengestimasi fungsi basis spline, sedang-kan Hall dan Opsomer (2005) menggunakan pendekatan penalti kuadrat terkecil.

Regresi partisi rekursif (RPR) merupakan pendekatan fungsi yang tidak diketahui dengan menggunakan pengembangan fungsi basis. Selain itu, RPR adalah suatu konsep geometri yang membagi daerah dengan dasar aritmetik, yaitu penjumlahan dan perkalian. Morgan dan Sonquist (1963) memperkenalkan RPR dalam riset *automatic interaction detection* (AID), selanjutnya Morgan dan Messenger (1973) menemukan *Theta* AID (THAID), yang digunakan untuk memproduksi pohon-pohon klasifikasi. Venables dan Ripley (1994) menggunakan aturan pengklasifikasi pada metode THAID untuk memprediksi suatu obyek.

Model MARS digunakan untuk mengatasi kelemahan RPR yaitu menghasilkan model yang kontinu pada knot. Beberapa perbaikan yang dilakukan untuk mengatasi keterbatasan RPR, antara lain menghasilkan fungsi basis menjadi:

$$B_m^{(q)}(\mathbf{x}) = \prod_{k=1}^{K_m} \left[s_{km} \cdot (x_{v(k,m)} - t_{km}) \right]_+^q$$

Estimasi dari kurva regresi $f(x)$ secara umum didapatkan melalui penalized least square (PLS) yakni meminimumkan persamaan berikut:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2 + \eta\int_a^b \left(f^{(m)}\left(x\right)\right)^2 dx$$

Setelah dilakukan modifikasi model RPR dan dikombinasikan dengan spline, estimator model MARS dapat ditulis sebagai berikut:

$$\hat{f}(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})] \qquad (1)$$

dengan

α_0 = konstanta (basis induk)

α_m = koefisien dari fungsi basis ke-m

M = banyaknya fungsi basis (*nonconstant basis function*)

K_m = derajat interaksi

s_{km} = nilainya ±1

$x_{v(k,m)}$ = variabel independen

t_{km} = nilai knot dari variabel prediktor $x_{v(k,m)}$

Dengan menggunakan estimator MARS, maka model MARS adalah :

$$y_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [s_{km}.(x_{v(k,m)} - t_{km})] + \varepsilon_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m B_m(\mathbf{x}) + \varepsilon_i \tag{2}$$

dengan

$$B_m(\mathbf{x}) = \prod_{k=1}^{K_m} [s_{km}.(x_{v(k,m)} - t_{km})]$$

Dalam bentuk matriks dapat ditulis menjadi:

$$\mathbf{Y} = \mathbf{B} + \tag{3}$$

dengan,

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^T, \quad \boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_m)^T, \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$$

$$\mathbf{B} = \begin{pmatrix} 1 & \prod_{k=1}^{K_1} s_{1m}(x_{1(1,m)} - t_{1m}) & \cdots & \prod_{k=1}^{K_M} s_{Mm}(x_{1(M,m)} - t_{Mm}) \\ 1 & \prod_{k=1}^{K_1} s_{1m}(x_{2(1,m)} - t_{1m}) & \cdots & \prod_{k=1}^{K_M} s_{Mm}(x_{2(M,m)} - t_{Mm}) \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 1 & \prod_{k=1}^{K_1} s_{1m}(x_{n(1,m)} - t_{1m}) & \cdots & \prod_{k=1}^{K_M} s_{Mm}(x_{n(M,m)} - t_{Mm}) \end{pmatrix}$$

3. Estimasi Dan GCV Model Mars Respon Kontinu

Model MARS dalam persamaan (3), $\hat{\boldsymbol{\alpha}}$ merupakan parameter yang akan diestimasi dari data, melalui *penalized least square* (PLS) yang telah dimodifikasi. Hal ini dijabarkan dalam bukti teorema dibawah ini.

Teorema 1. *Dengan menggunakan estimator MARS dalam Persamaan (1), dan* $\mathbf{B}$ *matrik non singular dan parameter smoothing* $\eta > 0$ *maka penduga dari* $\hat{\boldsymbol{\alpha}}$ *adalah*

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T\mathbf{Y} \tag{4}$$

dengan

$$\mathbf{B} = [1, \ (x_{v,(k,m)} - t_{km})_1^K]$$

Bukti: Perhatikan persamaan:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2 + \eta\int_a^b \left(f^{(m)}\left(x\right)\right)^2 dx$$

Dengan memperhatikan,

$$f(\mathbf{x}_i, \hat{\ }) = \sum_{m=0}^{M} \quad m \quad m(\ ) \tag{5}$$

$$f'(\mathbf{x}_i, \hat{\ }) = \sum_{m=0}^{M} \quad m \quad 'm(\ ) \tag{6}$$

$$f''(\mathbf{x}_i, \hat{\ }) = \sum_{m=0}^{M} \quad m \quad ''m(\ ) \tag{7}$$

dan,

$$\int_a^b \left(f^{(")}(x)\right)^2 dx = \int_a^b \sum_{m=0}^{M} \boldsymbol{\alpha}_m \mathbf{B}^{"}_m(\mathbf{x})\, dx$$
$$= \int_a^b [(\mathbf{B}^{"}(\mathbf{x}))^T \quad]^2 dx \tag{8}$$

misalkan,

$$(\mathbf{B}^{"}(\mathbf{x}))^T (\mathbf{B}^{"}(\mathbf{x})) = S(\mathbf{x}) \tag{9}$$

maka,

$$\int_a^b \left(f^{(")}(x)\right)^2 dx = \int_a^b [(\mathbf{B}^{"}(\mathbf{x}))^T \quad]^2 dx = \quad ^T \int_a^b S(\quad)\ dx = \quad ^T \quad = J(\psi) \tag{10}$$

dengan,

$$d_{ij} = \int_a^b \left(\frac{\partial^2 \mathbf{B}_i(\mathbf{x})}{\partial \mathbf{x}^2}\right)\left(\frac{\partial^2 \mathbf{B}_j(\mathbf{x})}{\partial \mathbf{x}^2}\right) d\mathbf{x} \tag{11}$$

Sehingga persamaan di atas menjadi

$$ASR + RP \tag{12}$$

atau

$$\frac{1}{n}\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \hat{\ }))^2 + \eta J(\psi) \tag{13}$$

dalam bentuk matriks

$$ASR(\boldsymbol{\alpha}) = (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) + \eta \boldsymbol{\alpha}^T \mathbf{D} \boldsymbol{\alpha} \tag{14}$$

Misalkan $ASR(\boldsymbol{\alpha}) = ASR + RP$, maka koefisien fungsi basis $\hat{\boldsymbol{\alpha}}$ diperoleh dengan meminimumkan persamaan (13) atau (14).

$ASR(\boldsymbol{\alpha})$ minimum maka

$$\frac{\partial ASR(\hat{\boldsymbol{\alpha}})}{\partial \hat{\boldsymbol{\alpha}}} = 0$$

$$\frac{\partial ASR(\hat{\boldsymbol{\alpha}})}{\partial \hat{\boldsymbol{\alpha}}} = \frac{\partial}{\partial \hat{\boldsymbol{\alpha}}}[(\mathbf{Y} - \mathbf{B} \quad)^T (\quad - \quad) + \eta \quad ^T \quad] = 0 \tag{15}$$

Sehingga,

$$\mathbf{B}^T\mathbf{Y} - \mathbf{B}\mathbf{Y}^T + 2\mathbf{B}^T\mathbf{B}\hat{\ } + 2\eta \quad \hat{\ } = 0$$
$$-2\mathbf{B}^T\mathbf{Y} + 2\mathbf{B}^T\mathbf{B}\hat{\ } + 2\eta \quad \hat{\ } = 0$$
$$\mathbf{B}^T\mathbf{B}\hat{\ } + \eta \quad \hat{\ } = \quad ^T$$
$$(\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})\hat{\ } = \quad ^T$$
$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T\mathbf{Y} \tag{16}$$

Penyelesaian yang diberikan dalam Persamaan (16) adalah meminimumkan $ASR(\boldsymbol{\alpha})$ dan merupakan estimator bagi $\hat{\boldsymbol{\alpha}}$. Jadi Teorema 1 terbukti. □

Teorema 2. *Dengan menggunakan estimator MARS dalam persamaan (1), dan* $\mathbf{B}$ *matrik non singular dan parameter smoothing* $\eta = 0$ *maka penduga dari* $\hat{\boldsymbol{\alpha}}$ *adalah*

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y} \qquad (17)$$

dengan $\mathbf{B} = [1,\ (x_{v,(k,m)} - t_{km})_1^K]$

Bukti: Perhatikan persamaan:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + \eta\int_a^b \left(f^{(m)}(x)\right)^2 dx$$

dengan $\eta = 0$, maka persamaan di atas menjadi:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 \qquad (18)$$

Dari Persamaan (3), maka $\hat{f}(x_i) = \mathbf{B}$, sehingga Persamaan (18) menjadi:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{B})^2 = (\quad - \quad)^{\mathrm{T}}(\quad - \quad) = Z$$

Untuk memperoleh estimator $\hat{\boldsymbol{\alpha}}$ digunakan metode kuadrat terkecil, yang pada prinsipnya meminumumkan Z, dinyatakan sebagai berikut:

$$Z = (\mathbf{Y} - \mathbf{B}\hat{\ })^{\mathrm{T}}(\quad - \quad\hat{\ })$$

$$Z = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \hat{\ }^{\mathrm{T}}\quad - \quad^{\mathrm{T}}\hat{\ } + \hat{\ }^{\mathrm{T}}\quad\hat{\ })$$

$$Z = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\hat{\ }^{\mathrm{T}}\quad + \hat{\ }^{\mathrm{T}}\quad\hat{\ })$$

Untuk memperoleh persamaan normal dilakukan dengan menurunkan secara parsial terhadap $\hat{\boldsymbol{\alpha}}$ dengan hasil sebagai berikut:

$$\frac{\partial Z}{\partial \hat{\boldsymbol{\alpha}}} = -2\mathbf{B}^T\mathbf{Y} + 2\mathbf{B}^T\mathbf{B}\hat{\ } = 0$$

$$-\mathbf{B}^T\mathbf{Y} + \mathbf{B}^T\mathbf{B}\hat{\ } = 0$$

$$\mathbf{B}^T\mathbf{B}\hat{\boldsymbol{\alpha}} = \mathbf{B}^T\mathbf{Y} \qquad (19)$$

karena $\mathbf{B}$ *matriks non singular,* maka

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y} \qquad (20)$$

sedangkan turunan kedua terhadap $\hat{\boldsymbol{\alpha}}$ adalah

$$\frac{\partial^2 Z}{\partial \hat{\boldsymbol{\alpha}}^2} = 2\mathbf{B}^T\mathbf{B} > \mathbf{0}.$$

Jadi Teorema 2 terbukti. □

Pada pemodelan MARS, pemilihan model digunakan metode stepwise. Forward stepwise dilakukan untuk mendapatkan fungsi dengan jumlah fungsi basis maksimum. Kriteria pemilihan fungsi basis pada *forward stepwise* adalah dengan meminimumkan ASR. Untuk memenuhi konsep parsemoni (model sederhana) dilakukan *backward stepwise* yaitu memilih fungsi basis yang dihasilkan dari forward stepwise dengan meminimumkan nilai *generalized cross-validation* (GCV).

Untuk membuktikan GCV pada model MARS diperlukan Teorema berikut.

Teorema 3. *Jika rank dari* $\mathbf{P}$ *adalah (M+1) maka* $\mathbf{P}$ *memiliki (M+1) unit nilai eigen yang tidak bernilai nol dan sebanyak [N-(M+1)] unit nilai eigen yang bernilai nol.*

Bukti: Misalkan λ adalah nilai eigen dan Γ vektor eigen, maka:

$$\mathbf{P}\Gamma = \lambda\Gamma \ (\Gamma \neq 0)$$

jika dikalikan Γ^T menjadi $\Gamma^T\mathbf{P}\Gamma = \lambda\Gamma^T\Gamma$ dan $\mathbf{P}$ adalah matrik idempoten ($\mathbf{P} = \mathbf{P}^2$) maka $\Gamma^T\mathbf{P}^T\mathbf{P}\Gamma = \lambda\Gamma^T\Gamma$, sehingga $(\lambda\Gamma)^T\lambda\Gamma = \lambda\Gamma^T\Gamma$ atau $\lambda(\lambda-1)\Gamma^T\Gamma = 0$. Hal ini menunjukkan bahwa nilai eigen dari $\mathbf{P}$ bernilai 1 sebanyak (M+1) dan bernilai 0 sebanyak [N-(M+1)], sehingga rank dari matrik $\mathbf{P}$ sama dengan (M+1). □

Teorema 4. *Misalkan* $\mathbf{R}$ *adalah matriks kuadratik dengan* $\mathbf{R}^{-1}(\mathbf{R}^{-1})^T = \mathbf{X}^T\mathbf{X}$, $\mathbf{B}^{-1}$ *adalah faktor Cholesky dari* $\mathbf{X}^T\mathbf{X}$. *Misalkan U dan Q matriks diagonal sedemikian hingga* $\mathbf{U}\mathbf{Q}^{-1}\mathbf{U}^T = \mathbf{R}D\mathbf{R}^T$. *selanjutnya,* $\mathbf{Z} = \mathbf{X}(\mathbf{R}^T\mathbf{U})$ *maka* $\mathbf{Z}^T = \mathbf{U}^T\mathbf{R}\mathbf{X}^T$ *dan misalkan* $\hat{\lambda} = \mathbf{U}(\mathbf{R}^{-1})^T\beta = (\mathbf{R}^T\mathbf{U})^{-1}\hat{\beta}$ *maka penyelesaian* $\hat{\lambda}$ *adalah:*

$$(\mathbf{I} + \delta^2\mathbf{Q})\hat{\lambda} = \mathbf{Z}^T\mathbf{Y} = (\mathbf{U}^T\mathbf{B})\mathbf{X}^T\mathbf{Y}$$

Selanjutnya $\mathbf{X}\hat{\beta} = \mathbf{Z}\hat{\lambda}$ *dan matriks Hat,* $S(\delta^2) = \mathbf{Z}(\mathbf{I} + \delta^2\mathbf{Q})^{-1}\mathbf{Z}^T$ *dengan derajat bebas,*

$$tr[S(\delta^2)] = tr\{(\mathbf{I} + \delta^2\mathbf{Q})^{-1}\mathbf{Z}^T\mathbf{Z}\} = tr\{(\mathbf{I} + \delta^2\mathbf{Q})^{-1}\} = \sum\nolimits_i (1 + \delta^2 Q_i)^{-1}$$

dimana Q_i *adalah matrik diagonal ke-i dari* $\mathbf{Q}$.

Bukti:

$$\mathbf{X}^T\mathbf{X} + \delta^2 D = \mathbf{R}^{-1}(\mathbf{R}^{-1})^T + \delta^2 D = \mathbf{R}^{-1}(\mathbf{R}^{-1})^T + \mathbf{R}^{-1}\mathbf{R}\delta^2 D\mathbf{R}^T(\mathbf{R}^{-1})^T$$
$$= \mathbf{R}^{-1}(I + \delta^2\mathbf{R}D\mathbf{R}^T)(\mathbf{R}^{-1})^T = \mathbf{R}^{-1}\mathbf{U}(I + \delta^2\mathbf{Q})\mathbf{U}^T(\mathbf{R}^{-1})^T$$

dan juga, $\mathbf{U}^T\mathbf{R}(\mathbf{X}^T\mathbf{X} + \delta^2 D)\mathbf{R}^T\mathbf{U} = I + \delta^2\mathbf{Q}$

ini berarti, $\mathbf{U}^T\mathbf{R}(\mathbf{X}^T\mathbf{X} + \delta^2 D)\mathbf{R}^T\mathbf{U}(\mathbf{U}^T(\mathbf{R}^{-1})^T\hat{\beta}) = \mathbf{U}^T\mathbf{R}(\mathbf{X}^T\mathbf{Y})$

atau, $(I + \delta^2\mathbf{R})\hat{\lambda} = \mathbf{Z}^T\mathbf{Y}$ juga $\mathbf{X} = \mathbf{Z}\mathbf{U}^T(\mathbf{R}^{-1})^T$

sehingga $\mathbf{X}\hat{\beta} = \mathbf{Z}\mathbf{U}^T(\mathbf{R}^{-1})^T\hat{\beta} = \mathbf{Z}\hat{\lambda}$. Jadi Teorema 4 terbukti. □

Teorema 5. *Misalkan digunakan model MARS pada Persamaan (2), maka* η *optimal diperoleh dengan kriteria GCV sebagai berikut:*

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i, \ ))^2}{\left\{1 - \frac{C(\tilde{M})}{n}\right\}^2} \qquad (21)$$

dengan, $\tilde{C}(M) = C(M) + d.M$, $\mathrm{C(M)} = \mathrm{Trace}(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T) + 1$, $2 \le d \le 4$.

Bukti: Perhatikan Persamaan (13) dan estimator fungsi MARS kontinu adalah $\hat{\boldsymbol{\alpha}} = (\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T\mathbf{Y}$, maka $f(\mathbf{x}_i, \hat{\ }) = H(\hat{\ })Y$ dengan, $H(\hat{\boldsymbol{\alpha}}) = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T$.

Selanjutnya dengan memperhatikan persamaan berikut,

$$CV(\eta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f^{(-i)}(\mathbf{x}_i, \hat{\ }))^2 \text{ dan } GCV(\eta) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \hat{\ }))^2}{[\frac{1}{n}tr(I - A(\eta))]^2}$$

maka banyaknya parameter model MARS selain basis induk, dapat diperoleh dengan mensubstitusikan persamaan (20) ke dalam persamaan (3) sebagai berikut:

$$\mathbf{Y} = \mathbf{B}[(\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{Y})] = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y} = \mathbf{P}\mathbf{Y} \qquad (22)$$

Dengan pembuktian Teorema 3, maka $\mathbf{P} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ berukuran $(M+1)\times(M+1)$. Karena $\mathbf{P}$ adalah matrik simetris dan idempoten maka Trace dari matriks $\mathbf{P}$ sama dengan rank dari $\mathbf{P}$ yang merupakan banyaknya parameter fungsi basis selain konstanta dan banyaknya parameter yang diestimasi, dinotasikan sebagai $C(M) = \text{Trace}(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T) + 1$.

Selanjutnya berdasarkan Teorema 4 dan persamaan (21), maka penyebut pada GCV, adalah

$$\mathbf{Y} = \mathbf{B}\hat{\ } = [(\ ^T\ + \eta\ )^{-1}(\ ^T\ )] = \ (\eta)$$

dengan, $\mathbf{A}(\eta) = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T$. Sehingga,

$$\begin{aligned}
\left\{\frac{1}{n}tr[I - A(\eta)]\right\}^2 &= \left\{\frac{1}{n}tr[I - \mathbf{B}(\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T]\right\}^2 \\
&= \left\{1 - \frac{1}{n}tr[\mathbf{B}(\mathbf{B}^T\mathbf{B} + \eta\mathbf{D})^{-1}\mathbf{B}^T]\right\}^2 \\
&= \left\{1 - \frac{1}{n}tr[\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T + \mathbf{B}\eta\mathbf{D}^{-1}\mathbf{B}^T]\right\}^2 \\
&= \left\{1 - \frac{1}{n}\left[tr(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T + 1) + tr(\mathbf{B}\eta\mathbf{D}^{-1}\mathbf{B}^T)\right]\right\}^2 \\
&= \left\{1 - \frac{1}{n}\left[\underbrace{tr(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T + 1) + d.M}_{C(\tilde{M})}\right]\right\}^2 = \left\{1 - \frac{C(\tilde{M})}{n}\right\}^2
\end{aligned}$$

penambahan 1 pada $tr(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ karena dalam model MARS selalu melibatkan basis induk (α_0), sedangkan *d* disarankan bernilai $2 \le d \le 4$. Jadi Teorema 5 terbukti. □

Berikut disajikan prosedur *forward* dan *backward*:

a. Prosedur Stepwise Forward MARS.

Algoritma 1. *Forward Stepwise*

(1) $B_1(x) \leftarrow 1$; $M \leftarrow 2$
(2) Loop until $M > M_{max}$: $lof^* \leftarrow \sim$
(3) For $m = 1$ to $(M-1)$ do:
(4) For $v \notin \{v(k,m) \mid 1 \le k \le K_m\}$
(5) For t $\in \{x_{vj} \mid B_m(x_j) > 0\}$
(6) $g \leftarrow \sum_{t=1}^{M-1} a_i B_i(x) + a_M B_m(x)[+(x_v - t)]_+ + a_{M+1} B_m(x)[-(x_v - t)]_+$
(7) lof $\leftarrow$ *min* $a_1,\ldots,a_{M+1}$ $LOF(g)$

(8) if (lof < lof*) then (lof* < lof);
$m^* \leftarrow m;\ v^* \leftarrow v;\ t^* \leftarrow t;$
(9) end if
(10) end for
(11) end for
(12) end for
(13) $B_M(x) \leftarrow B_{m^*}(x)[+(x_{v^*} - t^*)]_+$
(14) $B_{M+1}(x) \leftarrow B_{m^*}(x)[-(x_{v^*} - t^*)]_+$
(15) M $\leftarrow (M+2)$
(16) end loop
(17) end algoritma

Algoritma (1) mengimplementasikan bagian forward pada MARS dengan memodifikasi algoritma regresi partisi rekursi (RPR). Pangkat (q=1) pada fungsi basis disubstitusikan untuk step function pada baris 6, 12, 13. Parent basis function dimasukkan dalam memodifikasi model pada baris 6 dan sisanya di update melalui logika pada baris 12-14. Perkalian fungsi basis dipaksa untuk memperoleh faktor meliputi variabel yang berbeda dengan kontrol loop atas variabel dalam baris 4. Algoritma ini akan menghasilkan Mmaxq=1 truncated power. Fungsi basis merupakan subset dari tensor yang lengkap, basis dengan lokasi knot pada semua marginal data yang berbeda nilainya. Seperti pada algoritma regresi partisi rekursi (RPR) pada set basis function dihilangkan dengan backward stepwise untuk menghasilkan set basis function akhir. Lokasi knot dihubungkan dengan pendekatan ini, digunakan turunan basis piecewise cubic yang kontinu pada turunan pertama, sehingga menghasilkan model akhir dengan turunan yang kontinu.

b. Prosedur eliminasi backward MARS.
Algoritma (2). *Backward Stepwise*
(1) $J^*=\{1,2,\ldots, M_{max-1};\ K^* \leftarrow K$
(2) $lof \leftarrow \min_{\{a_j | j \in J^*\}} LOF[\sum_{j \in J^*} \alpha_j B_j(x)]$
(3) For M = M_{max} to 2 do: b $\leftarrow$ ~; $L \leftarrow K^*$
(4) For m = 2 to M do: K $\leftarrow L-\{m\}$~; $L \leftarrow K^*$
(5) $lof \leftarrow \min_{\{\alpha_k | k \in K\}} LOF[\sum_{k \in K} \alpha_k B_k(x)]$
(6) if (lof < b) then (b $\leftarrow lof$); $K^* \leftarrow K$; end if
(7) if (lof < lof*) then (lof* $\leftarrow lof$); $J^* \leftarrow K$; end if
(8) end for
(9) end for
(10) end algoritma

Tahap *backward stepwise* pada MARS menghasilkan region yang overlap. Pembuangan fungsi basis pada MARS tidak menghasilkan lubang pada ruang variabel prediktor (selama basis konstan tidak dihilangkan).

Pada algoritma (2), baris pertama menunjukkan inisial dari fungsi basis J^* yang dihasilkan oleh algoritma (1), setiap iterasi di luar *for loop* menyebabkan satu fungsi basis dibuang dan didalam *for loop* untuk memilih fungsi basis. Selama melakukan *backward stepwise* tidak pernah membuang konstanta basis $B_1(x)=1$. Algoritma ini akan membentuk barisan model sebanyak (M_{max-1}), setiap barisan mempunyai satu fungsi basis lebih sedikit dibandingkan dengan barisan sebelumnya. Model terbaik barisan ini berada pada kumpulan J^* saat selesai seleksi. Hal ini adalah bagian pada MARS yang menyeleksi submodel dalam meminimumkan prediksi estimasi *residual* dengan kriteria *generalized cross validation* (GCV) pada persamaan (21).

4. Hasil Empiris

Kajian empiris dalam penelitian ini meliputi dua bahasan. Pertama dilakukan untuk menunjukkan bahwa GCV dapat bekerja dengan baik dalam proses penentuan model MARS

terbaik. Bagian ini dilakukan dengan menggunakan suatu data simulasi untuk fungsi yang nonlinear. Kedua difokuskan pada pemodelan MARS respon kontinu pada data real yaitu emisi gas buang kendaraan berbahan bakas solar (opasitas)

Eksperimen simulasi dalam rangka menunjukkan bagaimana model MARS yang diperkenalkan dapat bekerja dengan baik. Kriteria model MARS dapat bekerja lebih baik dilihat dari nilai generalized cross validation (GCV) terkecil. Kajian empirik menggunakan data simulasi adalah membangun model kurva regresi, $y_i = f(x_i) + \varepsilon_i$, i = 1, 2, ..., n dengan fungsi:

SIMULASI 1: $f(x_i) = 5e^{-5x_1}$ dengan $n = 50,100,250$; $\sigma^2 = 0.025, 0.5, 1$; $\varepsilon_i \sim N(0,\sigma^2)$; $X_i \sim U(0,1)$

SIMULASI 2: $f(x_i) = \sin(2\pi x_1)$ dengan $n = 50,100,250$; $\sigma^2 = 0.025, 0.5, 1$; $\varepsilon_i \sim N(0,\sigma^2)$; $X_i \sim U(0,1)$

Nilai GCV dengan berbagai variansi pada fungsi, pengamatan (n) secara lengkap tersaji pada Tabel 1 berikut.

Tabel 1. Nilai GCV Pada Spline Truncated dan MARS

Fungsi	n	σ^2	GCV	
			Spline	**MARS**
SIMULASI 1	50	0,025	**0.00058**	0.0013
		0,5	0.13982	0.1636
		1	0.50940	0.7686
	100	0,025	0.00072	0.0011
		0,5	0.17331	0.2311
		1	0.67544	0.8576
	250	0,025	0.00073	0.0009
		0,5	0.20838	0.2550
		1	0.77183	0.9650
SIMULASI 2	50	0,025	0.00127	**0.0011**
		0,5	0.15802	0.1688
		1	0.45707	0.7298
	100	0,025	0.00125	0.0011
		0,5	0.18962	0.1997
		1	0.63835	0.9620
	250	0,025	0.00169	0.0009
		0,5	0.19760	0.2309
		1	0.78974	1.0030

Berdasarkan Tabel 1, pada model spline, ternyata nilai GCV semakin besar pada model simulasi 1, dengan pengamatan yang semakin besar dan varians σ2 konstan. Sedang pada model simulasi 2, dengan pengamatan yang semakin besar dan varians σ2 konstan memberikan nilai GCV yang semakin kecil. Sedangkan pada MARS, nilai GCV semakin kecil pada model simulasi 1, dengan pengamatan yang semakin besar dan varians σ2 konstan. Sedang pada model simulasi 2, dengan pengamatan yang semakin besar dengan varians σ2 = 0,025 memberikan nilai GCV yang semakin kecil, tetapi pada varians σ2 = 0,5 dan 1 nilai GCV bervariasi. Secara keseluruhan fungsi optimal terjadi pada varians σ2 kecil dengan n sembarang.

Sedangkan untuk pemodelan emisi gas buang berbahan bakar solar (0pasitas) dipengaruhi oleh variabel merk (X1), model (X2), jenis kendaraan (X3), tipe mesin (X4), jenis turbo (X5), tahun pembuatan (X6), isi silinder (cc) (X7) dan jarak tempuh (km) (X8). Dalam pemodelan MARS, untuk memperoleh model terbaik disarankan jumlah fungsi basis maksimum antara (2 sampai 4) kali jumlah variabel independen. Model MARS dengan variasi maksimum observasi dan interaksi tersaji dalam Tabel 2.

Berdasarkan Tabel 2, model terbaik adalah model 3. Hal dapat ditunjukkan pada nilai GCV yang paling kecil dan R^2 yang besar. Sehingga model MARS untuk opasitas dinyatakan dalam persamaan berikut:

$$\hat{f}(x) = 65.659 + 22.120\ BF3 - 2.060\ BF5$$

dengan,
BF1 = max(0, TAHUN_P - 1995.000);
BF3 = (MODEL = 2 OR MODEL = 7 OR MODEL = 8 OR MODEL = 11);
BF5 = (MODEL = 4 OR MODEL = 6 OR MODEL = 7 OR MODEL = 8 OR MODEL = 10 OR MODEL = 11 OR MODEL = 12) * BF1;

Persamaan di atas dapat diinterpretasikan sebagai berikut:

(i) Koefisien BF3
Setiap penambahan BF3 sebesar satu-satuan akan menaikkan Opasitas (Y) sebesar 22,120 persen dengan menganggap BF5 konstan. Lebih lanjut dapat dikatakan bahwa perubahan kenaikan opasitas (Y) terjadi pada Model kendaraan (Taft(2), Panther(7), Pregio(8) dan Kuda(11))

(ii) Koefisien BF5
Setiap penambahan BF5 sebesar satu-satuan akan menurunkan Opasitas (Y) sebesar 2,060 persen dengan menganggap BF3 konstan. Lebih lanjut dapat dikatakan bahwa perubahan penurunan opasitas (Y) terjadi pada model kendaraan (Everest(4), Elf(6), Panther(7), Pregio(8), Colt(10), Kuda(11), Dyna(12)) dengan tahun pembuatan di atas tahun 1995.

Variabel yang mempengaruhi opasitas adalah model kendaraan dan tahun pembuatan. Model kendaraan merupakan variabel yang penting pertama dengan kontribusi 100 persen dalam mempengaruhi emisi opasitas, selanjutnya tahun pembuatan merupakan variabel yang penting kedua dengan kontribusi 59,051 persen.

Tabel 2. Penentuan model MARS terbaik pada Opasitas

Model	BF	MI	MO	# BF	GCV		R^2	SE. Regression
					Linear	Kubik		
Model 1	16	3	0	2	*280,711*	283,329	0.960	15,859
Model 2	16	3	10	2	*274,532*	276,554	0,961	15,622
Model 3	16	2	10	2	*274,271*	276,291	0,961	15,622
Model 4	16	1	10	2	*275,785*	277,814	0,959	15,911

Keterangan: BF= fungsi basis, MI= maksimum interaksi, MO= minimum observasi pada setiap subregion

5. Kesimpulan

Estimasi parameter pada kurva regresi dilakukan dengan meminimumkan *penalized least-squares* (PLS) yang dimodifikasi bertujuan menemukan suatu penyelesaian α_0 dan α_m, $m = 1,...,M$. Sebagaimana telah dijabarkan dalam teorema 1 dan teorema 2 diperoleh estimasi untuk model MARS respon kontinu. Hasil kajian empiris menunjukkan bahwa GCV dapat bekerja dengan baik dalam menentukan pemilihan model terbaik yang diterapkan pada model MARS respon kontinu. Pada akhirnya, hasil kajian tentang GCV dikaitkan dengan data real memberikan temuan baru, yaitu diperolehnya pemilihan model terbaik didasarkan pada GCV dengan potongan regresi linear.

Daftar Pustaka

[1]. Biller, C., dan Fahrmeir, L. 2001. Bayesian varying-coefficient model using adaptive regression spline. *Statistical modeling*.

[2]. Friedman, J. H. 1991. Multivariate Adaptive Regression Splines (with Discussion). *The Annals of Statistics*, 19:1-141.

[3]. Hall, P., dan Opsomer, J. D. 2005. Theory for penalized spline regression. *Biometrika*, 92:1-105.

[4]. Hastie, T., Tibshirani, R., dan Friedman, J. H. 2001. *The Element of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in Statistics, New York.

[5]. He, X., dan Shi, P. 1988. Monote B-Spline Smoothing, *Journal of American Statistical Association*, 93.

[6]. Holmes, C. C., dan Mallick, B. K. 2003. Generalized Nonlinear Modeling with Multivariate Free-Knot Regression Spline, *Journal of American Statistical Association*, 98.

[7]. Morgan, J. N., dan Sonquist, J. S. 1963. Problem in the analysis of survey data and a proposal. *Journal of American Statistical Association*, 58:45-434.
[8]. Morgan, J., dan Messenger, R. 1973. THAID: A sequential search program for the analysis of nominal scale dependent variables. *(Technical Report). Institute for social Research*, University of Michigan, Ann Arbor, Michigan.
[9]. Venables, W. N. and Ripley, B. D. 1994. *Modern Applied Statistic with S Plus*. Springer Verlag, New York.
[10]. Wahba, G. 1990. *Spline Method for Observational Data: Society for industrial and Applied Mathematics*, Philadelphia, Pennyylvanic.