Jurnal Matematika UNAND
Vol. **VIII** No. **3** Hal. 68 – 76
Edisi Desember 2019
ISSN : 2303–291X

PENDUGAAN PERSENTASE PENDUDUK MISKIN DI PROVINSI SUMATERA BARAT MENGGUNAKAN *SMALL AREA ESTIMATION* DENGAN PENDEKATAN SEMIPARAMETRIK *PENALIZED SPLINE*

SHINTA MUTIA KARNEVA, HAZMIRA YOZZA, FERRA YANUAR

Program Studi S1 Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia.
email : Shintamutia39@gmail.com

Diterima 14 Oktober 2019 Direvisi 21 Oktober 2019 Dipublikasikan 3 Desember 2019

Abstrak. Kemiskinan merupakan masalah sosial yang belum teratasi oleh pemerintah hingga saat ini. Walaupun angka kemiskinan sudah menurun, tapi masi banyak penduduk di Indonesia dikatergorikan miskin. Hal ini dikarenakan tidak tepatnya sasaran kebijakan pemerintah. Agar hal tersebut tidak terjadi maka untuk mengimplementasikan program pengentasan kemiskinan diperlukan adanya informasi pada suatu daerah. Informasi yang diperlukan berupa persentase penduduk miskin yang didapat melalui survey. Persentase penduduk miskin merupakan penduduk yang memiliki rata-rata pengeluaran perkapita perbulan di bawah garis kemiskinan. Survei penduduk merupakan salah satu cara yang digunakan untuk memperoleh informasi mengenai kependudukan. Jika survei dilakukan di area yang besar, maka dapat dihasilkan pendugaan parameter yang cukup akurat. Keterbatasan objek survei menyebabkan data yang di duga dengan pendugaan parameter secara langsung tidak menghasilkan dugaan yang akurat. Untuk menghasilkan pendugan yang lebih baik maka digunakan metode pendugaan tidak langsung pada area kecil (*Small Area Estimation*). Pada SAE, ada informasi lain yang diasumsikan dapat dipinjam untuk memperbaiki pendugaan terhadap parameter yang menjadi perhatian. Informasi dapat berupa variabel yang sama pada area lain atau variabel lain pada area yang sama yang dipandang berkaitan dengan parameter yang akan diduga. Salah satu pendekatan yang dapat digunakan adalah pendekatan semiparametrik *Penalized Spline* (P-spline) yang memiliki model fleksibel karena keberadaan dua komponen dalam model yang mengakomodasi hubungan antara variabel respon dengan variabel prediktor yang bersifat linier dan hubungan antar variabel respon dengan variabel prediktor yang bersifat nonlinier. Pendugaan persentase kemiskinan dibandingkan dalam empat model, dimana tiga variabel prediktor diasumsikan parametrik dan satu variabel prediktor diasumsikan nonparametrik. Evaluasi hasil pendugaan persentase kemiskinan terbaik dapat dilihat berdasarkan nilai koefisien determinasi yang besar.

Kata Kunci: Semiparametrik, *Small Area Estimation*, *Penalized Spline*

1. Pendahuluan

1.1. *Kemiskinan*

Kemiskinan yang terjadi di Indonesia merupakan salah satu masalah sosial yang belum sepenuhnya teratasi oleh pemerintah hingga saat ini. Banyak faktor yang

menyebabkan hal tersebut diantaranya terbatasnya dana dan ketidaktepatan program pengentasan kemiskinan. Agar kebijakan tepat sasaran maka untuk mengimplementasikan program pengentasan kemiskinan diperlukan adanya informasi pada suatu daerah. Informasi yang didapatkan melalui survey pada suatu area kecil akan mengakibatkan adanya keterbatasan objek survey. Keterbatasan objek survey menyebabkan data yang di duga dengan pendugaan parameter secara langsung tidak menghasilkan dugaan yang akurat. Untuk menghasilkan pendugan yang lebih baik digunakan metode pendugaan tidak langsung pada area kecil (*Small Area Estimation*).

Small Area Estimation (SAE) merupakan suatu metode statistika untuk menduga parameter pada suatu subpopulasi jika jumlah contohnya berukuran kecil [4]. Metode ini memanfaatkan data dari domain besar untuk menduga variabel yang menjadi perhatian pada domain yang lebih kecil. Pada SAE, ada informasi lain yang diasumsikan dapat dipinjam untuk memperbaiki pendugaan terhadap parameter yang menjadi perhatian. Informasi dapat berupa variabel yang sama pada area lain atau variabel lain pada area yang sama yang dipandang berkaitan dengan parameter yang akan di duga.

Salah satu pendekatan nonparametrik yang digunakan adalah pendekatan semiparametrik *Penalized Spline* yang mempunyai model yang lebih fleksibel karena keberadaan dua komponen dalam model yang mengakomodasi hubungan antara respon dengan prediktor yang bersifat linier dan hubungan antar respon dengan prediktor yang bersifat nonlinier [2].

2. Landasan Teori

2.1. *Regresi Semiparametrik dengan Penalized Spline*

Regresi semiparametrik merupakan gabungan antara regresi parametrik dan regresi nonparametrik. Regresi parametrik merupakan metode yang digunakan untuk mengetahui pola hubungan antara variabel prediktor dengan variabel respon apabila bentuk kurva regresi diketahui, sedangkan regresi nonparametrik merupakan suatu metode statistika yang digunakan untuk mengetahui hubungan antara variabel respon dan prediktor, jika bentuk hubungan antara variabel respon dan prediktor tidak diketahui atau tidak didapatkan informasi sebelumnya [5].

Pada regresi semiparamterik dapat digunakan model *generalized additive*. Misalkan terdapat data berpasangan $(y_i, \mathbf{x}_i, t_i)$, hubungan antara y_i, x_i, t_i diasumsikan mengikuti model regresi semiparametrik seperti di bawah ini :

$$\mathbf{y}_i = \mathbf{X}_i\beta + f(t_i) + \varepsilon_i, i = 1, 2, \cdots, n \tag{2.1}$$

dengan $\mathbf{y}_i$ adalah variabel respon pada pengamatan ke-i, X_i adalah komponen parametrik, $f(t_i)$ adalah fungsi regresi nonparametrik dan ε adalah residual acak yang menyebar normal dengan nilai tengah 0 dan ragam σ^2 [5].

Penalized Spline (*P-spline*) merupakan salah satu model nonparametrik yang merupakan model polinomial terputus (*polynomial truncated*) tersegmen dimana sifat segmen inilah yang memberikan fleksibilitas yang lebih baik dibandingkan model polinomial biasa. Metode ini juga merupakan metode pemulusan yang

menarik karena mempunyai sifat sederhana. Pemodelan *P-spline* memberikan pemilihan *knot* yang fleksibel.

Secara umum fungsi *P-spline* dapat dinyatakan sebagai berikut:

$$\mathbf{y} = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p + \sum_{j=1}^{k} \gamma_j (x_i - K_j)_+^p \tag{2.2}$$

Bila dinyatakan dalam notasi matriks maka sistem persamaan tersebut dapat dinyatakan sebagai

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon, \tag{2.3}$$

dimana $\mathbf{y} = (y_1, \cdots, y_n)$ parameter β merupakan parameter koefisien parametrik dari parameter yang tidak diketahui, $\mathbf{Z}$ merupakan fungsi *truncated* pada *spline*, γ adalah vektor koefisien *P-spline* dan ε adalah galat. Fungsi *truncated* pada *P-spline* didefinisikan sebagai berikut

$$(x_i - K_j)_+^p = \begin{cases} (x_i - K_j)_+^p & untuk\ x_i \geqslant K_j \\ 0, & untuk\ x_i < K_j \end{cases} \tag{2.4}$$

Fungsi *spline* pada model menunjukkan bahwa *spline* merupakan polinomial terputus, tapi masih bersifat kontinu pada *knot-knot*nya. Model *spline* yang terbentuk akan mengalami *overparameterized* sehingga menyebabkan *overfitting*. Untuk menghindari hal tersebut ditambahkan *penalty* pada parameter *spline* [3].

Misalkan $\mathbf{C} = \left[\mathbf{X}, \mathbf{Z}\right]$ dan $\theta = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$ sehingga persamaan 2.3 dapat ditulis

$$\mathbf{y} = \mathbf{C}\theta + \varepsilon. \tag{2.5}$$

Dengan meminimalisasi jumlah kuadrat galat dan menggunakan pengganda Lagrange terhadap kendala $\|\gamma\|^2 \geqslant a$, Persamaan 2.3 dapat dinyatakan sebagai

$$min_{\beta,\gamma,\lambda} \|\mathbf{y} - \mathbf{C}\theta\|^2 + \lambda \theta^T D \theta, \lambda \geq 0 \tag{2.6}$$

dimana matriks D dapat dinyatakan matriks *penalty*.

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{(p\times 1)\times 2} & \mathbf{0}_{(p\times 1)\times k} \\ \mathbf{0}_{k\times(p+1} & \mathbf{I}_{k\times k} \end{bmatrix}$$

Selanjutnya minimumkan jumlah kuadrat galat tersbut dengan menggunakan metode *penalized least spline*, sehingga diperoleh

$$\widehat{\theta} = (\mathbf{C}^T\mathbf{C} + \lambda\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}.$$

Maka untuk $\widehat{y}$ pada Persamaan 2.5 diperoleh

$$\widehat{y} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}. \tag{2.7}$$

Nilai $\widehat{\theta}$ bergantung pada parameter pemulus λ. Jika nilai λ besar akan menghasilkan bentuk kurva regresi yang sangat halus. Sebaliknya, jika nilai λ kecil akan menghasilkan bentuk kurva regresi yang kasar [1].

2.2. *Pemilihan Jumlah Knot (k) Optimal*

Knot dapat diartikan sebagai suatu titik fokus dalam fungsi *spline*. Penentuan jumlah *knot* sangat berpengaruh dalam menentukan titik *knot* pada *P-spline*. Metode *fixed selection method* digunakan untuk menentukan jumlah titik *knot*, dimana jumlah *knot* k dihitung dari [5]

$$K = min\{\frac{1}{4} \times \textit{banyaknya } x_i \textit{ yang berbeda}; 35\}. \tag{2.8}$$

Suatu kriteria yang biasa digunakan dalam pemilihan model *spline* terbaik adalah *Generalized Cross Validation* (GCV). Nilai GCV dipakai karena aspek perhitungannya lebih sederhana dan cukup efisien. GCV didefinisikan sebagai berikut [1]

$$GCV(\lambda) = \frac{MSE(\lambda)}{(n^{-1}tr(\mathbf{I} - \mathbf{S}_\lambda))} \tag{2.9}$$

2.3. *Small Area Estimation*

Small Area Estimation (SAE) merupakan suatu teknik statistika untuk menduga parameter-parameter subpopulasi yang ukuran sampelnya kecil. Metode pendugaan ini memanfaatkan data dari skala yang besar untuk menduga parameter dari skala yang lebih kecil [4].

Dalam SAE terdapat dua jenis model dasar yang digunakan, yaitu model berbasis area dan model berbasis unit.

2.3.1. *Small Area Estimation Berbasis Area*

Pada model SAE berbasis area, data pendukung yang tersedia hanya sampai level area. Model level area menghubungkan penduga langsung area kecil dengan data pendukung dari domain lain untuk setiap area. Model SAE untuk level area sebagai berikut

$$\hat{\theta} = x_i^T\beta + b_i v_i + e_i, i = 1, 2, \cdots, m, \tag{2.10}$$

dengan $\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})^T$,$\beta = (\beta_l, \cdots, \beta_p)$ adalah koefisien regresi berukuran $p\times$ 1 , p merupakan banyak variabel prediktor, b_i adalah konstanta positif yang diketahui, dan v_i adalah pengaruh acak area kecil diasumsikan $v_i \sim iid$ $N(0, \sigma_v^2)$, e_i adalah galat yang diasumsikan $e_i \sim iid$ $N(0, \sigma_\varepsilon^2)$ [4].

2.4. *Small Area Estimation dengan Pendekatan Semiparametrik Penalized Spline*

Model *P-spline* merupakan pengaruh acak yang dapat dikombinasikan dengan model SAE berbasis area untuk mendapatkan estimasi area kecil secara semiparametrik berdasarkan model linier campuran [2]. Bentuk umum dari pendugaan area kecil dengan menggunakan *P-spline* adalah sebagai berikut :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + Du + e, \tag{2.11}$$

dengan $X\beta + Z\gamma$ adalah fungsi *spline* yang memuat komponen parametrik dan nonparametrik, Du adalah pengaruh acak area kecil, Setiap komponen acak diasumsikan independen satu sama lain dengan

$$\gamma \sim \ MVN(\mathbf{0}, \Sigma_\gamma) \text{ dengan } \Sigma_\gamma = \sigma_\gamma^2 I_k$$
$$u \sim \ MVN(\mathbf{0}, \Sigma_u) \text{ dengan } \Sigma_u = \sigma_u^2 I_T$$
$$\varepsilon \ \sim \ MVN(\mathbf{0}, \Sigma_\varepsilon) \ dengan \ \Sigma_\varepsilon = \sigma_\varepsilon^2 I_n$$

Dengan menggunakan fungsi *log likelihood* diperoleh penduga bagi β adalah

$$\widehat{\beta} = (\mathbf{X}^T V^{-1} \mathbf{X})^{-1} (\mathbf{X}^T V^{-1} \mathbf{y}), \tag{2.12}$$

dengan

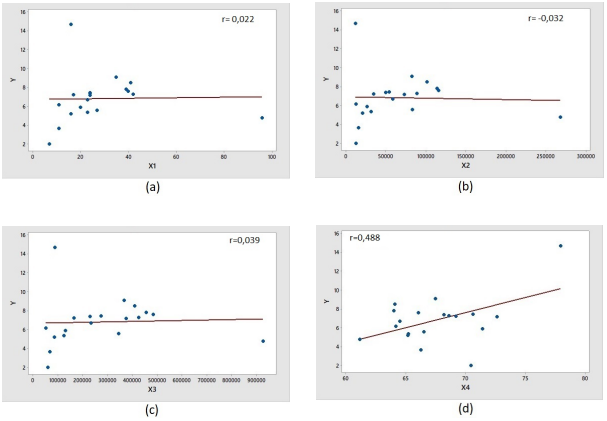
$$\hat{\gamma} = \Sigma_\gamma \mathbf{Z}^T V^{-1} (Y - \mathbf{X}\hat{\beta}), \hat{u} = \Sigma_u \mathbf{D}^T V^{-1} (Y - X\hat{\beta}). \tag{2.13}$$

Untuk area kecil U_T yang diberikan, maka akan dilakukan pendugaan terhadap

$$\bar{y}_t = \bar{x}_t \beta + \bar{z}_t \gamma + u_t. \tag{2.14}$$

3. Pembahasan

Hubungan antara variabel diperjelas oleh diagram pencar yang dibentuk antara variabel respon dengan masing-masing variabel prediktor sebagaimana yang diperlihatkan pada Gambar 1.



Gambar 1. Diagram pencar variabel respon Y dengan (a) jumlah lembaga pendidikan SMA sederajat (X_1), (b) Penduduk Usia Produktif Tidak Bekerja (X_2), (c) Jumlah Penduduk (X_3), dan (d) Tingkat Partisipasi Angkatan Kerja (X_4)

Pembentukan model dilakukan menggunakan metode *Generalized Additive Model*(GAM). *Generalized Additive Model* merupakan salah satu metode yang variabel responnya tidak harus berdistribusi normal dan hubungan dengan variabel prediktor tidak harus linier. Model *generalized Additive Model* yaitu

$$\mathbf{y} = f(\mathbf{X}) + m(t_i) + u, \tag{3.1}$$

dimana $f(\mathbf{X})$ merupakan komponen yang diduga dengan regresi parametrik menggunakan analisis linear berganda dan $m(t_i)$ merupakan komponen yang diduga dengan nonparametrik *P-spline*.

Dari Gambar 1 sebelumnya terlihat bahwa tidak terdapat bentuk hubungan tertentu antara variabel respon dengan variabel prediktor, sehingga pemodelan linear yang biasanya diperoleh dengan analisis regresi linear berganda tidak memberikan hasil yang baik sebagai alternatif. Hubungan tersebut dapat diatasi dengan menggunakan metode semiparametrik dengan pendekatan *P-spline*.

3.1. *Model 1*

Model semiparametrik dengan menggunakan metode *Generalized Additive Model* untuk model ini adalah

$$y = f(X_1, X_2, X_3) + f(X_4) + u. \tag{3.2}$$

Berikut merupakan nilai GCV yang didapatkan berdasarkan *spline* orde linear,kuadratik dan kubik.

Tabel 1. Nilai GCV Model 1

Banyak Knot	Linear	Kuadratik	Kubik
1	3,783535	3,969057	4,122396
2	3,38885	3,9998	4,122396
3	3,312154	3,536778	3,734098
4	2,872636	3,566566	3,859188

Dari Tabel 1 disimpulkan bahwa orde yang memiliki nilai GCV minimum adalah model *penalized spline* orde linear dengan jumlah titik *knot* sebanyak empat yaitu 64,03188;66,31688;67,50007 dan 70,43737, sehingga model ini yang digunakan untuk menduga persentase penduduk miskin.

3.2. *Model 2*

Model *Generalized Additive Model* yang terbentuk pada model ini adalah

$$y = f(X_1, X_2, X_4) + f(X_3) + u. \tag{3.3}$$

Dengan menggunakan langkah yang sama seperti langkah yang dilakukan pada model 1 diperoleh hasil sebagai berikut

Tabel 2. Nilai GCV Model 2

Banyak Knot	Linear	Kuadratik	Kubik
1	6,589564	7,6112288	8,630431
2	5,971366	6,953526	11,01113
3	2,666479	4,928749	6,488725
4	1,798299	3,714266	13,68431

Dari Tabel 2 dapat disimpulkan bahwa model yang memiliki nilai GCV minimum adalah model *P-spline* orde linear dengan jumlah titik *knot* adalah empat yaitu 85416,64; 88700,72; 131825,7 dan 410870,69. Setelah di dapatkan model terbaik berdasarkan GCV minimum, kemudian diduga nilai estimasi dari model 2 sehingga didapatkan pendugaan persentase penduduk miskin.

3.3. *Model 3*

Model *generalized Additive Model* pada model ini sebagai berikut

$$y = f(X_1, X_3, X_4) + f(X_2) + u. \tag{3.4}$$

Untuk masing-masing orde polinomial *spline*, diperoleh hasil sebagai berikut

Tabel 3. Nilai GCV Model 3

Banyak Knot	**Linear**	**Kuadratik**	**Kubik**
1	3,689326	1,828304	10,34255
2	1,76522	10,43231	10,76834
3	1,805813	12,1226	10,79565
4	2,099744	14,177768	16,79035

Berdasarkan Tabel 3 dapat disimpulkan bahwa model yang memiliki nilai GCV minimum adalah model *P-spline* orde linear dengan jumlah titik *knot* adalah sebanyak dua bernilai 13030,39 dan 44342,78. Setelah didapatkan titik *knot* dan model dari semiparametrik *P-spline* maka akan diestimasi pengaruh tetap dan pengaruh acaknya, sehingga didapatkan model menggunakan orde spline dan knot terbaik untuk menduga pesentase penduduk miskin.

3.4. *Model 4*

Model *Generalized Additive* pada model 4 sebagai berikut

$$y = f(X_2, X_3, X_4) + f(X_1) + u. \tag{3.5}$$

dengan asumsi X_2, X_3, X_4 sebagai parametrik dan X_1 diasumsikan nonparametrik. Nilai GCV untuk setiap model tersaji pada tabel berikut

Tabel 4. Nilai GCV Model 4

Banyak Knot	**Linear**	**Kuadratik**	**Kubik**
1	5,667391	6,385789	6,895403
2	5,437353	6,347444	7,043421
3	5,266518	6,240855	6,90462

Dengan langkah yang sama pada model sebelumnya didapatkan model dengan nilai estimasi berdasarkan GCV minimum dan knot yang diperoleh.

3.5. *Pemilihan Model Terbaik*

Pemilihan model terbaik ditentukan menggunakan koefisien determinasi.

Tabel 5. Koefisien Determinasi untuk 5 model terbaik

Model	R^2
Penduga *P-spline* Model 1	0,888
Penduga *P-spline* Model 2	0,593
Penduga *P-spline* Model 3	0,401
Penduga *P-spline* Model 4	0,771

Berdasarkan tabel diketahui bahwa model terbaik untuk kasus model persentase penduduk miskin di Provinsi Sumatera Barat yang baik adalah model 1 yang merupakan model yang diperoleh menggunakan *small area estimation* pendekatan semiparametrik *P-spline* dimana variabel X_1, X_2, X_3 dimodelkan secara parametrik sedangkan X_4 dimodelkan secara nonparametrik seperti yang disajikan pada tabel. Nilai estimasi model yang diperoleh berdasarkan model 1 adalah

$$\begin{aligned}\widehat{y} = {} & 4,689880 + 0,06355620X_1 - 0,0002455831X_2 + \\ & 0,00006185732X_3 - 0,6499306X_4 + \gamma_1(X_4 - 64,03188)_+^1 + \\ & \gamma_2(X_4 - 66,31688)_+^1 + \gamma_3(X_4 - 67,50007)_+^1 + \\ & \gamma_4(X_4 - 70,43737)_+^1 + u.\end{aligned} \tag{3.6}$$

dengan pendugaan persentase penduduk miskin dengan Model 1 adalah

Tabel 6. Pendugaan Persentase Penduduk Miskin di Provinsi Sumatera Barat dengan Model 1

No	**Kabupaten dan Kota**	**Persentase Penduduk Miskin**
1	Kab. Kepulauan Mentawai	9,20
2	Kab. Pesisir Selatan	6,96
3	Kab. Solok	8,41
4	Kab. Sijunjung	7,39
5	Kab.Tanah Datar	6,40
6	Kab. Padang Pariaman	7,17
7	Kab. Agam	7,61
8	Kab. Lima Puluh Kota	8,05
9	Kab. Pasaman	8,29
10	Kabupaten Solok Selatan	7,08
11	Kab. Dharmasraya	6,00
12	Kab. Pasaman Barat	8,92

No	Kabupaten dan Kota	Persentase Penduduk Miskin
13	Kota Padang	5,19
14	Kota Solok	4,80
15	Kota Sawahlunto	5,21
16	Kota Padang Panjang	5,14
17	Kota Bukittinggi	5,37
18	Kota Payakumbuh	6,94
19	Kota Pariaman	5,15

4. Kesimpulan

Berdasarkan hasil pengolahan data yang sudah dibahas pada bab pembahasan dapat diduga persentase penduduk miskin dengan menggunakan semiparametrik *P-spline* di Provinsi Sumatera Barat tahun 2017 berdasarkan model 1. Rata-rata persentase penduduk miskin di Provinsi Sumatera Barat adalah sebesar 6,805%, dimana persentase penduduk miskin terbesar terdapat pada Kabupaten Kepuluan Mentawai sebesar 9.20% dan persentase penduduk miskin terkecil adalah Kota Solok sebesar 4.80%. Sekitar 75% kabupaten dan kota di Sumatera Barat memiliki rata-rata persentase penduduk miskin sebesar 8,407%.

Daftar Pustaka

[1] Apriani, F., 2017, *Pemodelan Pengeluaran Perkapita Menggunakan Small Area Estimation dengan Pendekatan Semiparametrik Penalized Spline*, Thesis di Institut Teknologi Sepuluh Nopember, Tidak Diterbitkan

[2] Idhia, S., Sunandi, E., Rafflesia., U, 2017, Pemodelan Kemiskinan di Provinsi Bengkulu Menggunakan *Small Area Estimation* dengan Pendekatan Semiparametrik *Penalized Spline Jurnal MIPA*, **40** : 134-140

[3] Opsomer, DJ. dkk, 2008, Non-parametric Small Area Estimation Using Penalized Spline Regression, *Royal Statistical Society Journal*, : 4127-4129

[4] Rao JNK. 2003. *Small Area Estimation.* Wiley, London

[5] Ruppert, D, Wand, M.P, dan Carol, R.J. 2003. *Semiparametric Regression.* Cambridge University Press, New York