REGRESI KUANTIL SPLINE UNTUK PEMODELAN NILAI EKSTREM PADA PENCEMAR UDARA PM_{10} DI KOTA SURABAYA

Oleh : Anik Djuraidah[1)] dan La Ode Abdul Rahman[2)]

ABSTRAK

Regresi kuantil merupakan perluasan dari regresi median (pada kuantil 0,5) pada berbagai nilai kuantil. Metode ini dapat digunakan mengukur efek peubah penjelas tidak hanya di pusat sebaran data, tetapi juga pada bagian atas atau bawah ekor sebaran. Analisis ini sangat berguna dalam penerapan, khususnya bila nilai ekstrim merupakan permasalahan penting. Pada penelitian dikembangan model regresi kuantil dengan hubungan fungsional spline. Pendugaan model dikembangkan dari metode regresi median yaitu dengan metode metode simpleks, sedangkan untuk pendugaan selang kepercayaan dan pengujian hipotesis digunakan metode bootsrap. Data yang digunakan penelitian ini adalah data konsentrasi PM_{10} (μg/m3) berasal dari jaringan pemantau kualitas udara ambien kota Surabaya pada bulan Mei 2002 sampai Agustus 2002. Pemodelan diawali dengan penentuan jumlah simpul dan derajat spline, kemudian dilanjutkan dengan model regresi kuantil, dan seleksi model. Hasil penelitian menunjukkan pola hubungan fungsi tujuan dengan kuantil memiliki bentuk kuadratik, dengan nilai maksimum terjadi pada kuantil 0.6. Bentuk hubungan ini sama untuk jumlah simpul 6, 12, 24, dan 36 dan derajat spline 1, 2, dan 3. Model regresi kuantil spline terbaik mempunyai jumlah simpul 24 dengan derajat polinomial 2.

Kata kunci : regresi kuantil, spline, nilai ekstrem, simpleks, bootstrap

Pendahuluan

Pendekatan standar penentuan model regresi linear dan pendugaan parameternya adalah metode kuadrat terkecil (OLS) atau simpangan mutlak terkecil (*least absolute deviation*, selanjutnya disingkat LAD). Pendugaan parameter pada metode OLS diperoleh dengan minimisasi jumlah kuadrat galat, sedangkan pada metode LAD dengan minimisasi jumlah absolut galat. Penduga dari metode OLS adalah rataan fungsi sebaran bersyarat peubah respon, sedangkan dari metode LAD adalah median fungsi bersyarat peubah respon. Meskipun rataan dan median adalah dua ukuran pemusatan yang penting dari suatu sebaran, kedua statistik ini tidak menjelaskan tentang perilaku pada ekor (*tail*) suatu sebaran. Sehingga bila ingin mengetahui perilaku sebaran bersyarat, kurang memuaskan bila hanya mengamati perilaku rataan atau median saja.

Regresi kuantil dikemukakan oleh Koenker dan Bassett pada tahun 1978 (Buhai, 2004), merupakan perluasan model regresi pada kuantil bersyarat peubah respon. Pendekatan ini memungkinkan menduga fungsi kuantil dari sebaran bersyarat respon pada berbagai nilai kuantil yang diinginkan. Setiap kuantil mencirikan titik tertentu (pusat atau ekor) dari sebaran bersyarat. Kombinasi berbagai nilai kuantil akan menghasilkan deskripsi lengkap tentang sebaran bersyarat. Analisis ini sangat berguna untuk sebaran bersyarat yang tak simetrik, padat di ekor sebarannya, atau sebarannya terpotong.

[1)] Dosen Departemen Statistika FMIPA- IPB (e-mail : anikdjuraidah@gmail.com)
[2)] Dosen Departemen Statistika FMIPA- IPB

Keuntungan utama dari regresi kuantil dibandingkan regresi OLS adalah fleksibilitas dalam pemodelan data dengan sebaran bersyarat yang heterogen. Metode ini dapat digunakan mengukur efek peubah penjelas tidak hanya di pusat sebaran data, tetapi juga pada bagian atas atau bawah ekor sebaran. Hal ini sangat berguna dalam penerapan, khususnya bila nilai ekstrim merupakan permasalahan penting, seperti penelitian tentang curah hujan di Hwange National Park, Zimbabwe (Chamaille´-Jammesa *et al*, 2007), kecepatan angin pada siklon tropis di USA (Jagger dan Elsner, 2008)

Hubungan fungsional antara respon dengan peubah penjelas pada regresi kuantil merupakan fungsi linear, yaitu

$$\mathrm{Q}(\tau|\mathrm{X}=x)=\mathbf{x}'\boldsymbol{\beta}(\tau) \text{ dengan } 0<\tau<1 \qquad (1)$$

Bentuk hubungan fungsional antara kuantil bersyarat respon dengan peubah penjelas pada persamaan (1) dapat dikembangkan ke bentuk nonparametrik, sehingga model regresi kuantil nonparametrik adalah

$$\mathrm{Q}(\tau|\mathrm{X}=x)=s(x,\tau)$$

Salah satu bentuk hubungan fungsional nonparametrik adalah spline. Spline merupakan potongan polinomial yang kontinu, sehingga dapat menggambarkan karakteristik lokal pada data (Eubank,1988).

Pemodelan nilai ekstrem pada konsentrasi PM_{10} sangat penting ditinjau dari dampaknya terhadap kesehatan masyarakat. Berdasarkan hasil penelitian Djuraidah dan Aunuddin (2006) diketahui bentuk hubungan fungsional yang terbaik antara PM_{10} dengan waktu adalah spline. Permasalahan utama yang dikaji pada penelitian ini adalah pemodelan pencemar udara dominan PM_{10} di kota Surabaya dengan regresi kuantil spline, sehinga diperoleh gambaran lengkap tentang konsentrasi PM_{10} terutama pada nilai ekstrem.

Regresi Kuantil

Regresi kuantil merupakan generalisasi konsep kuantil pada peubah tunggal ke kuantil bersyarat untuk satu atau lebih peubah penjelas. Untuk peubah acak Y dengan fungsi sebaran peluang

$$\mathrm{F}(y)=\mathrm{P}(\mathrm{Y}\le y)$$

Kuantil ke–τ dari Y didefinisikan sebagai fungsi invers

$$\mathrm{Q}(\tau)=\inf\{y,\ \mathrm{F}(y)\ge\tau\}$$

dengan $\tau\in(0,1)$, sebagai contoh median adalah Q(0.5).

Untuk contoh acak berukuran n dari peubah acak Y, yaitu $(y_1,\cdots,y_n)$, median contoh adalah penduga yang meminimumkan jumlah mutlak galat yaitu

$$\min_{\xi\in R} \sum_{i=1}^{n} |y_i - \xi|$$

Seperti halnya median contoh, metode ini bisa dikembangkan untuk model regresi kuantil

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dengan $\mathbf{y} = (y_1,\cdots,y_n)'$ adalah vektor respon berukuran $(n\times 1)$, $\mathbf{X} = (\mathbf{x}_1,\cdots,\mathbf{x}_n)'$ adalah matriks prediktor berukuran $(n\times p)$, $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_p)'$, adalah vektor parameter berukuran $(p\times 1)$, dan $\boldsymbol{\varepsilon} = (\varepsilon_1,\cdots,\varepsilon_n)'$, adalah vektor galat berukuran $(n\times 1)$. Regresi L_1 disebut juga dengan regresi median yang merupakan perluasan dari median contoh. Penduga koefisien pada model regresi L_1 , $\hat{\boldsymbol{\beta}}_{\text{LAR}}$, disebut juga dengan penduga norma L_1, yang merupakan solusi dari minimisasi fungsi

$$\min_{\boldsymbol{\beta}\in R^p} \sum_{i=1}^{n} \left| y_i - \mathbf{x}_i'\boldsymbol{\beta} \right|$$

Secara umum menurut Koenker (2005) penduga regresi kuantil ke-τ untuk $\tau\in(0,1)$ merupakan solusi dari masalah minimisasi fungsi

$$\min_{\boldsymbol{\beta}\in R^p} \left[\sum_{i\in\{i:y_i\geq \mathbf{x}_i'\boldsymbol{\beta}\}}^{n} \tau \left| y_i - \mathbf{x}_i'\boldsymbol{\beta} \right| + \sum_{i\in\{i:y_i< \mathbf{x}_i'\boldsymbol{\beta}\}}^{n} (1-\tau) \left| y_i - \mathbf{x}_i'\boldsymbol{\beta} \right| \right] \qquad (2)$$

Solusi persamaan (2) dinotasikan $\hat{\boldsymbol{\beta}}(\tau)$, dan penduga norma L_1 adalah $\hat{\boldsymbol{\beta}}_{\text{LAR}} = \hat{\boldsymbol{\beta}}(0{,}5)$.

Regresi kuantil pada kuantil ke-τ merupakan perluasan kuantil ke-τ dari contoh acak, $\xi(\tau)$, yang diformulasikan sebagai solusi dari

$$\min_{\boldsymbol{\beta}\in R} \left[\sum_{i\in\{i:y_i\geq \xi\}}^{n} \tau \left| y_i - \xi \right| + \sum_{i\in\{i:y_i< \xi\}}^{n} (1-\tau) \left| y_i - \xi \right| \right] \text{ untuk } \tau\in(0,1)$$

Permasalahan optimasi pada regresi median diformulasikan dan dipecahkan dengan program linear sejak tahun 1950 dan variasi dari algoritma simpleks dikembangkan oleh Barrodale dan Roberts (1973). Beberapa metode lain yang digunakan untuk pemecahan regresi L_1 seperti algoritma titik interior (Karmarkar, 1984), dan metode pemulusan (Madsen & Nielsen, 1993).

Metode perhitungan selang kepercayaan untuk parameter regresi kuantil $\boldsymbol{\beta}(\tau)$ adalah pendugaan langsung (*sparsity*), pangkat (*rank*), dan *resampling*. Metode pendugaan

langsung merupakan metode sangat cepat, tetapi memerlukan pendugaan fungsi *sparsity*, yang tidak kekar untuk data dengan sebaran yang tidak bebas dan tidak identik. Metode pangkat menghitung selang kepercayaan dengan membalik uji skor pangkat, relatif tidak sulit hanya menggunakan algoritma simpleks. Pada metode resampling menggunakan bootsrap yang tidak stabil untuk data berukuran kecil.

Untuk pengujian hipotesis $\mathrm{H}_0 : \beta_2 = 0$, dengan β_2 menyatakan satu himpunan bagian parameter, dimana vektor parameter dibagi dalam $\beta'(\tau) = \left(\beta_1(\tau),\ \beta_2(\tau)\right)$ dan matriks ragam-peragam Ω untuk pendugaan parameter dibagi sebagai Ω_{ij} dengan $i = 1,\ 2\ ;\ j = 1,\ 2$ dan $\Omega_{22}^{-1} = \left(\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}^{-1}\right)^{-1}$. Statistik uji Wald adalah $T_W(\tau) = \hat{\beta}_2'(\tau)\ \hat{\Sigma}(\tau)^{-1}\hat{\beta}_2(\tau)$ dengan $\hat{\Sigma}(\tau)$ adalah penduga ragam-peragam $\beta_2(\tau)$. Penduga ragam $\hat{\Sigma}(\tau)$ dapat diperoleh dari pendekatan ragam peragam asimtotik dan bootstrap marjinal rantai Markov (*Markov chain marginal bootstrap* selanjutnya disingkat MCMB) yang dikembangkan oleh He & Hu (2002). Pada pendekatan asimtotik, penduga ragam peragamnya adalah $\hat{\Sigma}(\tau) = \frac{1}{n}\hat{\omega}(\tau)^2\Omega_{22}^{-1}$ dengan $\omega(\tau) = \sqrt{\tau(1-\tau)}\ \hat{s}(\tau)$ dan $\hat{s}(\tau)$ adalah fungsi sparsity. Penduga dari bootsrap adalah ragam-peragam empirik contoh MCMB.

Uji nisbah kemungkinan didasarkan pada beda antara nilai fungsi tujuan model yaitu $\mathrm{D}_0(\tau) = \sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_{1i}'\ \hat{\boldsymbol{\beta}}(\tau)\right)$ dan $\mathrm{D}_1(\tau) = \sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_{1i}'\ \hat{\boldsymbol{\beta}}_1(\tau)\right)$

dan misalkan $T_{LR}(\tau) = 2\left(\tau(1-\tau)\hat{s}(\tau)\right)^{-1}\ \left(D_1(\tau) - D_0(\tau)\right)$ dengan $\hat{s}(\tau)$ adalah fungsi sparsity. Koenker & Machado (1999) telah membuktikan bahwa kedua uji secara asimtotik ekivalen dan sebaran dari statistik uji jika H_0 benar akan konvergen ke χ_q^2 dengan q adalah dimensi dari β_2 .

Regresi Spline

Misalkan $(x_\mathrm{i}, y_\mathrm{i})$ adalah pengukuran pada peubah penjelas *x* dan peubah respon *y* untuk $1 \le \mathrm{i} \le \mathrm{n}$. Misalkan hubungan fungsional antara *x* dengan *y* dimodelkan sebagai

$$y_\mathrm{i} = s(x_\mathrm{i}) + \varepsilon_\mathrm{i} \qquad (3)$$

dengan *s* adalah fungsi mulus, ε_i bebas stokastik dengan ragam σ^2. Model (3) adalah bentuk regresi nonparametrik yang paling sederhana dan banyak metode pendekatan yang dapat digunakan seperti yang dibahas oleh Eubank (1988), Green dan Silverman (1994), dan Simonoff (1996). Misalkan fungsi mulus *s* diduga dengan model regresi spline yaitu :

$$s(\mathbf{x};\boldsymbol{\beta}) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} u_{pk}(x - \kappa_k)_+^p$$

dengan $\boldsymbol{\beta} = (\beta_0, ..., \beta_p, u_{p1}, ..., u_{pK})'$ adalah vektor koefisien regresi spline, $p \geq 1$ adalah bilangan bulat positif, $(w)_+^p = w^p \mathbf{I}(w \geq 0)$ adalah fungsi polinomial terpotong (*truncated polynomial function*, selanjutnya disingkat FPT), dan $\kappa_1 < ... < \kappa_K$ adalah simpul tetap. Untuk peubah tunggal, selain basis FPT dapat juga digunakan basis natural kubik spline, atau basis B-spline, sedangkan untuk peubah ganda umumnya digunakan fungsi basis radial. Pendugaan koefisien pada regresi spline menggunakan metode OLS

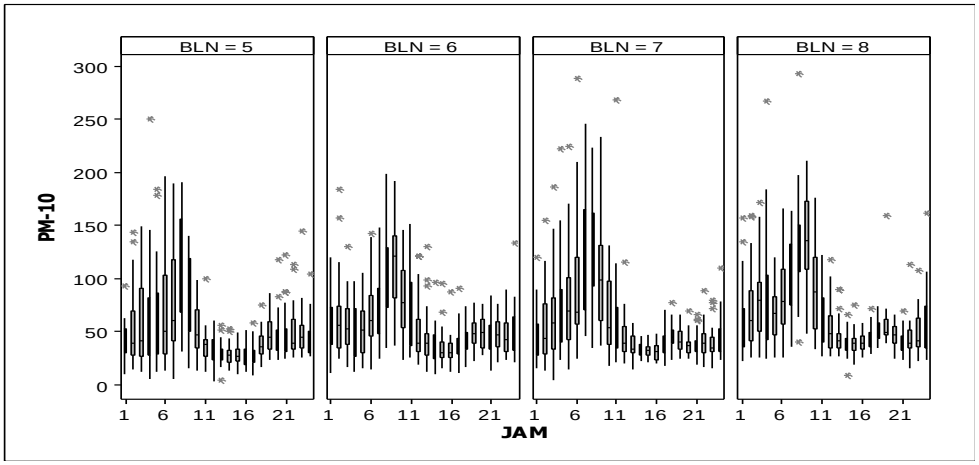
Data dan Metode Penelitian

Data yang digunakan penelitian ini adalah data konsentrasi PM_{10} ($\mu g/m^3$) di kota Surabaya pada bulan Mei 2002 sampai Agustus 2002 di SUF-1 yang terletak di halaman taman prestasi (Jl. Ketabang Kali). Lokasi ini mewakili daerah pusat kota, pemukiman, perkantoran, dan perdagangan Surabaya Pusat.

Tahap pertama pada pemodelan kuantil spline adalah menentukan hubungan fungsional spline terbaik antara waktu (jam) dengan peubah respon. Derajat dari basis FPT yang digunakan adalah 1, 2, dan 3 dengan jumlah simpul *nk* = 6, 12, 24, 36. Selanjutnya pada tahap kedua adalah pemodelan regresi kuantil dengan hubungan fungsional spline pada beberapa nilai kuantil yaitu τ = 0.50, 0.75, 0.90, dan 0.95. Regresi kuantil pada τ=0.50 untuk menggambarkan model di pusat data, pada kuantil τ=0.75 untuk menggambarkan model di kuartil ketiga, dan pada kuantil τ=0.90, 0,95 untuk menggambarkan model pada nilai ekstrem. Pada tahap ketiga dilakukan seleksi model berdasarkan nilai fungsi tujuan model. Tahap terakhir adalah prediksi model terbaik dan penentuan selang kepercayaan koefisien regresi kuantil.

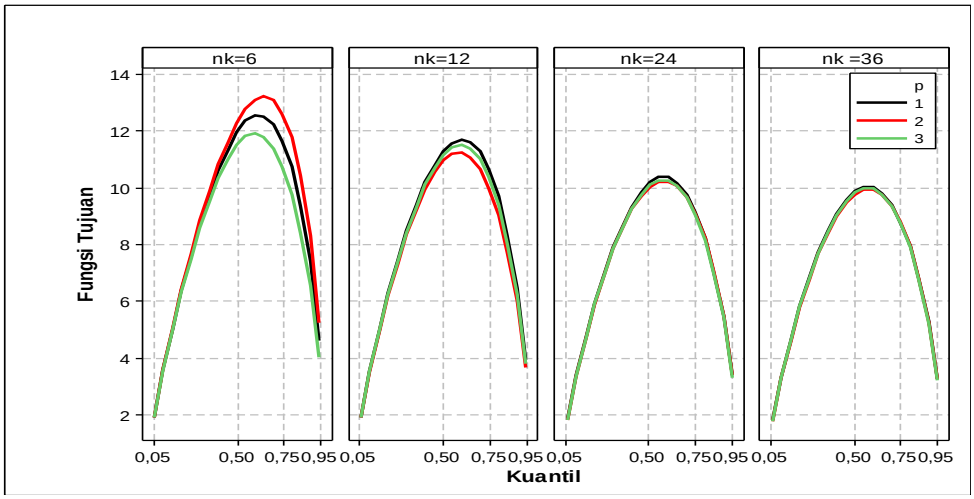
Hasil dan Pembahasan

Diagram kotak-garis konsentrasi PM_{10} bulan Mei sampai Agustus 2002 disajikan pada Gambar 1. Pada umumnya pola sebaran data tidak simetrik dengan banyak pencilan di nilai besar. Pada Gambar 1 tampak lebar kotak kuartil antar jam tidak sama, hal ini menunjukkan keragaman data antar jam tidak homogen. Secara umum pola kecenderungan konsentrasi PM_{10} dengan jam pada setiap bulan tampak mirip, yaitu mempunyai 2 puncak yang terdapat pada jam 8 dan jam 18. Kedua waktu puncak tersebut terutama disebabkan oleh rutinitas transportasi yang berhubungan dengan waktu berangkat kerja dan pulang kerja.



Gambar 1. Pola Konsentrasi PM_{10} pada Bulan Mei 2002 sampai Agustus 2002

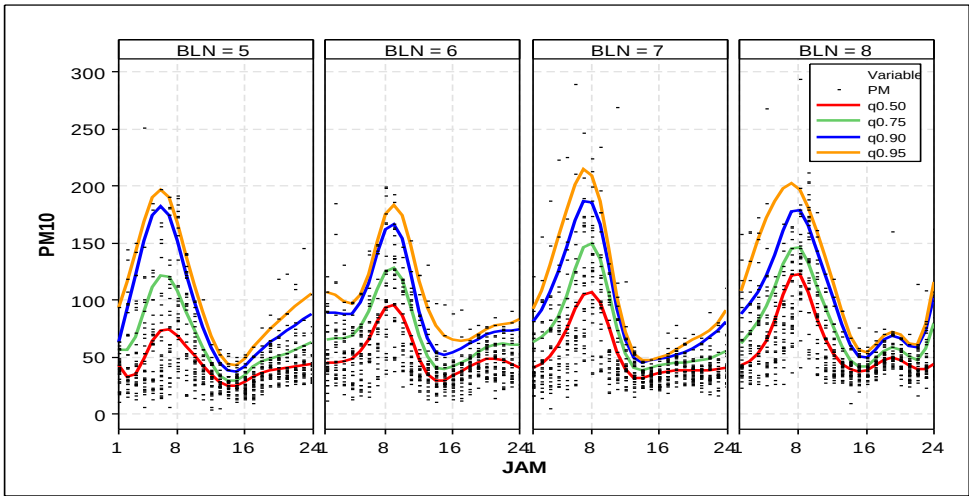
Fungsi tujuan pada regresi kuantil yang dinyatakan pada persamaan (2) merupakan jumlah simpangan mutlak galat, ekivalen dengan jumlah kuadrat galat pada regresi dengan metode OLS. Nilai fungsi tujuan pada kuantil 0,05 sampai 0,95 untuk jumlah simpul spline 6, 12, 24, dan 36 dengan derajat polinomial 1, 2, dan 3 disajikan pada Gambar 2. Fungsi tujuan cenderung menurun dengan semakin banyak jumlah simpul spline. Sedangkan pengaruh derajat polinomial pada fungsi tujuan tidak terlalu besar kecuali pada model dengan jumlah simpul 6 dan 12 antara kuantil 0,5 sampai 0,8. Nilai fungsi tujuan tertinggi terjadi pada kuantil 0,6. Hal ini menunjukkan pada kuantil tersebut keragaman terbesar. Pada kuantil 0,5 dan 0,95 tampak nilai fungsi tujuannya tidak sama, hal ini berarti disamping keragamannya yang tidak sama, juga bentuk sebaran data tidak simetrik.



Gambar 2. Nilai Fungsi Tujuan pada Kuantil 0,05 sampai 0,95 untuk jumlah simpul spline 6, 12, 24,36 dengan derajat polinomial 1,2, dan 3.

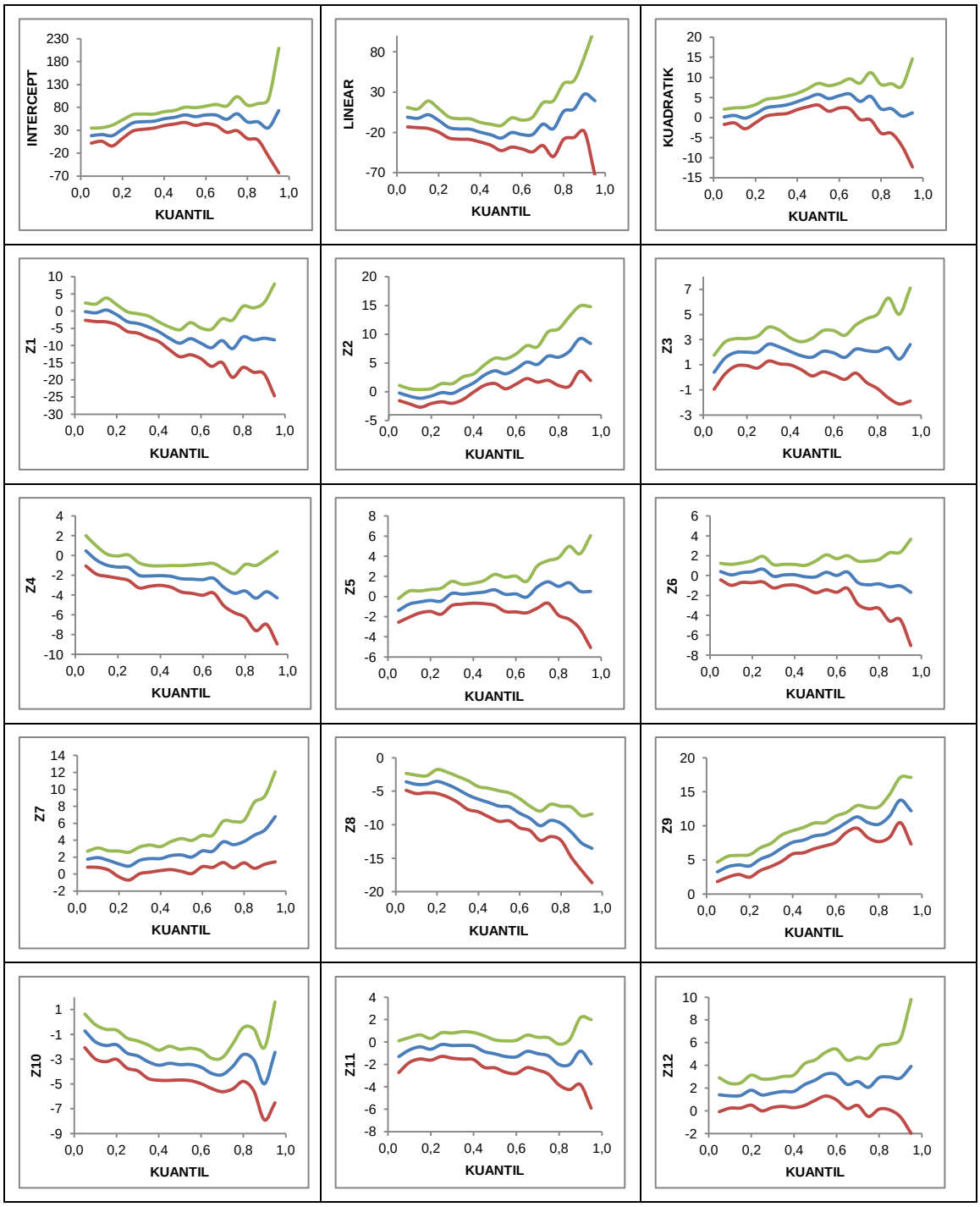
Perbedaan nilai fungsi tujuan pada jumlah simpul 24 ke 36 tidak terlalu besar. Sedangkan perbedaan fungsi tujuan pada 3 macam derajat polinomial pada simpul 24 tidak nyata, meskipun nilai fungsi tujuan terendah terjadi pada polinimial derajat 2. Dengan demikian model terbaik adalah pada jumlah simpul 24 dengan derajat polinomial 2.

Kurva prediksi model regresi kuantil spline terbaik pada kuantil 0,5, 0,75, 0,90, dan 0,95 disajikan pada Gambar 3. Pada Gambar tampak model regresi kuantil spline mendeskripsikan dengan baik data respon pada berbagai nilai kuantil. Lengkungan kurva model tampak mengikuti pola data dengan baik. Kurva model tidak bersifat paralel, pada jam 8, yang merupakan nilai tertinggi PM_{10}, tampak perbedaan koefisien model antar kuantil paling besar. Beberapa nilai pencilan masih tampak diluar kurva regresi kuantil spline pada kuantil 0,9. Hal ini disebabkan pencilan-pencilan tersebut tidak mengikuti pola kurva regresi kuantil yang diduga.



Gambar 3. Model Regresi Kuantil Spline pada kuantil 0,5, 0,75, 0,90, dan 0,95

Selang kepercayaan koefisien regresi kuantil spline diduga dengan metode bootstrap. Gambar selang kepercayaan untuk koefisien intersep, linear, kuadratik, dan basis spline Z1-Z12 disajikan pada Gambar 4. Pada umumnya keragaman koefisien pada kuantil 0,05 sampai 0,60 hampir sama, sedangkan pada kuantil lebih besar 0,6 keragaman koefisien cenderung meningkat searah dengan meningkatnya nilai kuantil. Hal ini disebabkan bentuk sebaran respon dengan kemiringan positip (menjulur ke nilai besar). Koefisien-koefisien model pada berbagai nilai kuantil tampak tidak ada yang sama dan pada umumnya mempunyai tren meningkat atau menurun.



Gambar 4. Selang Kepercayaan 95 % bagi Koefisien Regresi Kuantil Spline

Kesimpulan

Regresi kuantil spline bersifat fleksibel dan menghasilkan model yang berbasis data. Fungsi tujuan menurun dengan bertambahnya jumlah simpul spline, sedangkan derajat spline tidak berpengaruh terhadap fungsi tujuan. Keragaman koefisien regresi kuantil spline pada nilai ekstrem lebih besar dibandingkan pada nilai lainnya.

Daftar Pustaka

Barrodale I, Roberts FDK. 1973. An Improved Algorithm for Discrete L1 Linear Approximation. *SIAM Journal of Numerical Analysis* 10:839–848.

Buhai I S, 2004. Quantile Regression : Overview and Selected Applications. Roger Koenker's lecture notes from the recent Netherlands Network of Economics Workshop in Groningen, December 2003.

Chamaille´-Jammesa, S, Fritz H, Murindagomo F, 2007. Detecting climate changes of concern in highly variable environments : Quantile regressions reveal that droughts worsen in Hwange National Park, Zimbabwe. *Journal of Arid Environments* 71 :21–326

Djuraidah A, Aunuddin. 2006. Pendugaan Regresi Spline Terpenalti dengan Pendekatan Model Linear Campuran. *Statistika* Jurnal Statistika FMIPA-UNISBA 6(1): 39-46.

Eubank RL. 1988. *Spline Smoothing and Nonparametric Regression.* New York : Marcel Deker.

Green PJ, Silverman BW. 1994. *Nonparametric Regression and Generalized Linear Models : a Roughness Penalty Approach*. London: Chapman & Hall.

He X, Hu F. 2002. Markov Chain Marginal Bootstrap. Journal of the American Statistical Association 97:783–795.

Jagger TH, Elsner JB. 2008. Modeling tropical cyclone intensity with quantile regression. *Int. J. Climatol.* Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/joc.1804

Karmarkar, N. (1984), "A New Polynomial-time Algorithm for Linear Programming," Combinatorica, 4, 373–395.

Koenker R. 2005. *Quantile Regression*. Cambridge : Cambridge University Press.

Koenker R, Machado AF. 1999. Goodness of Fit and Related Inference Processes for Quantile Regression. JASA 94: 1296–1310.

Madsen, K. and Nielsen, H. B. (1993), "A Finite Smoothing Algorithm for Linear L1 Estimation," SIAM Journal on Optimization, 3, 223–235.

Simonoff JS. 1996. *Smoothing Methods in Statistics*. New York : Springer-Verlag