ESTIMASI MODEL UNTUK DATA DEPENDEN DENGAN METODE *CROSS VALIDATION*

Oleh: Tarno
Program Studi Statistika FMIPA UNDIP Semarang

Abstract

This paper discuss about application of cross-validation method for modeling of dependent data. One of the data that categorized into dependent data is a time series. To construct the mathematical model for a time series data, we must have at least 50 series. In practices we often have some problem as long as we collect the time series data. So we don't get ideal data related to number of sample. To solve this problem, we can generate observation data. There are several methods that can be used to generate data such as cross-validation and bootstrap. Application of cross-validation method to generate time series data can't be done randomly, but we must generate the data based on balanced incomplete block design. The basic principle of cross-validation method is the data divided into two parts those are construction data and validation data. Construction data are drawn from observation data based on moving block and then we construct the model with Box-Jenkins method and verify the model with validation data. Do this process for different blocks as replication samples of cross-validation method, such that we can construct the best model that minimized loss function for prediction errors.

Key words: time series data, estimate model, cross-validation

1. PENDAHULUAN

Model runtun waktu dibangun memiliki alasan utama yaitu untuk prediksi beberapa waktu ke depan yang mempunyai nilai strategis dan ekonomis. Untuk mendapatkan nilai prediksi yang akurat diperlukan sejumlah data historis masa lampau yang cukup panjang. Panjang runtun waktu yang ideal untuk membuat model prediksi yang akurat dibutuhkan minimal 50 deret observasi[2]. Namun dalam prakteknya, tidak jarang ditemui kendala untuk mendapatkan sederetan data observasi yang panjangnya ideal. Untuk mengatasi hal itu, tidak perlu dilakukan pengumpulan data ke lapangan karena banyak membutuhkan tenaga, waktu dan biaya, tetapi cukup dilakukan dengan cara pembangkitan data observasi secara acak dengan bantuan komputer. Salah satu metode yang dapat digunakan untuk membangkitakan data tersebut adalah metode validasi-silang[1],[2].

Metode validasi-silang biasanya banyak diterapkan pada data independent dan berdistribusi identik. Namun dalam praktek, data yang dikumpulkan tidak selalu memiliki sifat independent dan berdistribusi identik. Hal inilah yang mendorong perlunya dilakukan kajian tentang penerapan metode validasi-silang pada data runtun waktu atau data dependent yang lain. Suatu runtun waktu merupakan barisan observasi yang diindekkan dengan waktu dan biasanya berkorelasi. Data dependent lainnya termasuk m-dependent data, Markov chains, serta proses stokhastik stasioner lainnya tidak diindekkan dengan waktu[1].

Metode validasi-silang telah banyak diterapkan pada data independent dan berdistribusi identik, namun memiliki keterbatasan dalam hal penerapannya pada masalah data dependent. Secara umum, penerapan metode validasi-silang untuk data dependent, seringkali gagal untuk menangkap struktur ketergantungan data tersebut dan diperlukan adanya modifikasi nontrivial dalam hal menghasilkan estimator variansi yang valid dan prosedur inferensi lainnya[1].

Kendala yang dihadapi pada saat penerapan metode validasi-silang untuk pemodelan data runtun waktu adalah dalam hal proses pembangkitan data untuk konstruksi model, karena prinsip dasar pembangkitan data dengan validasi-silang didasarkan pada sampel acak sederhana. Sedangkan pembangkitan data untuk pemodelan runtun waktu harus berdasarkan prosedur rancangan acak blok tak lengkap berimbang[1].

Untuk menyusun model dari sekumpulan data observasi dengan metode validasi-silang, prosedur utama yang harus dilakukan adalah mengelompokkan data observasi menjadi dua bagian yaitu: data kontruksi dan data validasi. Apabila data observasi berukuran n, dan data kontruksi yang dibangkitkan berukuran d (d<n) maka data validasi berukuran (n-d). Data konstruksi berukuran d digunakan untuk menyusun model, sedangkan sisanya digunakan untuk validasi, sehingga nilai residual (error prediksi) dapat diestimasi[2].

Berdasarkan argument tersebut, muncul suatu permasalahan terkait dengan pemodelan data dependent khususnya data runtun waktu, yaitu bagaimana prosedur pembangkitan data konstruksi model berdasarkan rancangan acak blok tak lengkap berimbang serta proses validasi modelnya, sehingga dihasilkan model yang akurat dengan menimalkan fungsi kerugian untuk error prediksi.

2. RUNTUN WAKTU DAN m-DEPENDENT DATA

Barisan variabel random $\{Z_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ dikatakan stasioner kuat, apabila untuk sebarang bilangan bulat q dan $p > 0$, $\{Z_1, Z_2, \ldots, Z_p\}$ mempunyai distribusi yang sama dengan $\{Z_{1+q}, Z_{2+q}, \ldots, Z_{p+q}\}$. Semua barisan variabel random yang sedang dibicarakan diasumsikan stasioner. m-dependet adalah struktur ketergantungan yang paling sederhana dalam aplikasi statistik.

Barisan variabel random $\{Z_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ dikatakan sebagai m-dependent jika terdapat bilangan bulat nonnegative m sedemikian hingga untuk setiap bilangan bulat t, $\{\ldots, Z_{t-1}, Z_t\}$ dan $\{Z_{t+m+1}, Z_{t+m+2}, \ldots\}$ saling independent. Dari definisi tersebut, Z_i adalah independent dan berdistribusi identik apabila m=0. Jika $m \geq 1$, maka Z_i dependent. Sebagai contoh, model moving average.

Suatu model moving average (MA) adalah suatu model runtun waktu, dimana data $\{y_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ dapat dinyatakan sebagai

$$y_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_m\varepsilon_{t-m} \qquad (1)$$

dengan m bilangan bulat positif, μ dan $\theta_j, j = 1, 2, \ldots m$ adalah parameter yang tidak diketahui serta ε_t merupakan variabel random independent dan berdistribusi identik dengan mean 0 dan variansi σ^2. Hal ini dapat ditunjukkan bahwa $\{y_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ merupakan suatu deret m-dependent, yaitu:

$\operatorname{cov}(y_t, y_{t+p}) = \sigma^2 \sum_{j=p}^{m} \theta_j\theta_{j-p}$ untuk $1 \leq p \leq m$ ($\theta_0 = -1$) dan

$\operatorname{cov}(y_t, y_{t+p}) = 0$ untuk $p > m$.

Barisan variabel random dependent $\{y_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ sering disebut sebagai suatu runtun waktu (time series), walaupun variabel random tersebut tidak diindekkan dengan waktu. Dalam praktek banyak runtun waktu yang dinyatakan sebagai kombinasi linier dari variable random independent. Salah satu contohnya adalah model moving average pada contoh 1 di atas. Type runtun waktu yang sangat penting dan sering dibicarakan adalah model Autoregresive (AR).

Suatu runtun waktu $\{y_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ disebut sebagai suatu runtun waktu autoregressive order p apabila

$$y_t = \mu + \varepsilon_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} \qquad (2)$$

dengan p suatu bilangan bulat nonnegatip μ dan ϕ_j, $j = 1, 2, \ldots p$ adalah parameter yang tidak diketahui serta ε_t merupakan variabel random independent dan berdistribusi identik dengan mean 0 dan variansi σ^2. Suatu runtun waktu autoregressive dikatakan stasioner apabila:

akar-akar dari: $1 + \phi_1 B + \phi_2 B^2 + \ldots + \phi_p B^p = 0$ terletak di luar lingkaran satuan.

Model runtun waktu yang merupakan model campuran antara model autoregressive dan model average disebut model autoregresive moving average (ARMA). Suatu barisan variable random $\{y_t, t = 0, \pm1, \pm2, \pm3, \ldots.\}$ disebut sebagai suatu runtun waktu Autoregressive Moving Average order (p,q) dinyatakan sebagai ARMA(p,q) apabila:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \qquad (3)$$

dengan p dan q suatu bilangan bulat nonnegatip, μ, ϕ_i, θ_j; $i = 1, 2, \ldots p$; $j = 1,2, \ldots, q$ adalah parameter yang tidak diketahui serta ε_t merupakan variabel random independent dan berdistribusi identik dengan mean 0 dan variansi σ^2.

Suatu runtun waktu autoregressive moving average ARMA(p,q) dikatakan stasioner apabila:

akar-akar dari: $1 + \phi_1 B + \phi_2 B^2 + \ldots + \phi_p B^p = 0$ dan

akar-akar dari: $1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q = 0$ terletak di luar lingkaran satuan.

Secara umum model runtun waktu AR, MA dan ARMA tersebut merupakan model stasioner, sedangkan model runtun waktu nonstasioner dinyatakan sebagai model Autoregresive Integrited Moving Average (ARIMA). Jika didefinisikan w_t sebagai barisan selisih $w_t = y_t - y_{t-1}$ maka proses umum ARMA

$$w_t = \mu + \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \qquad (4)$$

dapat ditulis sebagai

$$y_t = \mu + y_{t-1} + \phi_1 (y_{t-1} - y_{t-2}) + \phi_2 (y_{t-2} - y_{t-3}) + \ldots + \phi_p (y_{t-p} - y_{t-p-1}) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \qquad (5)$$

Dari $w_t = y_t - y_{t-1}$ maka

$y_t = y_{t-1} + w_t$

$y_{t-1} = y_{t-2} + w_{t-1}$

$y_{t-2} = y_{t-3} + w_{t-2}$ dan seterusnya, sehingga diperoleh bahwa

$y_t = w_t + w_{t-1} + w_{t-2} + \ldots.$

Ini berarti bahwa y_t dapat dipandang sebagai integrasi runtun waktu w_t dan model runtun waktu (4) dipandang sebagai model proses ARIMA.

3. ESTIMASI MODEL DENGAN METODE BOX-JENKINS

Prosedur estimasi model AR, MA atau ARMA dengan metode Box-Jenkins, secara umum dapat dilakukan dengan langkah-langkah sebagai berikut.

1. Identifikasi Model

   Sebelum melakukan identifikasi model, berdasarkan data observasi terlebih dahulu dilakukan pengujian terhadap stasioneritas data. Apabila syarat stasioneritas dipenuhi maka dapat dilanjutkan dengan menentukan fungsi autokorelasi dan

fungsi autokorelasi parsial. Berdasarkan nilai autokorelasi dan autokorelasi parsial tersebut dapat dilakukan identifikasi model berdasarkan ciri-ciri model AR(p), MA(q) atau ARMA(p,q) seperti yang tercantum dalam tabel 1 berikut.

Tabel 1. Ciri-ciri teoritis F.a.k dan F.a.k.p untuk proses stasioner[3]

Model	Fungsi Autokorelasi (F.a.k)	Fungsi Autokorelasi Parsial (F.a.k.p)
AR(p)	Turun secara eksponensial atau berbentuk sinusoida	Terpotong setelah lag p
MA(q)	Terpotong setelah lag q	Turun secara eksponensial atau berbentuk sinusoida
ARMA(p,q)	Terpotong setelah lag (q-p)	Terpotong setelah lag (p-q)

2. Estimasi parameter

Apabila identifikasi model telah dilakukan, maka tahapan berikutnya adalah estimasi awal parameter model. Untuk menguji apakah parameter terkait dengan model yang telah diidentifikasi tersebut signifikan atau tidak, maka dilakukan langkah-langkah pengujian hipotesis sebagai berikut. Sebagai contoh, apabila model yang diidentifikasi adalah model AR maka langkah-langkah pengujian sigifikansi parameter modelnya adalah sebagai berikut:

a. Perumusan hipotesis

H_0: parameter (ϕ_j) = 0

H_1: parameter (ϕ_j) $\neq$ 0

b. Tingkat signifikansi α

c. Statistik uji

$t = \dfrac{\hat{\phi}_j}{se(\hat{\phi}_j)}$ berdistribusi t dengan derajat bebas (n-k-1)

d. Kriteria penolakan

Dengan menggunakan tingkat signifikansi α, maka H_0 akan ditolak apabila $|t_{hitung}| \geq t_{\alpha/2;(n-k-1)}$

e. Kesimpulan

Apabila H_0 ditolak maka dapat disimpulkan bahwa parameter model ϕ_j signifikan.

3. Verifikasi model

Verifikasi model dilakukan untuk memastikan apakah model yang telah diestimasi pada langkah 2 tersebut merupakan model terbaik atau bukan dengan cara melakukan underfit atau overfit. Model terbaik dipilih berdasarkan nilai fungsi kerugian yang minimal. Disamping itu juga dilakukan pengujian terhadap independensi nilai residual dengan uji chi-square (Box-Pierce).

4. Prediksi (forecasting)

   Model terbaik yang dipilih pada langkah ke-3 dapat digunakan untuk prediksi beberapa langkah ke depan.

4. ESTIMASI MODEL DENGAN VALIDASI SILANG

Prosedur estimasi model runtun waktu dengan menggunakan metode validasi-silang dapat dilakukan dengan langkah-langkah sebagai berikut.

1. Data observasi $\{y_1, y_2, \ldots, y_{t-1}, \ldots, y_n\}$ yang berukuran n dikelompokkan menjadi dua bagian yaitu: data konstruksi (DK) dan data validasi (DV). Pengambilan data konstruksi dilakukan dengan prinsip rancangan acak blok tak lengkap berimbang dengan ukuran sampel katakanlah d ($1<d<n$), sehingga sisanya merupakan data validasi berukuran $\alpha = (n-d)$. Dengan demikian jika A sebagai himpunan validasi alternative maka $A=\{\alpha\}$.

2. Berdasarkan data konstruksi tersebut, lakukan estimasi model dengan metode ARIMA dari Box-Jenkins.

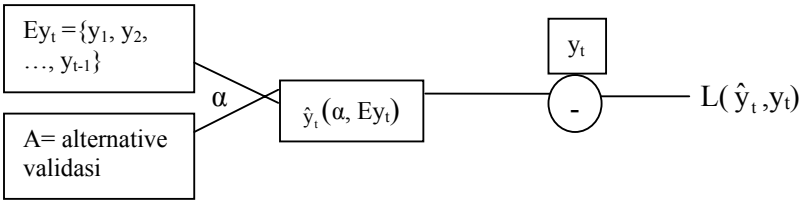
3. Lakukan validasi model yang diperoleh pada langkah ke-2 dengan membentuk fungsi kerugian (loss function) $L(\hat{y}_t, y_t)$. Fungsi kerugian yang dapat dibentuk berupa: $L(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$ yang disebut error kuadrat atau rata-rata error kuadrat.

4. Ulangi langkah 1 sampai dengan 3 tersebut, kemudian tentukan model terbaik berdasarkan nilai fungsi kerugian yang minimal.

5. Ulangi langkah 1 sampai dengan 4 dengan cara memvariasikan ukuran sampel d untuk data konstruksi, sehingga diperoleh estimasi model yang optimal yaitu yang meminimalkan fungsi kerugian untuk berbagai nilai d yang mungkin.

6. Tetapkan model terbaik yang dapat digunakan untuk prediksi beberapa langkah ke depan.

Jika diberikan data runtun waktu hasil observasi: $\{y_1, y_2, \ldots, y_{t-1}, \ldots, y_n\}$, dan DK: $Ey_t = \{y_1, y_2, \ldots, y_{t-1}\}$ data observasi hingga t-1, serta DV:$\{\alpha, \alpha \in A\}$: himpunan validasi alternatif, maka proses validasi model dapat digambarkan seperti bagan singkat berikut.



Gambar 1. Skema proses validasi model

$\hat{y}_t(\alpha, Ey_t)$: prediksi dari y_t untuk alternative estimasi α pada Ey_t.

$L(\hat{y}_t, y_t)$: fungsi kerugian untuk error prediksi

SIMULASI

Untuk memberikan ilustrasi tentang implementasi secara praktis estimasi model runtun waktu dengan metode validasi silang, Tabel 2 berikut memperlihatkan hasil simulasi terhadap sekumpulan data rata-rata jumlah produk cacat harian dari sebuah pabrik yang dicatat selama 45 hari [3].

Tabel 2. Hasil simulasi estimasi model AR dengan validasi-silang

Ukuran sampel		Estimasi Parameter Model AR(1)		Loss Function (MSE)
DK	DV	Parameter ($\hat{\phi}$)	Konstan ($\hat{\mu}$)	
30	15	0.6016	0.71186	0.306564866
31	14	0.5981	0.71486	0.305453267
32	13	0.5981	0.71483	0.329918638
33	12	0.5982	0.72215	0.360186517
34	11	0.6083	0.71338	0.394124836
35	10	0.5596	0.77919	0.400508023

Dari Tabel 2. terlihat bahwa untuk data observasi berukuran 45, dengan memvariasikan data konstruksi (DK) serta data validasi (DV) untuk berbagai ukuran diperoleh fungsi kerugian (MSE) minimal 0,305453267, sehingga estimasi modelnya adalah: $y_t = 0.71486+0.5981\ y_{t-1}$.

5. KESIMPULAN

Dalam proses pembangkitan data runtun waktu (data dependent) dengan metode validasi-silang tidak dapat dilakukan secara random terhadap data observasi, namun harus dilakukan dengan prinsip rancangan blok tak lengkap berimbang (balanced incomplete block design). Prinsip dasar dari metode validasi-silang adalah membagi data menjadi dua bagian yaitu data konstruksi dan data validasi. Data konstruksi diambil dari sekumpulan data secara blok kemudian dilakukan estimasi model dengan metode ARIMA Box-Jenkins. Setelah diperoleh estimasi model, dilakukan validasi model terhadap sisa data (data validasi). Proses tersebut diulang-ulang untuk blok yang berbeda sehingga diperoleh estimasi model yang terbaik dengan meminimalkan fungsi kerugian (loss function) untuk error prediksi.

DAFTAR PUSTAKA

1. Shao, J. & Tu, D, The Jackknife and Bootstrap, Springer-Verlag, New York, 1995.

2. Urban Hjorth, Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap, Chapman & Hall, London, 1994.

3. Wei, Time Series Analysis: Univariate and Multivariate Methods, Addison-Wesley Publishing Company-Inc. USA, 2006.

LAMPIRAN

Data rata-rata jumlah produk cacat dari suatu pabrik selama 45 hari[3].

No	y_t	No	y_t	No	y_t
1	1.20	16	2.25	31	1.85
2	1.50	17	2.50	32	1.82
3	1.54	18	2.05	33	2.07
4	2.70	19	1.46	34	2.32
5	1.95	20	1.54	35	1.23
6	2.40	21	1.42	36	2.91
7	3.44	22	1.57	37	1.77
8	2.83	23	1.40	38	1.61
9	1.76	24	1.51	39	1.25
10	2.00	25	1.08	40	1.15
11	2.09	26	1.27	41	1.37
12	1.89	27	1.18	42	1.79
13	1.80	28	1.39	43	1.68
14	1.25	29	1.42	44	1.78
15	1.58	30	2.08	45	1.84