BAB III
VALIDASI KEDEPAN
PADA MODEL PERAMALAN

Dalam peramalan menggunakan data runtun waktu, model yang dihasilkan dari analisa data runtun waktu harus memberikan ramalan yang lebih baik dan objektif. Untuk itu diperlukan validasi atau pengabsahan model tersebut. Terdapat banyak variasi dalam validasi ini, salah satunya adalah **VALIDASI KEDEPAN**, yaitu validasi derajat suatu model dari analisa data runtun waktu. Tujuannnya ialah untuk mendapatkan model yang lebih baik untuk peramalan dengan menggunakan data sebelumnya (yang disebut estimasi set) melalui suatu prosedur seleksi.

III.1 VALIDASI KEDEPAN.

Diberikan model runtun waktu yang dipengaruhi oleh trend linier dan musiman sebagai berikut:

$$Y(t)= b_0 + b_1(t) + \sum_{i:2}^{p} b_i Y_i \qquad (3.1)$$

dimana : $b_0 + b_1(t)$ adalah trend linier.

Dari (3.1) dapat dibentuk model peramalan τ langkah kedepan dengan menggunakan p parameter sebagai berikut :

$$\hat{Y}(t+\tau) = \beta_0 + \beta_1(t+\tau) + \sum_{i:2}^{p} \beta_i Y_{\beta(t-t_i)} \qquad (3.2)$$

dimana :

p = ukuran model

β_i = parameter yang diestimasi

ji = indek ke i pada prediktor terseleksi

t_i = lag pada prediktor terseleksi

Berapa jumlah parameter **p** dan bagaimana model peramalan yang dihasilkan merupakan model yang "terbaik" untuk peramalan, adalah masalah yang akan diselesaikan melalui metode Validasi kedepan.

Diberikan simbol-simbol sebagai berikut:

$ES_t=\{ y_1, y_2, \ldots\ldots, y_{t-1}\}$	Himpunan Observasi sampai data ke t-1 dan t=m,m+1,.....,n ,
$\{\alpha, \alpha \in A\}$	Himpunan validasi alternatif,
$\hat{y}_t(\alpha, ES_t)$	Prediksi y_t untuk alternatif α diestimasi menggunakan ES_t,
$L(y_t, \hat{y}_t)=(y_t - \hat{y}_t)^2$	Fungsi kerugian.

Simbol ES_t di atas menunjukkan himpunan observasi yang tersedia untuk satu langkah memprediksi y_t, ini berari ES_t menyatakan himpunan untuk estimasi, dan y_t sebagai himpunan tes. Sehingga bisa didefinisikan:

$$ES_t + TS_t = ES_{t+1}$$

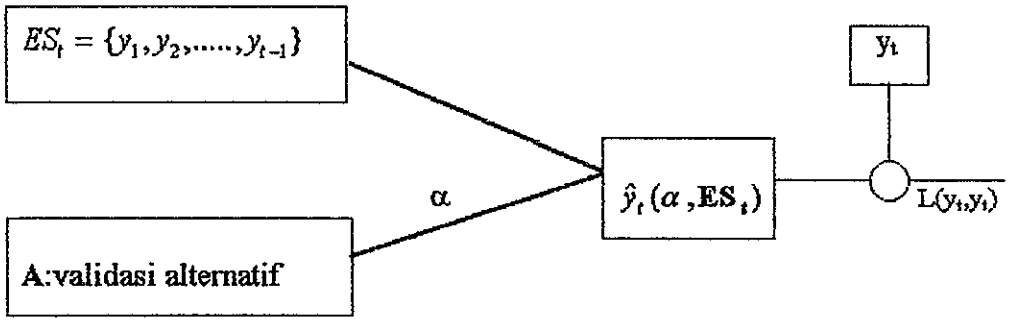
Langkah Validasi kedepan dimulai dengan memberikan data $y_1, y_2, \ldots, y_m, \ldots, y_n$ adalah runtun waktu ter-observasi. Runtun ini bisa skalar atau multivariat dan **diasumsikan** runtun waktu dipegaruhi Trend linier dan musiman

serta mengikuti distribusi normal. Disini digunakan notasi skalar dengan indeknya menyatakan harga observasi.

Dengan menggunakan ES_t, setelah waktu yang ke t-1 semua model yang berkorespondensi dengan alternatif $\{\alpha, \alpha \in A\}$, modelnya kemudian akan dicobakan pada observasi berikutnya yaitu y_t. Kemudian dihitung fungsi kerugian :

$$L(y_t, \hat{y}_t(\alpha, ES_t)) = (y_t - \hat{y}_t(\alpha, ES_t))^2. \quad (3.3)$$

Dimana $\hat{y}_t(\alpha, ES_t)$ menyatakan taksiran y dengan ukuran model α dan menggunakan ES_t untuk mendapatkan α. Proses perhitugan setelah waktu ke t-1 digambarkan seperti gambar 3.1 berikut ini:



Gambar 3.1 Perhitungan setelah t-1

Proses ini dilakukan sampai data observasi ke - n. Tiap pertambahan waktu t dihitung **ukuran keputusan (C(α))**, dan dicari α_0 yang berkorespondensi dengan bilangan parameter **P=P(α)** yang meminimalkan C(α). Bila $C(\alpha_0)$ harganya tidak melebihi **Estimasi hasil(CMF)**, maka α_0 ini dijadikan ukuran model peramalan. Model akhir peramalan τ langkah yang telah diabsahkan adalah :

$$\hat{Y}(t+\tau) = \beta_0 + \beta_1(t+\tau) + \sum_{i:2}^{\alpha_0} \beta_i Y_{j(t-t_i)} \quad (3.4)$$

dimana α_0 adalah banyaknya parameter terseleksi melalui metode seleksi kedepan.

III.2 ANALISA PREDIKSI ERROR SECARA REKURSI

Diambil n observasi (y_i,x_i'), i=1,2,...,n dan dianalisa menggunakan regresi linier $y=x_i'\beta+e_i$; dimana $x_i'=(x_{i1},x_{i2},...,x_{ip})$, $\beta'=(b_1,b_2,...,b_p)$ dan **e** independen berdistribusi $N(0,\sigma)$. Ambil $X'=(x_1,x_2,...,x_n)$, $Y'=(y_1,y_2,...,y_n)$ dan $\hat{\beta}=(X'X)^{-1}X'Y$. Anggap himpunan observasi yang baru dan independen (v_i,x_i') ,i=1,2,...,n dan akan dibawa ke bentuk vektor yang sama dengan x_i , dan ambil $V'=(v_1,v_2,...,v_n)$. **Rata-rata kuadrat error(Q)** menggunakan $\hat{\beta}$ untuk prediksi V adalah :

$$Q=\frac{1}{n}E[(V-X\hat{\beta})'(V-X\hat{\beta})|\hat{\beta}]$$

$$=\sigma^2+\frac{1}{n}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)$$

Ambil $Y'=(y_1,y_2,...,y_n)$, $X_i'=(x_1,x_2,...,x_i)$ untuk $i\geq m-1$ dan pilih m cukup besar untuk (X'X) menjadi tak singular. Anggap struktur model benar, didefinisikan:

$$z_i=y_i-x_i'\hat{\beta}_{i-1}$$

adalah **prediksi error secara rekursi** pada observasi ke i dan $\hat{\beta}=(X_i'X_i)^{-1}X_i'Y_i$

Sebagai estimasi variabel random Q digunakan **jumlah bobot kuadrat prediksi error :**

$$\hat{Q}=\sum_{i=m}^{n}\delta_i z_i^2$$

z_i independen dan berdistribusi normal ($N(0,\sigma d_i)$), dimana $d_i^2=1+\mathbf{x}_i'(\mathbf{X}_{i-1}\mathbf{X}_{i-1})^{-1}\mathbf{x}_i$ untuk $m\leq i\leq n$ dan bebas terhadap $\hat{\beta}_n$ yang mempunyai mean β dan kovarian matrik $\sigma^2(\mathbf{X}_n'\mathbf{X}_n)^{-1}$. Bobot δ_i dipilih yang meminimalkan $E(\hat{Q}-Q)^2$.

$$E(\hat{Q}-Q)^2 = E(\sum_{i=m}^{n} \delta_i z_i^2 - \sigma^2 - \frac{1}{n}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta))^2$$

Untuk mendapatkan δ_k dengan menghilangkan faktor persekutuan $\sigma^4\delta_k^2$ dan menggunakan $(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)/\sigma^2$ berdistribusi Chi-Kwadrat dengan derajat kebebasan **p** kita dapatkan :

$$\sum_{i=m}^{n} \delta_i d_i^2 + 2\delta_k d_k^2 = 1+p/n$$

$\delta_k d_i^2$ adalah konstan dan bobot yang optimal adalah :

$$\delta_k = \frac{1+p/n}{n-m+3}\,\frac{1}{d_k^2}$$

Jika $\mathbf{x}_k'(\mathbf{X}_{k-1}\mathbf{X}_{k-1})\mathbf{x}_k$ diganti dengan p/(k-1) yang merupakan harga rata-rata dari $\mathbf{x}_i'(\mathbf{X}_{k-1}\mathbf{X}_{k-1})\mathbf{x}_i$ untuk $1\leq i \leq k-1$, didapatkan :

$$\delta_k = \frac{1+p/n}{n-m+3}\,\frac{1}{1+p/(t-1)} \qquad (3.5)$$

untuk $m\leq k \leq n$.

III.3. PEMBOBOTAN

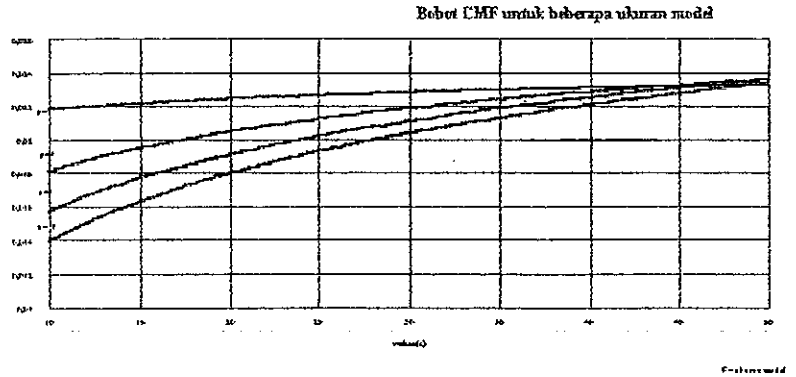
Estimasi hasil (CMF) yang diberikan bermaksud untuk memberikan sebuah estimasi pada hasil yang akan datang dari model terseleksi. Disini bobot δ_t persamaan (3.5), untuk i=t digunakan. Karena $\alpha_0(ES_t)$ yang muncul menjadi optimal pada saat t yang berbeda-beda, maka disini tidak diberikan harga konstan pada p didalam bobot, dan juga bobot harus berjumlah kurang dari satu.

Ambil $p(t) = (\alpha_0(ES_t))$ *sebagai ukuran model optimal* yang diestimasi pada waktu t, dari persamaan persamaan (3.5) menjadi :

$$\delta_t = \frac{1 + p(t)/n}{(n - m - 3)(1 + (p(t)/(t-1))} \qquad (3.6)$$

untuk $m \le t \le n$.

Untuk m=10, n=50 dan p=1,4,7,10 bobot δ_t digambarkan sebagai berikut :



Gambar 3.2. Bobot δ untuk CMF pada beberapa ukuran model pada waktu antara m=10, n=50

Bobot γ_{tn} dalam ukuran keputusan(c(α)) untuk seleksi model digunakan bagai penormalan dari bobot δ_t , maka jumlahnya sama dengan satu. Bobot γ_{tn} lefinisikan sebagai berikut :

$$= \delta_t \Big/ \sum_{i=m}^{n} \delta_i$$

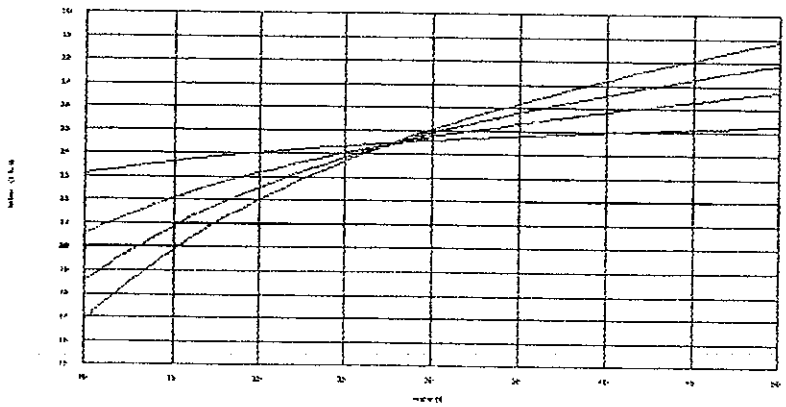
n dengan menggunakan persamaan (3.6) didapatkan:

$$= \frac{1 + p(t)/n}{(n-m+3)(1+(p(t)/(t-1))} \Big/ \sum_{i=m}^{n} \frac{1+p(t)/n}{(n-m+3)(1+p(t)/(i-1)}$$

$$= \frac{1}{1+\frac{p}{t-1}} \Big/ \sum_{i=m}^{n} \frac{1}{1+\frac{p}{i-1}} \qquad (3.7)$$

uk $m \le t \le n$

Prosedur seleksi C(α) untuk tiap keabsahan alternatif α ber-espondensi pada sebuah bilangan P=P(α), yang menyatakan banyaknya ameter yang digunakan untuk model yang berasosiasi dengan α . Bobot γ_{tn} digambarkan sebagai berikut :



Gambar 3.3. Bobot model seleksi γ untuk beberapa ukuran model P pada waktu antara m=10 , n=50.

Pada Gambar 3.2 dan Gambar 3.3 diatas terlihat pada bobot δ dan γ pada p(t) yang berbeda, grafik akan naik untuk tiap-tiap p(t) menurut waktu t. Kurva naik ini mempunyai alasan karena informasi dari ES_t bertambah sampai mendekati akhir observasi, ES_n. Pada proses mendekati akhir, data yang digunakan lebih banyak dibandingkan pada awal validasi, sehingga kemungkinan yang timbul adalah error yang ditimbulkan akan lebih besar. Meramal error pada saat akhir lebih relevan dari pada disaat awal, lagi pula pada saat akhir pengaruh error lebih kuat dan lebih besar terhadap bilangan parameter model peramalan yang dihasilkan. Bertambahnya resiko naiknya error juga menerangkan mengapa bobot δ berjumlah kurang dari satu

III.4. PROSEDUR SELEKSI

Prosedur seleksi adalah prosedur untuk menentukan ukuran model untuk peramalan. Ambil beberapa data sebelum model-model dapat diestimasi, misalkan ada **n** buah data : $y_1,y_2,....,y_n$. Ambil $ES_t=y_1,y_2,....,y_{t-1}$ observasi sebagai himpunan estimasi, dan validasi dimulai dari data item ke **m.** Dengan menggunakan metode **pemilihan kedepan (Forward Selection)** tanpa uji F, didapatkan beberapa prediktor-prediktor terseleksi dan estimasi β, yang akan digunakan untuk menaksir y_t. Fungsi kerugian $L(y_t,\hat{y}_t(\alpha, ES_n))$ yang dihasilkan pada perhitungan menggunakan ES_{t_n} dihitung **ukuran keputusan** yaitu :

$$C(\alpha) = \sum_{t=m}^{n} \gamma_{tn} L(y_t, \hat{y}_t(\alpha, ES_t)) \qquad (3.8)$$

Akhirnya dipilih alternatif α_0 yang **meminimalkan** $C(\alpha)$ yang berkorespondensi dengan bilangan parameter $P=P(\alpha)$. Dimana $P=P(\alpha)$ menyatakan ukuran model terpilih untuk waktu ke t.

III.5. ESTIMASI HASIL

Seleksi yang dibuat setelah waktu ke **n** dapat dibuat dengan waktu lebih awal setelah **t-1**. Ukuran keputusan $C(\alpha)$ kemudian dapat diganti dengan jumlah yang lebih pendek sampai **n'=t-1**

Ambil $\alpha_0(ES_t)$ yang meminimalkan $C(\alpha)$ sebagai *hasil dari prosedur seleksi*. Karena α_0 berkorespondensi dengan banyaknya parameter untuk peramalan, maka dengan menggunakan $\alpha_0(ES_t)$ dibuat taksiran $\hat{y}_t(\alpha_0(ES_t))$ dan dicobakan pada y_t dihasilkan fungsi kerugian :

$$L(y_t, \hat{y}_t(\alpha_0 ES_t), ES_t))$$

dan dengan menggunakan bobot yang berbeda dengan γ_{tn} didefinisikan **estimasi hasil :**

$$CMF = \sum_{t=m}^{n} \delta_{tn} L(y_t, \hat{y}_t(\alpha_0(ES_t), ES_t) \qquad (3.9)$$

yang merupakan **estimasi rata-rata kuadrat prediksi error** untuk validasi kedepan yang berkorespondensi dengan $\alpha_0(S_n)$.

CMF ini bertujuan untuk memberikan estimasi yang valid pada nilai fungsi kerugian pada model terseleksi, maka **nilai ukuran keputusan** $C(\alpha_0)$ **valid bila mempunyai nilai yang lebih kecil dari nilai CMF.**

Selisih $\Delta = CMF - C(\alpha_0)$ (3.10)

dapat digunakan sebagai estimasi bias minimum pada $C(\alpha)$, menggunakan α konstan melalui data runtun waktu dan tergantung pada bobot berbeda yang digunakan.

III.6. ANALISA RESIDUAL

Penelaahan nilai sisa (error of fit) sangat penting untuk memutuskan kecocokan model peramalan yang diberikan. Jika kesalahan secara esensial bersifat random, maka model tersebut mungkin baik. Jika kesalahan menunjukkan suatu pola, maka berarti model tersebut tidak memperhatikan semua informasi sistematis pada semua himpunan data.

Model peramalan yang baik residunya tidak menunjukkan suatu pola. Untuk menunjukkan ini maka digunakan metode **autokorelasi residual dengan k kelambatan**(r_k). Definisi autokorelasi residual (nilai sisa yang ber-autokorelasi) adalah bahwa nilai sisa yang tertinggal setelah penerapan suatu metode peramalan ternyata ber-autokorelasi, berarti bahwa model peramalan tersebut belum menghilangkan seluruh pola data *(Makridrakis,1983)*. Maka model yang baik tidak mempunyai autokorelasi atau autokorelasi mendekati harga nol. Untuk mengitung autokorelasi dengan k kelambatan diberikan rumus :

$$r_k = \frac{\sum_{t=k+1}^{n} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \qquad (3.11)$$

autokorelasi bernilai sekitar antara -1 dan 1.

Koefisien korelasi dari data random mempunyai distribusi sampling yang mendekati kurva normal N(0,1) dengan nilai tengah nol dan kesalahan standar(Se_{rk}) didefinisikan :

$$Se_{rk} = \frac{1}{\sqrt{n}} \tag{3.12}$$

Ini berarti bahwa 95% dari seluruh koefisien korelasi berdasarkan sampel harus terletak didalam daerah nilai tengah ditambah atau dikurangi 1,96 kali kesalahan standar Se_{rk}. Nilai 1,96 menyatakan luas daerah dibawah kurva normal. Karena mendekati 2, maka biasanya dibulatkan menjadi 2. Sebagai kesimpulan adalah bahwa *residual bersifat random apabila koefisien korelasi dihitung berada didalam selang :*

$$-1{,}96\, Se_{rk} \leq r_k \leq +1{,}96\, Se_{rk} \tag{3.13}$$