JURNAL MATEMATIKA DAN KOMPUTER
Vol. 4. No. 1, 46 - 58, April 2001, ISSN : 1410-8518

PEMILIHAN MODEL REGRESI LINIER DENGAN BOOTSTRAP

Tarno
Jurusan Matematika FMIPA UNDIP Semarang

Subanar
Jurusan Matematika FMIPA UGM Yogyakarta

Abstrak

Tulisan ini membicarakan tentang penerapan bootstrap untuk pemilihan model regresi linier terbaik. Model regresi linier terbaik yang terpilih adalah model dengan estimasi sesatan prediksi kuadrat minimal atas semua model regresi yang mungkin yaitu sebanyak 2^p-1 model dengan p: banyaknya variabel prediktor. Metode Bootstrap memilih suatu model dengan meminimalkan rata-rata sesatan prediksi kuadrat berdasarkan resampling data yang dibangkitkan melalui pasangan data dan residual, dengan mempertimbangkan juga variabel prediktor yang terlibat sesedikit mungkin. Pemilihan variabel berdasarkan bootstrap pasangan data dan bootstrap residual dengan n ukuran sampel bootstrap adalah konsisten. Dan jika ukuran sampel bootstrap diambil m dengan $\frac{m}{n} \to 0$ and $m \to \infty$, pemilihan variabel bootstrap juga konsisten. Hasil dari suatu simulasi dengan SPLUS disajikan dalam tulisan ini.

Kata kunci : pemilihan model, bootstrap dan sasatan prediksi.

1. PENDAHULUAN

Salah satu model yang sangat berguna dalam berbagai bidang aplikasi adalah model linier umum :

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1,2,...,n \qquad (1.1)$$

dengan $\mathbf{y}_i$ adalah respon ke-i, $\mathbf{x}_i$: p-vektor variabel prediktor yang berkaitan dengan y_i, $\boldsymbol{\beta}$: p-vektor parameter yang tidak diketahui dan ε_i : sesatan random. Masalah regresi linier dapat diformulasikan sebagai kasus khusus dari model (1.1)

tersebut. Jika $\mathbf{x}_i$ dalam model (1.1) deterministik, maka diasumsikan bahwa ε_i independen dengan mean 0 dan variansi σ^2. Jika $\mathbf{x}_i$ tersebut random, maka model (1.1) dikatakan sebagai model korelasi. Dalam suatu model korelasi, $(y_i, \mathbf{x}_i')$ diasumsikan independen dan berdistribusi identik dengan momen kedua berhingga dan $E(y_i \mid \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$, σ_i^2 menyatakan variansi bersyarat dari y_i diberikan $\mathbf{x}_i$.

Parameter regresi $\boldsymbol{\beta}$ dapat diestimasi dengan menggunakan metode kuadrat terkecil. Apabila estimasi parameter telah ditentukan berarti diperoleh estimasi model untuk respon **y** yang tergantung pada prediktor **x**. Tetapi beberapa komponen dari **x** kemungkinan tidak berpengaruh secara signifikan terhadap respon **y**, sehingga perlu dilakukan pemilihan variabel prediktor. Pemilihan variabel dalam model regresi linier ini dapat dilakukan dengan beberapa metode : AIC, Cp, BIC, Jackknife (Validasi Silang) dan Bootstrap (Shao, Tu, 1995). Permasalahan yang diuraikan disini adalah pemilihan model yang lebih kompak yaitu model yang memiliki estimasi rata-rata sesatan prediksi kuadrat minimum. Dan pemilihan modelnya dilakukan dengan menggunakan metode yang berdasarkan data-resampling yaitu bootstrap. Karena beberapa komponen dari $\boldsymbol{\beta}$ mungkin sama dengan nol, maka model yang lebih kompak memiliki bentuk:

$$\mathbf{y} = \mathbf{x}_\alpha'\boldsymbol{\beta}_\alpha + \boldsymbol{\varepsilon}, \qquad (1.2)$$

dengan α himpunan bagian dari $\{1,2,\ldots,p\}$.

Metode bootstrap memilih model dengan meminimalkan estimasi rata-rata jumlah kuadrat dari sesatan prediksi $\overline{mse}$ atas semua α berdasarkan sampel bootstrap residual dan bootstrap pasangan data pengamatan yang dibangkitkan dari fungsi distribusi empiris (Shao, Tu, 1995).

2. PEMILIHAN MODEL DAN SESATAN PREDIKSI

Prediksi nilai respon untuk masa yang akan datang, secara aktual mungkin tidak tergantung pada semua komponen **x**, artinya terdapat beberapa komponen $\boldsymbol{\beta}$ yang sama dengan nol (Shao, Tu, 1996). Oleh karena itu, didapatkan model yang lebih kompak yang berbentuk :

$$y_i = \mathbf{x}_{i,\alpha}^{'}\boldsymbol{\beta}_\alpha + \varepsilon_i, i = 1,2,...,n \qquad (2.1)$$

dengan $\alpha \subset \{1,2,...,p\}$. Jika $\boldsymbol{\beta}_\alpha$ dan $\mathbf{x}_{i,\alpha}$ sebagai subvektor yang memuat komponen-komponen dari $\boldsymbol{\beta}$ dan $\mathbf{x}_i$ berada dalam α, maka terdapat (2^p-1) model berbeda yang mungkin yang berbentuk (2.1), masing-masing terkait dengan suatu himpunan bagian α dan dinotasikan dengan $\hat{\alpha}$. Dimensi (ukuran) dari $\hat{\alpha}$ adalah banyaknya prediktor dalam $\hat{\alpha}$ (Shao dan Tu, 1995).

Jika ***A*** menyatakan semua himpunan bagian dari {1,2,…,p} dan setiap komponen dari $\boldsymbol{\beta}$ diketahui sama dengan 0 atau tidak, maka model-model $\hat{\alpha}$ dapat diklasifikasikan menjadi dua kategori :

- Kategori I (an incorrect model) : minimal satu komponen dari $\boldsymbol{\beta}$ yang tidak nol tidak berada dalam $\boldsymbol{\beta}_\alpha$.
- Kategori II (a correct model) : $\boldsymbol{\beta}_\alpha$ memuat semua komponen $\boldsymbol{\beta}$ yang tidak nol.

Model optimal adalah model (2.1) dengan α_0 sedemikian hingga $\boldsymbol{\beta}_{\alpha 0}$ memuat semua komponen $\boldsymbol{\beta}$ yang semuanya tidak nol (model dalam kategori II dengan dimensi terkecil). Model optimal tersebut tidak diketahui karena $\boldsymbol{\beta}$ tidak diketahui, sehingga perlu memilih model dari (2.1) berdasarkan data (y_1,x_1), (y_2,x_2), …, (y_n,x_n) yang memenuhi (1.1). Jika diasumsikan bahwa $\boldsymbol{\varepsilon}$ i.i.d dengan mean 0 dan variansi σ^2, maka dibawah model α, dengan Estimator Kuadrat Terkecil diperoleh:

$\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = (\mathbf{X}_{\boldsymbol{\alpha}}^{'}\mathbf{X}_{\boldsymbol{\alpha}})^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^{'}\mathbf{y}$ dengan $\mathbf{y}=(y_1,y_2,\ldots,y_n)^{'}$ dan $\mathbf{X}_\alpha = (x_{1\alpha}, x_{2\alpha},..., x_{n\alpha})$.

Jika dianggap bahwa $\mathbf{y}_f$: nilai respon yang akan datang untuk suatu nilai prediktor $\mathbf{x}_f$, maka $\hat{y}_{f\alpha} = \mathbf{x}_{f\alpha}^{'}\hat{\boldsymbol{\beta}}_\alpha$. Hal ini berakibat bahwa mean sesatan prediksi kuadrat $mse(\mathbf{x}_f, \alpha)$ adalah :

$$mse(\mathbf{x}_f, \alpha) = E(y_f - \hat{y}_{f\alpha})^2 = \sigma^2 + \sigma^2\mathbf{x}_{f\alpha}^{'}(\mathbf{X}_\alpha^{'}\mathbf{X}_\alpha)^{-1}\mathbf{x}_{f\alpha} + \Delta(\mathbf{x}_f, \alpha),$$

dengan $\Delta(\mathbf{x}_f, \alpha) = [\mathbf{x}_f^{'}\boldsymbol{\beta} - \mathbf{x}_{f\alpha}^{'}(\mathbf{X}_\alpha^{'}\mathbf{X}_\alpha)^{-1}\mathbf{X}\boldsymbol{\beta}]^2$.

Jika α dalam kategori II maka $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_\alpha\boldsymbol{\beta}_\alpha$, $\mathbf{x}_f^{'}\boldsymbol{\beta} = \mathbf{x}_{f\alpha}^{'}\boldsymbol{\beta}_\alpha$ dan $\Delta(\mathbf{x}_f, \alpha) = 0$. Dengan demikian jika $mse(\mathbf{x}_f, \alpha)$ diketahui, maka model optimal dapat dipilih dengan meminimalkan $mse(\mathbf{x}_f, \alpha)$ atas semua $\alpha \in \boldsymbol{A}$. Model optimal dapat juga

ditentukan dengan meminimalkan rata-rata dari sesatan prediksi kuadrat $\text{mse}(\mathbf{x}_f, \alpha)$ atas $X=\{x_1,x_2,\ldots,x_n\}$:

$$\overline{\text{mse}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\text{mse}(\mathbf{x}_i, \alpha) = \sigma^2 + \frac{\sigma^2 p}{n} + \Delta(\alpha),$$

dengan $\Delta(\alpha) = \frac{1}{n}\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}_n - \mathbf{H}_\alpha)\mathbf{X}\boldsymbol{\beta}$ dan $\mathbf{H}_\alpha = \mathbf{X}_\alpha(\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha'$.

Namun, $\text{mse}(\mathbf{x}_f, \alpha)$ dan $\overline{\text{mse}}(\alpha)$ kedua-duanya tidak diketahui. Sehingga mengestimasi $\overline{\text{mse}}(\alpha)$ lebih mudah dari pada mengestimasi $\text{mse}(\mathbf{x}_f, \alpha)$ dengan menggunakan $\overline{\wedge\text{mse}}(\alpha)$, kemudian memilih model dengan meminimalkan $\overline{\wedge\text{mse}}(\alpha)$ atas $\alpha \in \boldsymbol{A}$.

3. ESTIMASI PARAMETER

Jika parameter $\boldsymbol{\beta}$ merupakan parameter regresi yang akan diestimasi dengan $\hat{\boldsymbol{\beta}}$, maka di lingkungan boostrap $\hat{\boldsymbol{\beta}}$ dapat diestimasi dengan $\hat{\boldsymbol{\beta}}^*$. Untuk mengestimasi parameter regresi dapat dilakukan dengan beberapa prosedur bootstrap, antara lain: bootstrap berdasarkan residual, bootstrap pasangan data pengamatan .

Bootstrap Residual

Bootstrap berdasarkan residual (residuals bootstrap) disingkat RB (Efron, 1979). Jika diketahui model regresi (1.1) :

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1,2,...,n$$

dengan $\mathbf{y}_i$ adalah respon ke-i , $\mathbf{x}_i$: p-vektor variabel prediktor yang berkaitan dengan y_i, $\boldsymbol{\beta}$: p-vektor parameter yang tidak diketahui dan ε_i : sesatan random, dan apabila $\mathbf{x}_i$ nonrandom, dengan asumsi bahwa ε_i i.i.d dengan mean 0 dan variansi σ^2. Maka untuk memperoleh estimasi parameter dapat dilakukan prosedur bootstrap sebagai berikut :

a. Model regresi (1.1) diidentifikasi sebagai model parameter, yaitu $(\boldsymbol{\beta}, F_\varepsilon)$, dengan F_ε merupakan distribusi dari ε_i yang tak diketahui.

b. Dengan metode kuadrat terkecil, parameter $\boldsymbol{\beta}$ diestimasi dengan $\hat{\boldsymbol{\beta}}$ dan F_ε diestimasi dengan fungsi distribusi empiris $\hat{F}_\varepsilon$ dengan mengambil massa peluang n^{-1} terhadap $r_i - \bar{r}, i = 1,2,...,n$ dengan $r_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ merupakan residual ke-i dan $\bar{r} = n^{-1}\sum_{i=1}^{n} r_i$. $\hat{F}_\varepsilon$ dipusatkan pada 0 karena F_ε mempunyai mean 0.

c. Data bootstrap dibangkitkan dari model tersebut dengan $(\boldsymbol{\beta}, F_\varepsilon)$ diganti dengan $(\hat{\boldsymbol{\beta}}, \hat{F}_\varepsilon)$. Dengan kata lain dibangkitkan data independen dan berdistribusi identik $\varepsilon_1^*, \varepsilon_2^*, ..., \varepsilon_n^*$ dari $\hat{F}_\varepsilon$ dan didefinisikan $y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \varepsilon_i^*$.

d. Dengan metode kuadrat terkecil, dihitung $\hat{\boldsymbol{\beta}}^*$ berdasarkan data $(y_1^*, x_1'), (y_2^*, x_2'), ...(y_n^*, x_n')$, yaitu : $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$.

e. Ulangi langkah diatas sebanyak B kali sebagai replikasi bootstrap.

Karena $E_*(\varepsilon_i^*) = 0, untuk setiap\ i$, maka

$$E_*(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_*(\mathbf{y}^*) = \hat{\boldsymbol{\beta}}, \tag{3.1}$$

dan juga ,

$$var_*(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ var_*(\mathbf{y}^*)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.2}$$

dengan $\hat{\sigma}^2 = var(\varepsilon_i^*) = n^{-1}\sum_{i=1}^{n}(r_i - \bar{r})^2$.

Dari persamaan (3.1) dan (3.2), prosedur bootstrap ini menghasilkan estimator variansi dan estimator bias dari $\hat{\boldsymbol{\beta}}$ yang konsisten.

Suatu asumsi yang penting untuk bootstrap residual adalah ε_i i.i.d . Bahkan jika asumsi ini berlaku fungsi distribusi empiris $\hat{F}_\varepsilon$ tidak didasarkan pada data i.i.d secara eksak. Estimator (3.2) mempunyai bentuk secara eksplisit tetapi tidak sama dengan estimator yang diberikan oleh:

$$var(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n-p}\sum_{i=1}^{n} r_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.3}$$

dengan $r_i = y_i - x_i'\hat{\boldsymbol{\beta}}$ merupakan residual ke-i.

Estimator dalam persamaan (3.2) bias menurun, karena $E(\hat{\sigma}^2) = (1 - \frac{k_n}{n})\sigma^2$, dengan $k_n = p + 1 - n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i'(\mathbf{X}'\mathbf{X})^{-1}x_j \geq p \geq 0$. Untuk menghilangkan bias negatif ini, Shao dan Tu (1995) menyatakan bahwa data bootstrap diambil dari fungsi distribusi empiris berdasarkan $(r_i - \bar{r})/\sqrt{1-p/n}, i = 1,2,...,n$. Ketentuan ini masih mengarah kepada suatu estimator variansi yang bias menurun, karena $k_n > p$ bilamana $\bar{r}=0$. Kemudian data bootstrap dibangkitkan dari fungsi distribusi empiris $\tilde{F}_\varepsilon$ dengan mengambil massa peluang n^{-1} terhadap residual yang teratur $(r_i - \bar{r})/\sqrt{1-k_n/n}, i = 1,2,...,n$. Maka (3.2) berlaku dengan $\hat{\sigma}^2$ diganti dengan :

$$\tilde{\sigma}^2 = \frac{1}{n-k_n}\sum_{i=1}^{n}(r_i - \bar{r})^2 .$$

Jika $\bar{r}=0$ maka $k_n = p$, $\tilde{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2$ dan estimator variansi bootstrap untuk $\hat{\boldsymbol{\beta}}$ sama seperti estimator variansi tak bias yang diberikan oleh (3.3). Penetapan $1-k_n/n$ tidak memberikan dampak substansial terhadap estimasi variansi jika n cukup besar.

Bootstrap Data Berpasangan

Bootstrap berpasangan (paired bootstrap) disingkat PB, nampaknya merupakan suatu prosedur yang sangat alami apabila $\mathbf{x}_i$ random dan $(y_i, x_i'), i = 1,2,...,n$, independen dan berdistribusi identik (i.i.d). Dalam kasus ini, untuk mengestimasi parameter regresi dapat dilakukan prosedur sebagai berikut :

a. Model diidentifikasi dengan distribusi bersama dari $(y_i, x_i'), i = 1,2,...,n$ dan diestimasi dengan fungsi distribusi empiris dengan massa peluang n^{-1} untuk setiap $(y_i, x_i'), i = 1,2,...,n$.
b. Dibangkitkan data bootstrap dari fungsi distribusi empiris ini, yaitu :
   $(y_1^*, x_1^*), (y_2^*, x_2^*), ...(y_n^*, x_n^*)$.

c. Dengan metode kuadrat terkecil ditentukan estimasi parameter regresi :

d. $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{y}^*$.

e. Ulangi langkah diatas sebanyak B sebagai replikasi bootstrap.

4. PREDIKSI

Suatu penerapan yang sangat penting dari model (1.1) adalah prediksi dari respon yang akan datang y_f untuk suatu nilai prediktor x_f yang diberikan. Bootstrap dapat digunakan untuk menentukan sesatan prediksi dalam masalah prediksi titik. Dibawah model (1.1) dengan sesatan ε_i i.i.d, suatu prediksi titik untuk $\mathbf{y}_f$ adalah :

$$\hat{y}_f = x_f'\hat{\boldsymbol{\beta}},$$

dan ini dapat dievalusi dengan mean sesatan prediksi kuadrat,

$$\text{mse}(x_f) = E(y_f - \hat{y}_f)^2 .$$

Jika y_f dan $y_1, y_2, \ldots, y_n$ independen, maka

$$\text{mse}(x_f) = \text{var}(y_f) + \text{var}(x_f'\hat{\boldsymbol{\beta}}) = \sigma^2 + \sigma^2 x_f'(\mathbf{X}'\mathbf{X})^{-1}x_f .$$

Suatu estimator dari mse(x_f), berdasarkan bootstrap residual (RB) dapat diperoleh sebagai berikut. Jika $(\varepsilon_1^*, \varepsilon_2^*, ..., \varepsilon_n^*)$ dan ε_f^* i.i.d dari $\tilde{F}_\varepsilon$, yaitu fungsi distribusi empiris dari residual yang diatur $(r_i - \bar{r})/\sqrt{1-k_n/n}, i = 1,2,...,n$. Jika $y_f^* = x_f'\hat{\boldsymbol{\beta}} + \varepsilon_f^*$ suatu nilai respon bootstrap yang akan datang dan $\hat{y}_f^* = x_f'\hat{\boldsymbol{\beta}}^*$ prediksi bootstrap dari y_f^*. Maka estimator bootstrap untuk $\text{mse}(x_f)$ adalah:

$$\overset{\triangle}{\text{mse}}_{\text{BOOT}}(x_f) = E_*(y_f^* - \hat{y}_f^*)^2 = \text{var}_*(y_f^*) + \text{var}_*(x_f'\hat{\boldsymbol{\beta}}^*)$$

$$= \tilde{\sigma}^2 + \tilde{\sigma}^2 x_f'(\mathbf{X}'\mathbf{X})^{-1}x_f$$

dengan $\tilde{\sigma}^2 = \dfrac{1}{n-k_n}\sum_{i=1}^{n}(r_i - \bar{r})^2$.

Karena $E(\tilde{\sigma}^2) = \sigma^2$ maka $\overset{\wedge}{\text{mse}}_{BOOT}(x_f)$ merupakan estimator tak bias.

Kadang-kadang nilai yang akan datang ingin diprediksikan untuk suatu himpunan $\boldsymbol{X}$ dari x_f. Apabila $x_1, x_2, \ldots, x_n$, x_f random serta independen dan berdistribusi identik (i.i.d) maka mean sesatan prediksi kuadrat adalah:

$$\text{mse}(x_f) = \sigma^2 + \sigma^2 \text{tr}[E(\mathbf{X}'\mathbf{X})^{-1} E(x_f x_f')] = \sigma^2 + \frac{\sigma^2 p}{n} + o(n^{-1}).$$

Berdasarkan bootstrap pasangan Efron (Shao dan Tu,1995) mengusulkan estimator bootstrap untuk $\overline{\text{mse}}$. Didefinisikan sesatan ekses harapan dengan

$$e = \left[(y_f - \hat{y}_f)^2 - \frac{1}{n}\sum_{i=1}^{n} r_i^2 \right]$$

dan estimator boostrapnya dengan

$$\hat{e} = E_* \left[\frac{1}{n}\sum_{i=1}^{n} (y_i - x_i' \hat{\boldsymbol{\beta}}^*)^2 - \frac{1}{n}\sum_{i=1}^{n} (y_i^* - x_{i\alpha}^{*'} \hat{\boldsymbol{\beta}}^*)^2 \right]$$

Dengan demikian estimator boostrapnya adalah:

$$\overset{\wedge}{\text{mse}}_{BOOT} = \frac{1}{n}\sum_{i=1}^{n} r_i^2 + \hat{e} \quad . \tag{4.1}$$

Dan nilai harapannya adalah :

$$E(\overset{\wedge}{\text{mse}}_{BOOT}) = \sigma^2 + \frac{\sigma^2 p}{n} + o(n^{-1}).$$

Jadi estimator bootstrap tersebut hampir tak bias.

5. PEMILIHAN VARIABEL

Prosedur pemilihan model dengan bootstrap dapat diturunkan dari estimator untuk $\overline{\text{mse}}\,(\alpha)$. Pandang estimator bootstrap berbentuk :

$$\overset{\wedge}{\text{mse}}_{BOOT}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_{i\alpha}' \hat{\boldsymbol{\beta}}_\alpha)^2 + \hat{e}(\alpha) \tag{5.1}$$

dengan $\hat{e}(\alpha)$: estimator bootstrap dari sesatan ekses harapan :

$$e(\alpha) = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_{f,i} - x_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - x_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2 \right] = \frac{2\sigma^2 p_\alpha}{n}$$

dengan $y_{f,i}$ respon yang akan datang pada x yang independen dengan y_i .

Jika $\hat{\boldsymbol{\beta}}_\alpha^*$ estimator dari $\hat{\boldsymbol{\beta}}_\alpha$ di lingkungan bootstrap maka diperoleh :

$$\hat{\boldsymbol{\beta}}_\alpha^* = (\mathbf{X}_\alpha^{*'}\mathbf{X}_\alpha^*)^{-1}\mathbf{X}_\alpha^{*'}\mathbf{y}_\alpha^*$$

untuk bootstrap berpasangan, dan

$$\hat{\boldsymbol{\beta}}_\alpha^* = (\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha'\mathbf{y}_\alpha^*$$

untuk bootstrap residu, dengan $\mathbf{y}_\alpha^* = (y_{1\alpha}^*, y_{2\alpha}^*, ..., y_{n\alpha}^*)$, $y_{i\alpha}^* = x_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha + \varepsilon_i^*$

dan ε_i^* independen dan berdistribusi identik dari $\hat{F}_\varepsilon$ yaitu fungsi distribusi empiris dari $\boldsymbol{\varepsilon}$ dengan massa peluang 1/n. Dengan demikian diperoleh estimator bootstrap dari $e(\alpha)$ untuk bootstrap berpasangan adalah :

$$\hat{e}(\alpha) = E_*\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_{i\alpha}^{*'}\hat{\boldsymbol{\beta}}_\alpha^*)^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i^* - x_{i\alpha}^{*'}\hat{\boldsymbol{\beta}}_\alpha^*)^2 \right] \qquad (5.2)$$

dan untuk bootstrap residu :

$$\hat{e}(\alpha) = E_*\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha^*)^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i^* - x_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha^*)^2 \right] \qquad (5.3)$$

Estimator bootstrap untuk $\overline{mse}(\alpha)$ diberikan oleh persamaan (5.1) dengan $\hat{e}(\alpha)$ yang bersesuaian. Perlu diketahui bahwa $e(\alpha) = e_n(\alpha)$ tergantung pada ukuran sampel n. Untuk bootstrap residual perhitungan secara langsung menghasilkan

$$\hat{e}(\alpha) = \frac{2\hat{\sigma}^2 p_\alpha}{n} \qquad (5.4)$$

yang merupakan estimator tak bias asimptotis dan konsisten untuk $e(\alpha)$. Dan $\hat{e}(\alpha)$ dengan menggunakan bootstrap pasangan akan mendekati ruas kanan pada persamaan (5.4) tetapi penurunannya cukup rumit.

Lebih lanjut untuk mendapatkan prosedur pemilihan model bootstrap yang konsisten , dapat juga dipilih ukuran sampel m sedemikian hingga $e_m(\alpha)$ dapat diestimasi dengan $\hat{e}_m(\alpha)$ dengan $m/n \to 0$ dan kemudian meminimalkan :

$$\overset{\wedge}{\text{mse}}_{\text{BOOT-m}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\alpha)^2 + \hat{e}_m(\alpha) \text{ atas } \alpha \in \boldsymbol{A}. \qquad (5.5)$$

Untuk mengestimasi $e_m(\alpha)$ pada bootstrap pasangan , terlebih dahulu dibangkitkan sampel berpasangan $(y_1^*, x_1^*), (y_2^*, x_2^*), ..., (y_n^*, x_m^*)$ dan menggunakan :

$$\hat{e}_m(\alpha) = E_*\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_{i\alpha}^{*'}\hat{\boldsymbol{\beta}}_{m,\alpha}^*)^2 - \frac{1}{m}\sum_{i=1}^{m}(y_i^* - x_{i\alpha}^{*'}\hat{\boldsymbol{\beta}}_{m,\alpha}^*)^2\right] \qquad (5.6)$$

dengan $\hat{\beta}_{m,\alpha}^*$ didefinisikan seperti pada $\hat{\beta}_\alpha^*$ yang didasarkan pada m pasangan data bootstrap. Jika m dipilih sedemikian hingga $m/n \to 0$ dan $m \to \infty$ maka :

$$\hat{e}_m(\alpha) = \frac{2\hat{\sigma}^2 p_\alpha}{m} + o(m^{-1})$$

Pada bootstrap residual, digunakan secara langsung $\hat{e}_m(\alpha) = \frac{2\hat{\sigma}^2 p_\alpha}{m}$ dan mengestimasi $e_m(\alpha)$ dengan $\frac{2\hat{\sigma}^2 p_\alpha}{m}$.

6. KONSISTENSI BOOTSTRAP

Teorema 6.1 : (Shao dan Tu, 1995)

Jika diasumsikan bahwa ε_i i.i.d dan $\max_{i \le n} h_{i\alpha} \to 0$ untuk semua $\alpha \in \boldsymbol{A}$,

dengan $h_{i\alpha} = x_{i\alpha}'(\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}x_{i\alpha}$.

(i) Pandang suatu estimator bootstrap $\overset{\wedge}{\text{mse}}_{\text{BOOT}}(\alpha)$ dalam persamaan (5.1) dengan $\hat{e}(\alpha)$ seperti yang diberian pada persamaan (5.2) untuk PB dan (5.4) untuk RB. Maka, apabila α dalam kategori I (an incorrect model),

$$\overset{\wedge}{\text{mse}}_{\text{BOOT}}(\alpha) = \overline{\text{mse}}(\alpha) + o_p(1); \qquad (6.1)$$

sedangkan apabila α dalam kategori II (a correct model),

$$\overset{\wedge}{\text{mse}}_{\text{BOOT}}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{2\sigma^2 p_\alpha}{n} - \frac{\varepsilon' H_\alpha \varepsilon}{n} + o_p(n^{-1}).$$

(ii) Jika $\overset{\wedge}{\text{mse}}_{\text{BOOT}-\text{m}}(\alpha)$ yang didefinisikan dalam persamaan (5.4) dengan $\hat{e}_m(\alpha)$ seperti dalam persamaan (5.5) untuk PB dan $\frac{2\hat{\sigma}^2 p_\alpha}{m}$ untuk RB. Lebih lanjut ukuran sampel bootstrap m dipilih sedemikian hingga

$$\frac{m}{n} \to 0, \quad \frac{n}{m} \max_{i \le n} h_{i\alpha} \to 0 \text{ untuk semua } \alpha \in \boldsymbol{A} \qquad (6.2)$$

Maka apabila α dalam kategori I (an incorrect model),

$$\overset{\wedge}{\text{mse}}_{\text{BOOT}-\text{m}}(\alpha) = \overline{\text{mse}}(\alpha) + o_p(1);$$

sedangkan apabila α dalam kategori II (a correct model),

$$\overset{\wedge}{\text{mse}}_{\text{BOOT}-\text{m}}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{\sigma^2 p_\alpha}{m} + o_p(m^{-1}).$$

(iii) Lebih lanjut diasumsikan bahwa $\lim_n \inf \inf_{\alpha \text{ dlm kategori I}} \Delta(\alpha) > 0$.

Maka $\lim_{n\to\infty} P\{\hat{\alpha}_{\text{BOOT}} \text{ dalam kategori I}\} = 0$, dan $\lim_{n\to\infty} P\{\hat{\alpha}_{\text{BOOT}} = \alpha_0\} < 1$ kecuali bila $\alpha = \{1,2,...,p\}$ dalam kategori II (a correct model); dan $\hat{\alpha}_{\text{BOOT}-\text{m}}$ konsisten, yaitu $\lim_{n\to\infty} P\{\hat{\alpha} = \alpha_0\} = 1$ berlaku untuk $\hat{\alpha}_{\text{BOOT}-\text{m}}$, dengan $\hat{\alpha}_{\text{BOOT}}$ dan $\hat{\alpha}_{\text{BOOT}-\text{m}}$ adalah model-model terpilih dengan masing-masing meminimalkan $\overset{\wedge}{\text{mse}}_{\text{BOOT}}(\alpha)$ dan $\overset{\wedge}{\text{mse}}_{\text{BOOT}-\text{m}}(\alpha)$.

7. SIMULASI

Sebagai suatu implementasi secara praktis dilakukan simulasi terhadap "data semen" yang diambil dari (Hjorth, 1994) dengan menggunakan metode bootstrap residual dan pasangan data dengan replikasi bootstrap B=200 dan B=400, memberikan hasil seperti terlihat pada Tabel.1 dibawah ini. Adapun perhitungan estimasi rata-rata sesatan prediksi kuadrat dilakukan dengan program SPLUS. Variabel prediktor yang terlibat didalam model sebanyak 4 variabel, sehingga model yang mungkin seluruhnya ada 15 model. Dari 15 model tersebut

akan dipilih satu model terbaik untuk masing-masing ukuran dan kemudian dari 4 model terbaik tersebut dipilih satu model yang memiliki $\overline{\wedge}$ mse yang terkecil.

Tabel.1: Estimasi $\overline{mse}$ untuk 15 model yang mungkin

No	Variabel-variabel dalam Model				mse RB		mse-PB n=13		mse-PB m=6	
	x1	x2	x3	x4	B=200	B=400	B=200	B=400	B=200	B=400
1	x1				119,28	119,14	110,18	110,73	123,17	122,95
2		x2			85,32	85,11	78,82	78,96	89,51	90,14
3			x3		183,15	183,18	169,26	168,58	190,98	190,25
4				x4	**83,29**	**83,26**	**77,04**	**77,30**	**86,87**	**86,74**
5	x1	x2			**5,69**	**5,71**	**5,25**	**5,25**	**5,88**	**5,81**
6	x1		x3		120,03	120,24	111,08	110,80	122,92	121,43
7	x1			x4	7,30	7,35	6,80	6,79	7,37	7,52
8		x2	x3		40,68	41,09	37,61	37,50	41,73	41,44
9		x2		x4	84,62	85,61	78,07	78,41	85,27	85,11
10			x3	x4	17,16	17,26	15,84	15,93	17,70	17,42
11	x1	x2	x3		4,89	4,86	4,57	4,53	4,59	4,52
12	x1	x2		x4	**4,82**	**4,83**	**4,49**	**4,52**	**4,58**	**4,47**
13	x1		x3	x4	5,12	5,16	4,84	4,78	4,85	4,89
14		x2	x3	x4	7,42	7,46	6,91	6,96	7,09	7,27
15	x1	x2	x3	x4	**4,95**	**4,90**	**4,61**	**4,64**	**4,14**	**4,02**

Dari Tabel diatas diperoleh model terbaik untuk masing-masing ukuran sebagai berikut.

- Model terbaik dengan 1 prediktor: y = 117,57 – 0,74 x4
- Model terbaik dengan 2 prediktor: y = 52,58 + 1,47 x1 + 0,66 x2
- Model terbaik dengan 3 prediktor: y = 71,65 + 1,45 x1 + 0,42 x2 – 0,24 x4
- Model terbaik dengan 4 prediktor: y = 62,41+1,55x1+0,51x2+0,10x3– 0,144x4.

Dengan mempertimbangkan sesatan prediksi dan prinsip parsimonius (melibatkan variabel prediktor sesedikit mungkin) dapat dipilih satu model terbaik dari 4 model tersebut yaitu :

y = 71,65 + 1,45 x1 + 0,42 x2 – 0,24 x4 .

8. KESIMPULAN

Model regresi terbaik merupakan model regresi yang memiliki kemampuan prediksi terbaik, yaitu memiliki estimasi rata-rata sesatan prediksi kuadrat minimum dengan melibatkan variabel prediktor sesedikit mungkin. Dari hasil simulasi dengan menggunakan metode bootstrap residual dan bootstrap data berpasangan diperoleh model terbaik untuk masing-masing ukuran adalah sama, baik untuk ukuran sampel bootstrap n maupun m.

DAFTAR PUSTAKA

1. Efron, B, *Bootstrap Methods : Another Look at Jackknife, Annals of Statistics,* 1979, 7 : 1 - 26.

2. Efron B and Tibshirani, *An Introduction to Bootstrap,* Chapman and Hall, New York, 1993.

3. Hjorth J. S. U, *Computer Intensive Statistical Methods, Validation Model Selection and Bootstrap,* Chapman and Hall, New York, 1994.

4. Searle S. R, *Linier Models*, John Wiley and Sons, New York, 1971.

5. Serfling R. J, *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.

6. Shao J, *Linier Model Selection by Cross-Validation, Journal American Statistics Assosiation*, 1993, 88 : 486 - 494.

7. Shao J dan Tu D, *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995.

8. Shao J, *Bootstrap Model Selection, Journal American Statistics Assosiation*, 1996, 9 : 655 - 665.