

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 $m \rightarrow \infty$, pemilihan variabel bootstrap juga konsisten. Hasil dari suatu simulasi dengan SPLUS disajikan dalam tulisan ini.

Kata kunci : pemilihan model, bootstrap dan sesatan prediksi.

1. PENDAHULUAN

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

[]

(1.1)

[]

dengan y_i adalah respon ke- i , \mathbf{x}_i : p -vektor variabel prediktor yang berkaitan dengan y_i , β : 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 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, diasumsikan

[]

independen dan berdistribusi identik dengan momen kedua berhingga dan menyatakan variansi bersyarat dari y_i diberikan x_i .

[]

Parameter regresi 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 mungkin sama dengan nol, maka model yang lebih kompak memiliki bentuk:

[]

(1.2)

[]

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

[]

Metode bootstrap memilih model dengan meminimalkan estimasi rata-rata jumlah kuadrat

[]

dari sesatan prediksi atas semua 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 yang sama dengan nol (Shao, Tu, 1996). Oleh karena itu, didapatkan model yang lebih kompak yang berbentuk :

[]

(2.1)

[]

[]

[]

[]

[]

dengan . Jika sebagai subvektor yang memuat komponen-komponen dari dan x_i berada dalam , maka terdapat (2^p-1) model berbeda yang mungkin yang berbentuk (2.1), masing-masing terkait

[]

[]

[]

dengan suatu himpunan bagian dan dinotasikan dengan . Dimensi (ukuran) dari adalah

[]

banyaknya prediktor dalam (Shao dan Tu, 1995).

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

- Kategori I (an incorrect model) : minimal satu komponen dari β yang tidak nol tidak berada dalam A .

- Kategori II (a correct model) : memuat semua komponen β yang tidak nol.

Model optimal adalah model (2.1) dengan sedemikian hingga memuat semua komponen β yang semuanya tidak nol (model dalam kategori II dengan dimensi terkecil). Model optimal tersebut

tidak diketahui karena β tidak diketahui, sehingga perlu memilih model dari (2.1) berdasarkan data $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ yang memenuhi (1.1). Jika diasumsikan bahwa ϵ_i i.i.d dengan mean 0 dan

variansi σ^2 , maka dibawah model (2.1), dengan Estimator Kuadrat Terkecil diperoleh:

dengan $\mathbf{y}=(y_1, y_2, \dots, y_n)'$ dan $\mathbf{X}=(x_1, x_2, \dots, x_n)'$.

Jika dianggap bahwa \mathbf{y}_f : nilai respon yang akan datang untuk suatu nilai prediktor \mathbf{x}_f ,

maka $\hat{\mathbf{y}}_f$ berakibat bahwa mean sesatan prediksi kuadrat adalah :

Jika \mathbf{A} dalam kategori II maka $\hat{\beta}$ dan $\hat{\mathbf{y}}_f$. Dengan demikian jika diketahui \mathbf{A} , maka model optimal dapat

dipilih dengan meminimalkan $\sum_{i \in A} \hat{\epsilon}_i^2$ atas semua \mathbf{A} . Model optimal dapat juga ditentukan dengan

meminimalkan rata-rata dari sesatan prediksi kuadrat atas $X = \{x_1, x_2, \dots, x_n\}$:

dengan dan

Namun, dan kedua-duanya tidak diketahui. Sehingga mengestimasi lebih mudah dari pada

mengestimasi dengan menggunakan , kemudian memilih model dengan meminimalkan atas A .

3. ESTIMASI PARAMETER

Jika parameter merupakan parameter regresi yang akan diestimasi dengan , maka di lingkungan

bootstrap dapat diestimasi dengan . 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) :

dengan y_i adalah respon ke- i , x_i : p -vektor variabel prediktor yang berkaitan dengan y_i , p -

vektor parameter yang tidak diketahui dan : sesatan random, dan apabila x_i nonrandom, dengan

asumsi bahwa i.i.d dengan mean 0 dan variansi . Maka untuk memperoleh estimasi parameter dapat dilakukan prosedur bootstrap sebagai berikut :

[]
[]

a. Model regresi (1.1) diidentifikasi sebagai model parameter, yaitu (β_0, β_1) , dengan ϵ_i merupakan distribusi dari ϵ_i yang tak diketahui.

[]
[]
[]

b. Dengan metode kuadrat terkecil, parameter β_0 diestimasi dengan $\hat{\beta}_0$ dan β_1 diestimasi dengan fungsi

[]
[]
[]

distribusi empiris dengan mengambil massa peluang n^{-1} terhadap ϵ_i dengan ϵ_i merupakan residual ke- i

[]
[]

dan $\hat{\beta}_0$ dipusatkan pada 0 karena mempunyai mean 0.

[]
[]

c. Data bootstrap dibangkitkan dari model tersebut dengan ϵ_i diganti dengan ϵ_i^* . Dengan kata lain

[]
[]

dibangkitkan data independen dan berdistribusi identik dari ϵ_i dan didefinisikan

[]
[]

d. Dengan metode kuadrat terkecil, dihitung berdasarkan data $\{y_i, x_i\}$, yaitu :

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

[]

Karena maka

[]

(3.1)

dan juga,

[]

(3.2)

[]

dengan .

Dari persamaan (3.1) dan (3.2), prosedur bootstrap ini menghasilkan estimator variansi dan

[]

estimator bias dari yang konsisten.

[]

Suatu asumsi yang penting untuk bootstrap residual adalah . Bahkan jika asumsi ini

[]

berlaku fungsi distribusi empiris tidak didasarkan pada data i.i.d secara eksak. Estimator (3.2) mempunyai bentuk secara eksplisit tetapi tidak sama dengan estimator yang diberikan oleh:

[]

[]

$$\text{var}() = \quad (3.3)$$

[]

dengan merupakan residual ke-i.

[]

[]

Estimator dalam persamaan (3.2) bias menurun, karena , dengan . Untuk menghilangkan bias negatif ini, Shao dan Tu (1995) menyatakan bahwa data bootstrap diambil dari fungsi distribusi

[]

empiris berdasarkan . Ketentuan ini masih mengarah kepada suatu estimator variansi yang bias

[]

menurun, karena $k_n > p$ bilamana $=0$. Kemudian data bootstrap dibangkitkan dari fungsi distribusi

[]

[]

empiris dengan mengambil massa peluang n^{-1} terhadap residual yang teratur . Maka (3.2)

[]

berlaku dengan diganti dengan :

[]

[]

[]

[]

Jika $=0$ maka $k_n = p$, dan estimator variansi bootstrap untuk 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 x_i random dan , 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 dan diestimasi dengan fungsi distribusi

empiris dengan massa peluang n^{-1} untuk setiap .

b. Dibangkitkan data bootstrap dari fungsi distribusi empiris ini, yaitu :

.

c. Dengan metode kuadrat terkecil ditentukan estimasi parameter regresi :

d. .

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 i.i.d, suatu prediksi titik untuk y_f adalah :

dan ini dapat dievaluasi dengan mean sesatan prediksi kuadrat,

.

Jika y_f dan y_1, y_2, \dots, y_n independen, maka

.

Suatu estimator dari $mse(x_f)$, berdasarkan bootstrap residual (RB) dapat diperoleh sebagai berikut.

Jika yaitu fungsi distribusi empiris dari residual yang diatur . Jika suatu nilai respon bootstrap

yang akan datang dan prediksi bootstrap dari . Maka estimator bootstrap untuk adalah:

[Empty box]

[Empty box]

Karena $\hat{\beta}$ merupakan estimator tak bias.

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

[Empty box]

Berdasarkan bootstrap pasangan Efron (Shao dan Tu,1995) mengusulkan estimator bootstrap

[Empty box]

untuk β . Didefinisikan sesatan ekspekstasi dengan

[Empty box]

dan estimator bootstrapnya dengan

[Empty box]

Dengan demikian estimator bootstrapnya adalah:

[Empty box]

(4.1)

Dan nilai harapannya adalah :

[Empty box]

Jadi estimator bootstrap tersebut hampir tak bias.

5. PEMILIHAN VARIABEL

[Empty box]

Prosedur pemilihan model dengan bootstrap dapat diturunkan dari estimator untuk β . Pandang estimator bootstrap berbentuk :

Teorema 6.1 : (Shao dan Tu, 1995)

Jika diasumsikan bahwa $i.i.d$ dan A ,

dengan .

i) Pandang suatu estimator bootstrap dalam persamaan (5.1) dengan seperti yang diberikan pada persamaan (5.2) untuk PB dan (5.4) untuk RB. Maka, apabila dalam kategori I (an incorrect model),

(6.1)

sedangkan apabila dalam kategori II (a correct model),

ii) Jika yang didefinisikan dalam persamaan (5.4) dengan seperti dalam persamaan (5.5) untuk

PB dan untuk RB. Lebih lanjut ukuran sampel bootstrap m dipilih sedemikian hingga

A (6.2)

Maka apabila dalam kategori I (an incorrect model),

sedangkan apabila dalam kategori II (a correct model),

[Redacted]

[Redacted]

iii) Lebih lanjut diasumsikan bahwa [Redacted]

[Redacted]

Maka, dan [Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted]

kecuali bila dalam kategori II (a correct model); dan konsisten, yaitu berlaku untuk , dengan

[Redacted]

[Redacted]

dan adalah model-model terpilih dengan masing-masing meminimalkan dan .

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

[Redacted]

model yang memiliki yang terkecil.

[Redacted]

Tabel.1: Estimasi 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	1 119,28 119,14 110,18 110,73 123,17 122,95
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: $y = 117,57 - 0,74 x_4$
- Model terbaik dengan 2 prediktor: $y = 52,58 + 1,47 x_1 + 0,66 x_2$
- Model terbaik dengan 3 prediktor: $y = 71,65 + 1,45 x_1 + 0,42 x_2 - 0,24 x_4$
- Model terbaik dengan 4 prediktor: $y = 62,41 + 1,55x_1 + 0,51x_2 + 0,10x_3 - 0,144x_4$.

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

$$y = 71,65 + 1,45 x_1 + 0,42 x_2 - 0,24 x_4 .$$

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.