BOOTSTRAP PADA ANALISIS REGRESI

Bootstrap on Regression Analysis

I Ketut Tunas[1] dan Subanar[2]

Program Studi Matematika
Program Pasca Sarjana Universitas Gadjah Mada

ABSTRACT

Considered a regression model $Y(n) = X(n)\beta + \varepsilon(n)$, where β is a px1 vektor of unknown parameters; $\varepsilon(n)$ is an nx1 unobservable vektor of random errors. The problem in regression analysis is how to estimate the parameter vektor β based on the observable data.

For the regression model, let $\hat{\beta}^*(m)$ be the least squares estimate based on the resample: $\hat{\beta}^*(m) = \{X(m)^T X(m)\}^{-1} X(m)^T Y^*(m)$. The distribution of $\sqrt{m}(\hat{\beta}^*(m)-\hat{\beta}(n))$, which can be computed directly from the data consistent for the distribution of $\sqrt{n}(\hat{\beta}(n)-\beta)$. For the correlation model, the least squares estimate is $\hat{\beta}^*(m) = \{X^*(m)^T X^*(m)\}^{-1} X^*(m)^T Y^*(m)$ and the law of $\sqrt{m}(\hat{\beta}^*(m)-\hat{\beta}(n))$ close to the law of $\sqrt{n}(\hat{\beta}(n) - \beta)$. It is shown that under mild conditions, the bootstrap approximation to the distribution of the least squares estimates is valid. The bootstrap give the same asymtotic result as normal approximation.

Key Words: Regression, correlation, least squares estimation, bootstrap.

PENDAHULUAN

Pada model regresi

$$Y(n) = X(n)\beta + \varepsilon(n), \qquad (1)$$

β merupakan vektor parameter tak diketahui dan berukuran p x 1, yang diestimasi dari data; Y(n) vektor data n x 1; X(n) matrik data bertipe n x p

[1] FKIP Universitas Pattimura, Ambon.
[2] Fakultas MIPA Universitas Gadjah Mada, Yogyakarta.

dengan rank penuh $p \leq n$; $\varepsilon(n)$ vektor sesatan random tak terobservasi dengan ukuran nx1. Permasalahan dalam analisis regresi adalah bagaimana mengestimasi parameter β, yang diperoleh dari sampel. Masalah ini bisa diatasi dengan cara mencari estimator yang konsinten dan sahih, dan biasanya parameter β diestimasi dengan metode kuadrat terkecil, yaitu $\hat{\beta}(n)$. Tetapi, seberapa kedekatannya terhadap β. Oleh karena itu, diperlukan suatu estimator untuk deviasi ini. Jadi permasalahan sekarang ialah mengestimasi variabilitas statistik $\hat{\beta}(n)$-β. Salah satu metode untuk mengestimasi adalah penyampelan kembali pada sampel aslinya atau yang lebih dikenal dengan metode bootstrap.

Metode bootstrap pertama kali diperkenalkan oleh Efron (1979). Pendekatan regresi bootstrap banyak dibahas oleh Freedman (1981), Bickel dan Freedman (1983), dan Wu (1986). Bickel dan Freedman (1981) membahas teori asimtotis bootstrap

Berdasarkan uraian di atas, maka dalam tulisan ini akan dibahas konsintensi dan konvergensi metode bootstrap dan perbandingannya dengan pendekatan normal.

Freedman (1981) membahas dua model yang berbeda, yaitu " model regresi " dan " model korelasi ". Asumsi dasar yang diperlukan,

Matrik X(n) tidak random. (2)

Konponen ε_1 , ε_2 , . . . , ε_n dari ε (n) independen, dengan distribusi bersama F, mempunyai mean 0 dan variansi berhingga σ^2; F dan σ^2 tidak diketahui. (3)

$$\frac{1}{n} X(n)^T X(n) \rightarrow V \text{ definite positif.} \quad (4)$$

Untuk model korelasi, baris ke-i data (X,Y) dinotasikan dengan (X_i,Y_i); jadi (X_i, Y_i) merupakan vektor baris random berdimensi $p + 1$.

Vekotr (X_i , Y_i) diasumsikan independen, dengan distribusi bersama μ dalam R^{p+1}; $E\{\|(X_i,Y_i)\|^4\} < \infty$. dengan $\| \cdot \|$ panjang Euclidean. (5)

PENGERTIAN DASAR

Model regresi dalam persamaan (1), β merupakan vektor parameter tidak diketahui berukuran $p \times 1$, yang diestimasi dari data. Y(n) vektor $n \times 1$, X(n) matrik $n \times p$ dengan rank penuh $p \leq n$, $\varepsilon(n)$ vektor $n \times 1$ dan diasumsikan berdistribusi independen, dengan mean nol dan variansi berhingga σ^2.

Salah satu metode untuk memperoleh suatu estimasi vektor paramater β adalah metode kuadrat terkecil dan dengan metode ini diperoleh persamaan normal $X^TX\,\hat{\beta} = X^TY$. Dengan asumsi bahwa X matrik bertipe *n x p* dengan rank p, X^TX matrik definite positif, sehingga X^TX merupakan matrik non singular. Akibatnya persamaan normal di atas mempunyai penyelesaian tunggal $\hat{\beta} = (X^TX)^{-1}\,X^TY$. Estimasi kuadrat terkecil $\hat{\beta}$ merupakan estimator tak bias linear untuk β dengan matrik varian-kovarian $\sigma^2(X^TX)^{-1}$.

Teorema

Dengan asumsi (1)-(4), $\sqrt{n}\,(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 V^{-1})$.

Bukti

$\hat{\beta} - \beta = [\frac{1}{n}X^TX]^{-1}.\frac{1}{n}X^T\varepsilon \xrightarrow{p} 0$. Berdasarkan teorema limit pusat multivariat maka $X^T\varepsilon/\sqrt{n} \xrightarrow{d} N(0, \sigma^2 V)$. Akibatnya $\sqrt{n}\,(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 V^{-1})$. □

Teorema

Jika $\Sigma = E(X_i^TX_i)$ dan M didefinisikan dengan $M_{jk} = E(X_{ij}X_{ik}\,\varepsilon_i^2)$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma^{-1}M\Sigma^{-1}).$$

Bukti

Dengan teorema limit pusat diperoleh $X^T\varepsilon/\sqrt{n} \xrightarrow{d} N(0, M)$, dan $\frac{1}{n}X^TX \rightarrow \Sigma$, akibatnya teorema terbukti.

Misalkan $X_1, X_2, \ldots, X_n$ (i.i.d) berdistribusi *F*. Distribusi empiris $\hat{F}_n$ merupakan suatu fungsi dari observasi yang memberikan probabilitas sama, yaitu 1/n untuk setiap nilai atau observasi; $\hat{F}_n$ merupakan estimator *F*, yang didefinisikan dengan :

$$\hat{F}_n(x) = \frac{1}{n}\sum I(X_i \le x), \quad \text{untuk semua nilai real x,}$$

dengan $I(X_i \le x)$ fungsi indikator, yaitu:

$$I(X_i \le x) = \begin{cases} 1 & \text{untuk } X_i \le x \\ 0 & \text{untuk } X_i > x \end{cases}$$

Bootstrap adalah prosedur resampling untuk mengestimasi distribusi probabilitas suatu statistik. Bootstrap diperkenalkan oleh Bradley Efron pada tahun 1979. Prinsip dasar pembentukan sampel dengan metode bootstrap sebagai berikut :

1. Konstruksi distribusi probabilitas sampel, yaitu $\hat{F}_n$, dengan massa $\frac{1}{n}$ pada setiap titik $x_1, x_2, \ldots, x_n$.
2. Dengan $\hat{F}_n$ tetap, ambil sampel random berukuran n dari F_n, sebut
$$X_i^* = x_i^*, \; X_i^* \sim_{ind} F_n \qquad i = 1, 2, \ldots, n.$$
Sebut sampel ini dengan *sampel bootstrap*, $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$, $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$. Dalam hal ini tidak didapatkan distribusi permutasi sebab nilai dari X^* dipilih atau diambil dengan pengembalian dari $\{x_1, x_2, \ldots, x_n\}$.
3. Hampiri distribusi sampling $R(X,F)$ dengan distribusi bootstrap $R^*(X^*,F_n)$.

Prosedur bootstrap untuk estimasi

(1) Estimasi F dengan $\hat{F}_n$ dan hitung $\theta_n = \theta(\hat{F}_n)$.
(2) Diberikan $X_1, X_2, \ldots, X_n$, Anggap $X_1^*, X_2^*, \ldots, X_n^*$ suatu sampel i.i.d. dengan distribusi $\hat{F}_n$.
(3) Ambil $Y_n^* = b_n(T_n(X_1^*, X_2^*, \ldots, X_n^*) - \theta_n)$ versi bootstrap dari Y_n.
(4) Distribusi $F(Y_n)$ diestimasi dengan $\hat{F}_n(Y_n^*)$.

REGRESI BOOTSTRAP

Bootstrapping

Ada dua metode yang berbeda untuk bootstrapping model regresi. Pertama bootstrapping residu, dengan himpunan data bootstrap berbentuk

$$Z^* = \{(X_1, X_1\hat{\beta} + \varepsilon_i^*), (X_2, X_2\hat{\beta} + \varepsilon_i^*), \ldots, (X_n, X_n\hat{\beta} + \varepsilon_i^*)\},$$

metode ini disebut dengan model regresi. Sedangkan metode kedua bootstrapping pasangan (X_i, Y_i), dan himpunan data bootstrap berbentuk

$$Z^* = \{(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_n^*, Y_n^*)\},$$

untuk selanjutnya metode ini disebut dengan model korelasi.
Diberikan Y(n), misalkan $\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*$ dikondisikan independen, dengan distribusi bersama $\hat{F}_n$; $\varepsilon^*(n)$ n-vektor yang komponen ke i-nya ε_i^* dan

$$Y^*(n) = X(n)\hat{\beta}(n) + \varepsilon^*(n). \qquad (6)$$

Dengan diberikannya data berbintang (X,Y*), maka akan diestimasi vektor parameter. Estimasi kuadrat terkecil:

$$\hat{\beta}^*(n) = (X(n)^T X(n))^{-1} X(n)^T Y^*(n). \quad (7)$$

Prinsip dasar pendekatan bootstrap, distribusi $\sqrt{n}\,(\hat{\beta}^*(n) - \hat{\beta}(n))$, yang dapat dihitung langsung dari data, merupakan pendekatan distribusi $\sqrt{n}\,(\hat{\beta}(n) - \beta)$.

Pandang sekarang model korelasi, pada umumnya ada suatu ketergantungan antara ε_i dan X_i. Ambil μ_n distribusi empiris (X_i, Y_i) untuk i = 1, 2, . . . , n. Oleh karena itu μ_n merupakan suatu probabilitas dalam R^{p-1}, dengan massa $\frac{1}{n}$ pada setiap vektor (X_i, Y_i). Diberikan $\{X(n), Y(n)\}$, anggap (X_i^*, Y_i^*) independen, dengan distribusi bersama μ_n, untuk i = 1, 2, . . . , m. Jika $\hat{\beta}^*(m)$ estimasi kuadrat terkecil yang berdasarkan pada resampel, maka

$$\hat{\beta}^*(m) = \{X^*(m)^T X^*(m)\}^{-1} X^*(m)^T Y^*(m).$$

Model Regresi

Asumsikan model regresi

1. $Y(n) = X(n)\beta + \varepsilon(n)$.
2. Matriks X(n) tidak random.
3. Komponen $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ dari $\varepsilon(n)$ independen, dengan distribusi bersama F, mempunyai mean 0 dan variansi berhingga σ^2; F dan σ^2 tidak diketahui.

Anggap $\psi_n(F)$ distribusi $\sqrt{n}\,(\hat{\beta}(n) - \beta)$, dengan F distribusi ε. Jadi $\psi_n(F)$ merupakan probabilitas $\in R^p$. Misalkan G distribusi alternatif untuk ε: asumsikan G juga mempunyai mean 0 dan varian berhingga σ^2.

Teorema

$$d_2{}^p\{\psi_n(F), \psi_n(G)\}^2 \le n \cdot \text{trace}\,\{X(n)^T X(n)\}^{-1} \cdot d_2(F, G)^2.$$

Bukti

$\psi_n(F)$ distribusi $\sqrt{n}\,(\hat{\beta}(n) - \beta) = \sqrt{n}\,\{X(n)^T X(n)\}^{-1} X(n)^T \varepsilon(n)$, dengan $\varepsilon(n)^T = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, ε_i variabel random independen dengan distribusi bersama F. Dengan cara yang sama untuk G. Jadi $d_2{}^p\{\psi_n(F), \psi_n(G)\}^2 = n \cdot \text{trace}\,\{X(n)^T X(n)\}^{-1} \cdot d_2(F, G)^2$ □

Misalkan F_n fungsi distribusi empiris $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$; $\tilde{F}_n$ distribusi empiris residu $\hat{\varepsilon}_1(n)$, $\hat{\varepsilon}_2(n)$, $\cdots$, $\hat{\varepsilon}_n(n)$ dari regresi original pada n

vektor data, dan $\hat{F}_n$ adalah $\tilde{F}_n$ yang terpusat pada meannya $\hat{\mu}_n = \frac{1}{n}\sum_i^n \hat{\varepsilon}_i(n)$.

Diberikan barisan $X_1, X_2, \ldots$, misalkan X(n) merupakan n pertama barisan tersebut, yaitu $X_1, X_2, \ldots, X_n$. Demikian juga untuk $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ n pertama barisan tak hingga variabel random independen dengan fungsi distribusi bersama F. Dalam hal ini, dilakukan resampel berukuran m untuk membedakannya dengan n.

Ambil $\hat{\beta}(n)$ estimator untuk β, berdasarkan atas n data pertama. Data berbintang diberikan oleh

$$Y^*(m) = X(m)\hat{\beta}(n) + \varepsilon^*(m).$$

dengan Y*(m) vektor berukuran *mx1*, X(m) matrik bertipe *mxp*, $\hat{\beta}(n)$ vektor berukuran *px1*, dan ε*(m) vektor berukuran *mx1*. $\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_m^*$, independen dengan distribusi bersama $\hat{F}_m$, distribusi empiris residu himpunan data asli, yang terpusat pada mean μ_n. $\hat{\beta}^*(m)$ estimator parameter yang berdasarkan pada data berbintang.

$$\hat{\beta}^*(m) = [X(m)^T X(m)]^{-1} X(m)^T Y^*(m).$$

Residu berbintang

$$\hat{\varepsilon}^*(m) = Y^*(m) - X(m)\hat{\beta}^*(m).$$

Lemma

$d_2(\hat{F}_n, F) \to 0$ hampir di mana-mana

Bukti

Karena d_2 metrik, maka $d_2(\hat{F}_n, F) \le d_2(\hat{F}_n, F_n) + d_2(F_n, F) \to 0$ hampir di mana-mana. □

Teorema

Asumsikan model regresi, dengan (1)-(4). Untuk seluruh barisan sampel $X_1, X_2, \ldots$, diberikan $X_1, X_2, \ldots, X_n$, dengan m dan n menuju ∞.

(a) Distribusi $\frac{1}{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ konvergen secara lemah ke distribusi normal dengan mean 0 dan matrik varian-kovarian $\sigma^2 V^{-1}$.

(b) Distribusi $\hat{\sigma}_m^*$ konvergen ke massa titik pada σ.

(c) Distribusi $\{X(m)^T X(m)\}^{1/2}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}/\hat{\sigma}_m^*$ konvergen ke distribusi normal standar dalam R^p.

Bukti

(a) Dari Teorema di atas ganti n dengan m dan G dengan $\hat{F}_n$. Selanjutnya

$$d_2^p \{\psi_n(F), \psi_n(\hat{F}_n)\}^2 \le m \cdot \text{trace}\{X(m)^T X(m)\}^{-1} \cdot d_2(F_n, \hat{F}_n)^2$$
$$= \text{trace}[(1/m)\, X(m)^T X(m)]^{-1} \cdot d_2(F, \hat{F}_n)^2$$
$$\to 0 \cdot \text{hampir di mana-mana.}$$

Selanjutnya diperoleh $\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ konvergen secara lemah ke distribusi normal dengan mean 0 dan matrik varian-kovarian $\sigma^2 V^{-1}$.

(b) $\hat{\sigma}_n^2 = \frac{1}{n}\sum \hat{\varepsilon}_i(n)^2 - \mu_n^2$; $\mu_n = \frac{1}{n}\sum \hat{\varepsilon}_i(n)$ dan $\sigma_n^2 = \frac{1}{n}\sum \varepsilon_i^2 - [\frac{1}{n}\sum \varepsilon_i]^2$,

maka $(\hat{\sigma}_n - \sigma_n)^2 \le \frac{1}{n}\| \hat{\varepsilon}(n) - \varepsilon(n) \|^2 \to 0$ hampir di mana-mana.

Ambil $\sigma_m^{*2} = \frac{1}{m}\sum \varepsilon_i^{*2} - [\frac{1}{m}\sum \varepsilon_i^*]^2$. Selanjutnya

$$E(|\hat{\sigma}_m^* - \sigma_m^*| \,|\, Y_1, Y_2, \ldots, Y_n)^2$$
$$\le E(\hat{\sigma}_m^* - \sigma_m^*)^2 \,|\, Y_1, Y_2, \ldots, Y_n\}$$
$$\le E[\frac{1}{m}\sum\{\hat{\varepsilon}_i^*(m) - \varepsilon_i^*\}^2 \,|\, Y_1, Y_2, \ldots, Y_n]$$
$$= \frac{1}{m} p\, \sigma_n^2 \to 0 \quad \text{hampir di mana-mana.}$$

Karena $d_1\{\frac{1}{m}\sum \varepsilon_i^{*2}, \frac{1}{m}\sum \varepsilon_i^2\} \le d_1(\varepsilon_1^{*2}, \varepsilon_i^2)$, akibatnya distribusi $\hat{\sigma}_m^{*2}$ mendekati massa titik pada σ^2.

(c) Karena $\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ konvergen secara lemah ke $N(0, \sigma^2 V^{-1})$.

$\hat{\sigma}_m^*$ konvergen ke σ, dan $\frac{1}{m} X(m)^T X(m) \to V$, maka

$$\{X(m)^T X(m)\}^{1/2}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}/\hat{\sigma}_m^* \to N(0, I_{p \times p}).$$

Jadi distribusi $\{X(m)^T X(m)\}^{1/2}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}/\hat{\sigma}_m^*$ konvergen ke distribusi normal standar dalam R^p. □

Model Korelasi

Sekarang Ambil μ_n dan μ probabilitas pada R^{p+1}, dengan pangkat empat dari norm Euclidean terintegral. Suatu titik dalam R^{p+1} ditulis (x,y), dengan $x \in R^p$, sebagai vektor baris dan $y \in R^1$. Asumsikan

$$\Sigma(\mu) = \int x^T x \, \mu(dx, dy) \quad \text{definite positif,}$$

dan bentuk

$$\beta(\mu) = \Sigma(\mu)^{-1} \int x^T y \, \mu(dx, dy);$$
$$\varepsilon(\mu, x, y) = y - x\beta(\mu).$$

Diberikan $\{X(n), Y(n)\}$, vektor resampel $(X_i^*, Y_i^*)'$ independen, dengan distribusi bersama μ_n, untuk $i = 1, 2, \ldots, m$. Misalkan $X^*(m)$ matrik $m \times p$ dengan baris ke-i nya X_i^*; dan $Y^*(m)$ vektor kolom Y_i^* berukuran $mx1$. $\hat{\beta}(n)$ Estimasi kuadrat terkecil berdasarkan data asli; pada data berbintang, $\hat{\beta}^*(m)$; residu terobservasi

$$\hat{\varepsilon}(n) = Y(n) - X(n)\hat{\beta}(n).$$

ε^* vektor kolom sesatan untuk data berbintang, dengan

$$\varepsilon_i^* = Y_i^* - X_i^*(m)\hat{\beta}(n),$$

sedangkan vektor kolom residu, $\hat{\varepsilon}^*(m)$, dengan

$$\hat{\varepsilon}_i^*(m) = Y_i^* - X_i^*(m)\hat{\beta}^*(m)$$

Ambil $\Sigma = E(X_i^T X_i)$, matrik varian-kovarian $p \times p$. Asumsikan

$$\Sigma \text{ definite positif.} \tag{8}$$

Definisikan matrik definite nonnegatif M dengan

$$M_{jk} = E(X_{ij} X_{ik} \varepsilon_i^2) \tag{9}$$

Teorema

Asumsikan model korelasi dengan kondisi (5), (8), dan (9). Untuk seluruh barisan sampel $(X_1, Y_1), (X_2, Y_2), \ldots$, diberikan (X_i, Y_i) untuk $1 \le i \le n$, dengan m dan n menuju tak hingga.

(a) $\frac{1}{m} X^*(m)^T X^*(m)$ konvergen ke Σ dalam probabilitas.

(b) Distribusi $\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ konvergen secara lemah ke normal dengan mean 0 dan matrik varian-kovarian $\Sigma^{-1}M\Sigma^{-1}$.

Bukti

Mengingat $\hat{\beta}^*(m)$ dan $\hat{\beta}(n)$, maka

$$\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\} = W^*(m)^{-1} Z^*(m)$$

dengan $$W^*(m) = \frac{1}{m} X^*(m)^T X^*(m) = \frac{1}{m} \Sigma X_i^{*T} X_i^*$$

dan $$Z^*(m) = \frac{1}{\sqrt{m}} X^*(m)^T \varepsilon^*(m) = \frac{1}{\sqrt{m}} \Sigma X_i^{*T} \varepsilon_i^*.$$

Disini W*(m) matrik $p \times p$; sedangkan Z*(m) vektor kolom $p \times 1$. Kuantitas tanpa bintang yang bersesuaian dengannya diberikan oleh

$$W(m) = \frac{1}{m} X(m)^T X(m) = \frac{1}{m} \sum X_i^T X_i$$

dan $$Z(m) = \frac{1}{\sqrt{m}} X(m)^T \varepsilon(m) = \frac{1}{\sqrt{m}} \sum X_i^T \varepsilon_i.$$

Selanjutnya W*(m) merupakan suatu jumlah vektor dalam $R^{p \times p}$; kondisikan W*(m) pada {X(n), Y(n)}.

$$\therefore d_1^{p \times p} \{W^*(m), W(m)\} = d_1^{p \times p} \{X_i^{*T} X_i^*, X_i^T X_i\}.$$

Terbukti bahwa

Distribusi W*(m) terpusat dekat Σ.

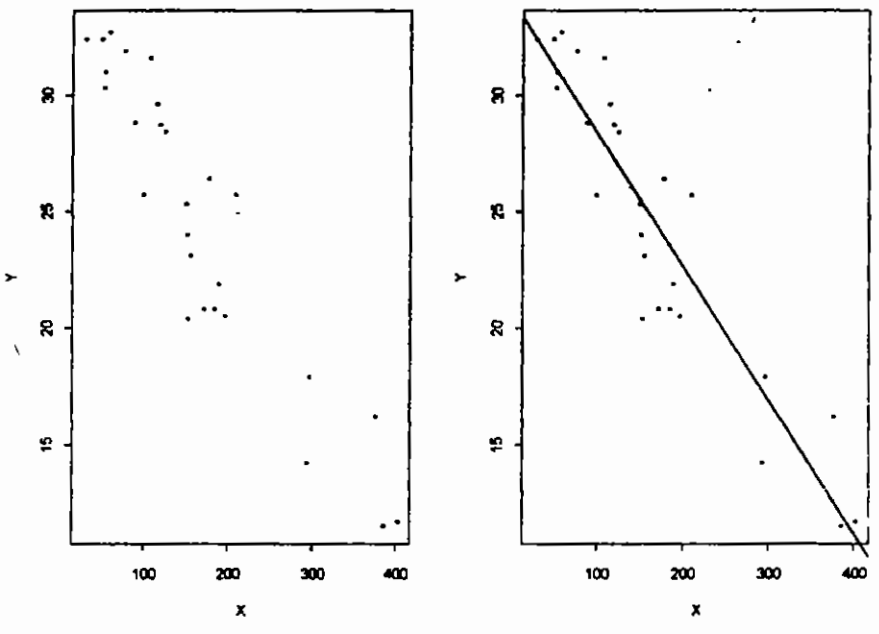
Dengan cara yang sama, Z*(m) merupakan suatu jumlah vektor dalam R^p, dan $d_2^{\,p} \{Z^*(m), Z(m)\}^2 \le d_2^{\,p} \{X_i^{*T} \varepsilon_i^*, X_i^T \varepsilon_i\}^2$, sehingga distribusi Z*(m) merupakan distribusi normal multivariate dengan mean 0 dan matrik varian-kovarian M, sehingga distribusi $\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ konvergen secara lemah ke distribusi normal dengan mean 0 dan matrik varian-kovarian $\Sigma^{-1} M \Sigma^{-1}$. □

Contoh Penggunaan Bootstrap

Untuk memberikan penjelasan pada pembahasan di atas, maka pada bagian ini akan diberikan suatu contoh analisis data yang menggunakan metode Bootstrap. Dalam bagian ini akan diperbandingkan antara pendekatan normal dan pendekatan bootstrap. Tabel 1 menyajikan data hormon, yang diambil dari Tabel 9.1 (Efron and Tibshirani, 1993) dan diolah dengan menggunakan paket program S-PLUS for Windows.

Tabel 1. Data Hormon

No.	jam	jumlah	No.	jam	jumlah	No.	jam	jumlah
1.	99	25,8	10.	376	16,3	19.	119	28,8
2.	152	20,5	11.	385	11,6	20.	188	22,0
3.	293	14,3	12.	402	11,8	21.	115	29,7
4.	155	23,2	13.	29	32,5	22.	88	28,9
5.	196	20,6	14.	76	32,0	23.	58	32,8
6.	53	31,1	15.	296	18,0	24.	49	32,5
7.	184	20,9	16.	151	24,1	25.	150	25,4
8.	171	20,9	17.	177	26,5	26.	107	31,7
9.	52	30,4	18.	209	25,8	27.	125	28,5



(a) (b)

Gambar 1. Fitting data hormon (a). Garis regresi data hormon (b)

Gambar 1(a) scatter plot 27 titik data, yang diambil dari Tabel 1, yaitu $(x_i, y_i) = (jam_i, jumlah_i)$. Plot pada Gambar 1(b) menunjukkan bahwa suatu himpunan data kecil, yang sangat baik untuk analisis regresi.

Model regresi dengan $p = 2$, dan parameter β diberikan: $Y = X\beta + \varepsilon$, dengan $X_i = (1, x_i)$ dan $\beta^T = (\beta_0, \beta_1)$. Persamaan normal memberikan estimasi kuadrat terkecil $\hat{\beta}^T = (34{,}17, -0{,}0574)$. Ringkasan statistik estimasi kuadrat terkecil selengkapnya diberikan dalam tabel di bawah ini.

Tabel 2. Ringkasan statistik estimasi kuadrat terkecil

parameter	estimasi	sesatan standar
β_0	34,1675282	0,834461
β_1	-0,0574463	0,004296

Tabel 3 dan Tabel 4 berikut ini menyajikan ringkasan statistik estimasi kuadrat terkecil dengan menggunakan pendekatan bootstrap, dengan jumlah replika bootstrap B = (100, 200, 300, 400, 500, 600).

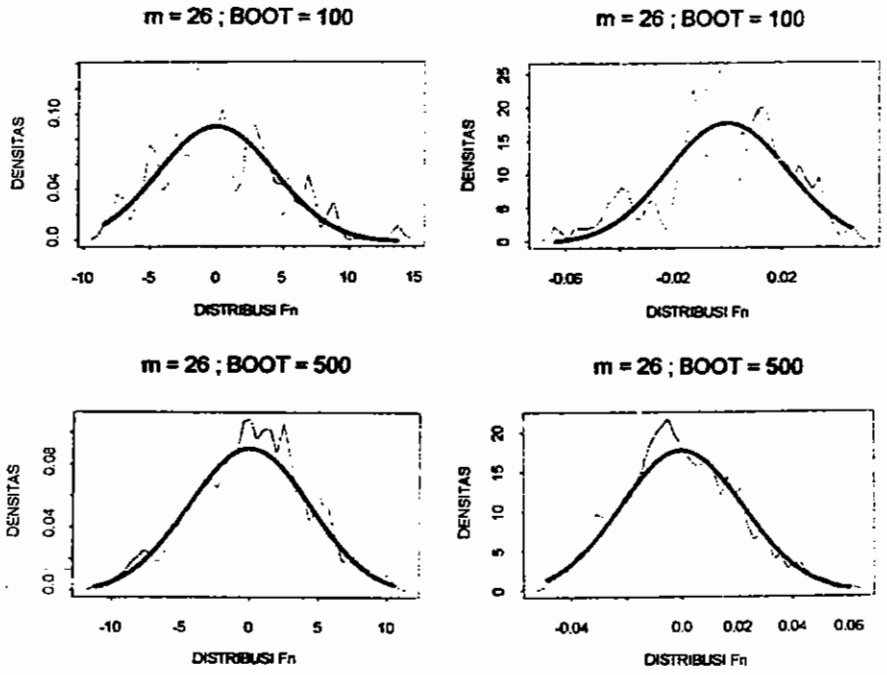
Tabel 3. Ringkasan statistik estimasi kuadrat terkecil berdasarkan bootstrapping residu

jumlah replika bootstrap	sesatan standar	
	β_0	β_1
100	0,8428236	0,004201863
200	0,8395171	0,004227824
300	0,8302135	0,004290505
400	0,8332706	0,004323663
500	0,8253492	0,004290892
600	0,8233578	0,004168954

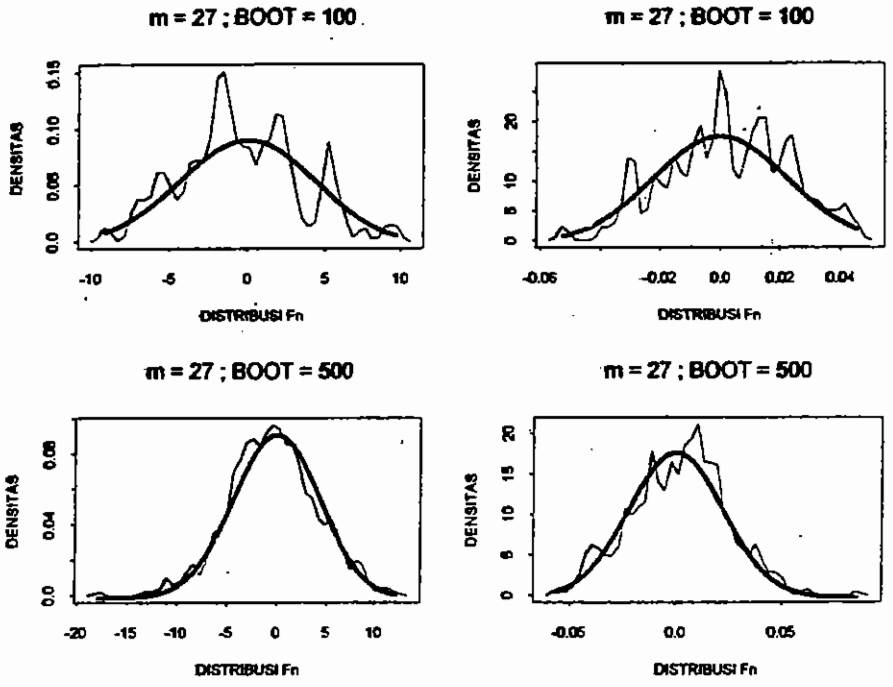
Tabel 4. Ringkasan statistik estimasi kuadrat terkecil berdasarkan bootstrapping pasangan

jumlah replika bootstrap	sesatan standar	
	β_0	β_1
100	0,7839182	0,004666107
200	0,7734073	0,004570576
300	0,7861598	0,004666561
400	0,7818962	0,004606659
500	0,7669171	0,004624214
600	0,7729119	0,004471365

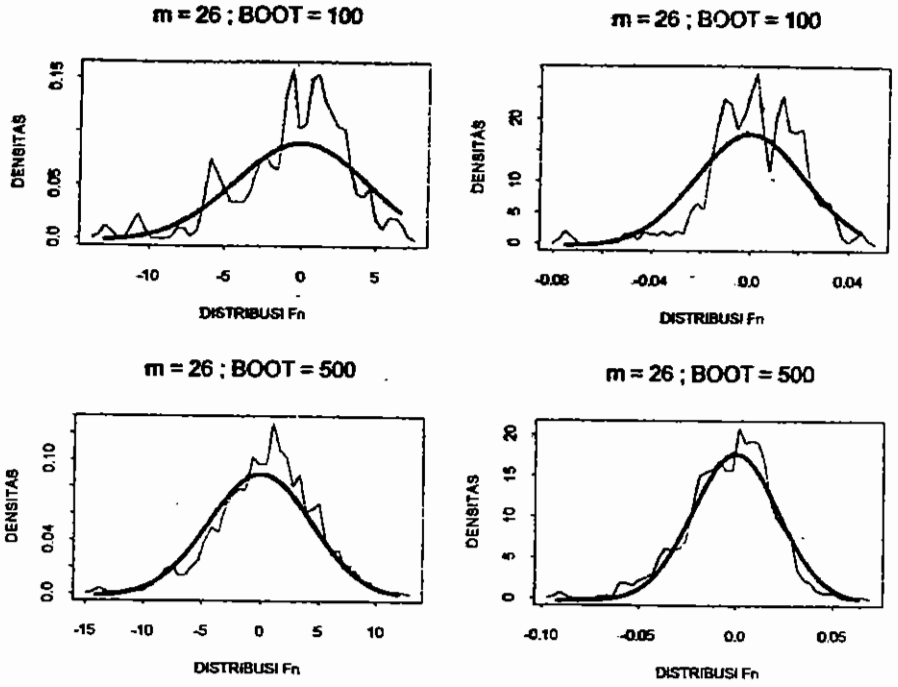
Dalam Tabel 3 dan Tabel 4, terlihat bahwa sesatan standar $\hat{\beta}^*$ tidak jauh berbeda dengan sesatan standar $\hat{\beta}$, yang tercantum dalam Tabel 2. Berarti pendekatan bootstrap dalam hal ini cukup baik. Selanjutnya pada Gambar 2 sampai dengan Gambar 5 disajikan fungsi densitas pendekatan distribusi F_n, yaitu fungsi densitas pendekatan bootstrap dan pendekatan normal, berdasarkan bootstrapping residu dan pasangan.



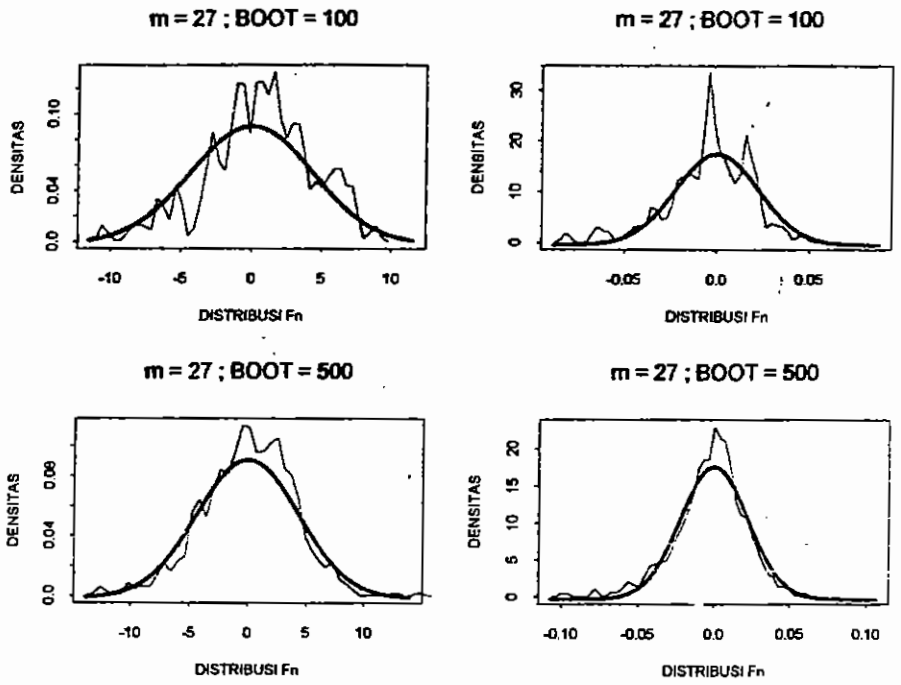
Gambar 2. Plot Pendekatan Normal dan Pendekatan Bootstrap distribusi $\sqrt{n}(\hat{\beta}(n)-\beta)$. Kolom pertama dan kedua, masing-masing bersesuaian dengan β_0 dan β_1. Berdasarkan bootstrapping residu dengan m = 26.
——— Bootstrap; ——— Normal.



Gambar 3. Plot Pendekatan Normal dan Pendekatan Bootstrap distribusi $\sqrt{n}\,(\hat{\beta}(n)-\beta)$. Kolom pertama dan kedua, masing-masing bersesuaian dengan β_0 dan β_1. Berdasarkan bootstrapping residu dengan m = 27.
——— Bootstrap; ——— Normal.



Gambar 4. Plot Pendekatan Normal dan Pendekatan Bootstrap distribusi $\sqrt{n}\,(\hat{\beta}(n)-\beta)$. Kolom pertama dan kedua, masing-masing bersesuaian dengan β_0 dan β_1. Berdasarkan bootstrapping pasangan dengan $m = 26$.
——— Bootstrap; ——— Normal.



Gambar 5. Plot Pendekatan Normal dan Pendekatan Bootstrap distribusi $\sqrt{n}\,(\hat{\beta}(n)-\beta)$. Kolom pertama dan kedua, masing-masing bersesuaian dengan β_0 dan β_1. Berdasarkan bootstrapping pasangan dengan m = 27.
——— Bootstrap; ——— Normal.

KESIMPULAN

Diberikan model regresi $Y(n) = X(n)\beta + \varepsilon(n)$

1. Diketahui $\hat{\beta}^*(m)$ estimasi kuadrat terkecil bootstrap untuk $\hat{\beta}(n)$: $\hat{\beta}^*(m) = \{X(m)^T X(m)\}^{-1} X(m)^T Y^*(m)$. $\sqrt{m}\,(\hat{\beta}^*(m)-\hat{\beta}(n))$ merupakan estimator konsisten untuk $\sqrt{n}\,(\hat{\beta}(n) - \beta)$. Distribusi $\sqrt{m}\,(\hat{\beta}^*(m)-\hat{\beta}(n))$ konvergen secara lemah ke distribusi normal dengan mean 0 dan matrik varian-kovarian $\sigma^2 V^{-1}$.

2. Untuk model korelasi, Distribusi $\sqrt{m}\,(\hat{\beta}^*(m) - \hat{\beta}(n))$ konvergen secara lemah ke distribusi normal dengan mean 0 dan matrik varian-kovarian $\Sigma^{-1}M\Sigma^{-1}$.
3. Dengan kondisi yang lebih lemah, pendekatan bootstrap memberikan hasil asimtotis yang sama dengan pendekatan normal.

DAFTAR PUSTAKA

Ash, R.B., 1972, *Real Analysis and Probability*, Academic Press, New York.

Bickel, P.J. and Freedman, D.A., 1981, Some Asymtotic Theory for the Bootstrap, *The Annals of Statistics*, 9:1196-1217.

Bickel, P. J. and Freedman, D. A., 1983, Bootstrapping Regression Models with Many Parameters, P. J. Bickel, K. A. Doksum, and J. L. Holges Jr., eds. *A festschrift for E.L. Lehman*, Wadsworth, Belmont, CA. pp.28-48.

Billingsley, P., 1979, *Probability and Measure*, Wiley, New York.

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

Efron, B. and Tibshirani, R. J., 1993, *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Freedman, D. A., 1981, Bootstrapping Regression Models, *The Annals of Statistics*, 9:1218-1228.

Putter, H., 1994, *Consistency of Resampling Methods*, Ph.D. Dissertation, University of Leiden.

Seber, G.A.F., 1977, *Linear Regression Analisis*, John Wiley & Sons, New York.

Singh, K., 1981, On the Asymtotic Accuracy of Efron's Bootstrap, *The Annals of Statistics*, 9:1187-1195.

Statistical Sciences, Inc., 1993, *S-PLUS for Windows User's Manual, Version 3.1*, Seattle: Statistical Sciences, Inc.

Wu, C.F.J., 1986, Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis, *The Annals of Statistics*, 14:1261-1295.

Zulaela, at al., 1995, *Bootstrapping Linear Regression Models*, Research Workshop in Statistic. Unpublished manuscript.