

PENENTUAN UKURAN CONTOH DAN REPLIKASI *BOOTSTRAP* UNTUK MENDUGA MODEL REGRESI LINIER SEDERHANA

OLIVIA ATINRI, HAZMIRA YOZZA, YUDIANTRI ASDI

*Program Studi Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia,
oliviatinri@gmail.com*

Abstrak. Analisis regresi adalah suatu teknik statistika yang digunakan untuk menarik suatu kesimpulan mengenai hubungan antara satu variabel tak bebas (Y) dengan satu atau lebih variabel bebas (X) dalam suatu sistem. Dalam analisis regresi, diperlukan beberapa asumsi yang harus dipenuhi oleh sebaran komponen galat (ε). Metode *bootstrap* merupakan salah satu metode yang dapat digunakan untuk menduga nilai-nilai penduga parameter jika asumsi sebaran komponen galat tidak terpenuhi. Dalam metode *bootstrap* biasanya dilakukan *resampling* sebanyak B kali ulangan *bootstrap*. Penelitian ini menggunakan metode simulasi Monte Carlo. Penelitian ini dilakukan dengan tujuan menentukan banyaknya ulangan *bootstrap* untuk menduga β_0 dan β_1 pada analisis regresi linier sederhana. Hasil dari penelitian ini diperoleh, untuk ulangan *bootstrap* $B=5$ menghasilkan nilai ragam penduga *bootstrap* relatif lebih tinggi. Untuk ukuran contoh berukuran lebih atau sama dari 40, tidak perlu dilakukan banyak ulangan *bootstrap*, karena dengan melakukan ulangan yang sedikit penduga parameter regresi sederhana sudah memiliki ragam yang kecil. Untuk sampel berukuran yang lebih kecil, dapat dilakukan 25 kali ulangan *bootstrap*. Namun perlu dicatat untuk ukuran sampel sebesar 10 sebaiknya tidak digunakan, karena menghasilkan penduga parameter regresi dengan ragam yang masih cukup besar.

Kata Kunci: Analisis regresi, metode Bootstrap, metode simulasi Monte Carlo.

1. Pendahuluan

Analisis regresi adalah suatu teknik statistika yang digunakan untuk menarik suatu kesimpulan mengenai hubungan antara satu variabel tak bebas (Y) dengan satu atau lebih variabel bebas (X) dalam suatu sistem. Hubungan antara variabel-variabel tersebut biasanya dinyatakan dalam suatu model regresi yang secara umum dinyatakan sebagai $Y = f(X_1, X_2, \dots, X_k) + \varepsilon$, ε adalah komponen galat. Model tersebut menghubungkan variabel bebas dan tak bebas melalui suatu parameter yang dinamakan sebagai parameter regresi, dinotasikan dengan β .

Dalam pemodelan regresi, metode kuadrat terkecil sering digunakan untuk mendapatkan nilai-nilai penduga parameter. Pada penggunaan metode kuadrat terkecil diperlukan beberapa asumsi klasik yang harus dipenuhi oleh komponen galat (ε), diantaranya galat harus menyebar normal, ragam galat yang homogen dan tidak terjadi autokorelasi antar galat.

Jika asumsi tidak terpenuhi, diperlukan suatu metode pendugaan lain yang kekar

terhadap ketidakterpenuhiannya asumsi tersebut. Metode *bootstrap* merupakan salah satu pilihan yang dapat digunakan untuk menghasilkan model yang lebih baik. Metode *bootstrap* bekerja tanpa membutuhkan asumsi mengenai distribusi karena pada pendekatan ini, sampel asli yang diperlakukan sebagai populasi. Sebagai pengganti, *bootstrap* menggunakan distribusi empiris untuk menduga sebaran penarikan contoh tersebut. Dalam metode *bootstrap* biasanya dilakukan *resampling* sebanyak B kali ulangan *bootstrap*. Yang menarik untuk dilihat adalah berapa banyaknya ulangan *bootstrap* yang diperlukan dalam pendugaan parameter regresi linier sederhana. Penelitian ini dilakukan dengan tujuan menentukan banyaknya ulangan *bootstrap* untuk menduga β_0 dan β_1 pada analisis regresi linier sederhana dalam kondisi galat menyebar menurut Sebaran Khi-kuadrat, sebaran Eksponensial, Sebaran- t , Sebaran- F , Sebaran Gamma, dan Sebaran Normal namun dengan ragam yang tidak sama dan ukuran contoh (n) = 10, 15, 20, 25, 30, 35, 40, 45, 50.

2. Analisis Regresi

2.1. Analisis Regresi Linier Sederhana

Analisis regresi berguna untuk menelaah pola hubungan antara dua atau lebih variabel yang modelnya belum diketahui dengan sempurna. Model regresi linier yang menggambarkan hubungan antara variabel tak bebas atau variabel yang dijelaskan (*dependent variable*) dengan satu variabel bebas atau variabel penjelas (*independent variable*) disebut model regresi linier sederhana, secara umum dapat dinyatakan sebagai

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (2.1)$$

dengan y_i merupakan nilai variabel tak bebas untuk pengamatan ke- i , x_i merupakan nilai variabel bebas untuk pengamatan ke- i , β_0, β_1 merupakan parameter regresi, dan ε_i merupakan nilai galat untuk pengamatan ke- i .

Tujuan utama dari analisis regresi adalah untuk mendapatkan dugaan dari model regresi. Pendugaan model regresi pada persamaan (2.1) dilakukan dengan menggunakan model regresi sampel yang dinyatakan sebagai:

$$y_i = b_0 + b_1 x_i + e_i. \quad (2.2)$$

dengan y_i merupakan nilai variabel terikat untuk pengamatan ke- i , x_i merupakan nilai variabel bebas untuk pengamatan ke- i , b_0, b_1 merupakan koefisien regresi sampel, dan e_i merupakan nilai sisaan untuk pengamatan ke- i .

Metode kuadrat terkecil adalah salah satu metode yang sering digunakan untuk menduga parameter pada regresi linier sederhana dengan meminimumkan jumlah kuadrat sisaan sehingga diperoleh :

$$\begin{aligned} b_0 &= \bar{Y} - b_1 \bar{X} & \text{dengan} & \quad \text{Var}(b_0) = \frac{s^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\ b_1 &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} & \text{dengan} & \quad \widehat{\text{Var}}(b_1) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \quad (2.3)$$

Terdapat beberapa asumsi yang harus dipenuhi oleh model regresi linier sederhana. Asumsi tersebut meliputi asumsi-asumsi mengenai distribusi dari galat, ε , dan

asumsi mengenai hubungan antara X dan Y . Asumsi tersebut menyatakan bahwa ε merupakan variabel acak yang saling bebas dengan nilai harapan sama dengan nol dan ragam yang homogen, yakni sebesar σ^2 , untuk semua nilai X . Selain itu X dianggap konstan dari sampel ke sampel dan untuk model regresi linier sederhana, Y merupakan fungsi linier X_i .

2.2. Metode Bootstrap

Metode *bootstrap* merupakan metode yang digunakan untuk mengestimasi suatu distribusi populasi yang tidak diketahui dengan menggunakan distribusi empiris yang diperoleh dari proses pengambilan sampel ulang dari sampel asli dengan ukuran sama dengan ukuran sampel asli dan dilakukan dengan pengembalian [3]. Kedudukan sampel asli dalam metode *bootstrap* dipandang sebagai populasi. Metode peyampelan ini biasa disebut dengan *resampling bootstrap*.

Prinsip dasar pembentukan *bootstrap* sebagai berikut:

- Mengkontruksi distribusi empiris $\hat{F}_n$ dari suatu sampel dengan memberikan peluang $\frac{1}{n}$ masing-masing $x_1, x_2, \dots, x_n$.
- Dengan $\hat{F}_n$ tetap, ambil sampel acak berukuran n dari $\hat{F}_n$ sebut, $X_i^* = x_i^*$, $X_i^* \sim \hat{F}_n$, $i = 1, 2, \dots, n$.

Prinsip dasar *resampling bootstrap* sebagai berikut:

- (1) Mengambil sampel *bootstrap* berukuran n secara acak dengan pengembalian dari distribusi empiris $\hat{F}_n$ disebut sebagai sampel *bootstrap* pertama X_1^* .
- (2) Menghitung statistik $\hat{\theta}$ yang diinginkan dari sampel *bootstrap* X_1^* disebut $\hat{\theta}_1^*$.
- (3) Mengulangi langkah 2 dan 3 sebanyak B kali sehingga diperoleh $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$.
- (4) Mengkonstruksikan suatu distribusi peluang dari $\hat{\theta}_B^*$ dengan memberikan peluang $\frac{1}{B}$ pada setiap $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$. Distribusi tersebut merupakan penduga *bootstrap* untuk distribusi sampling $\hat{\theta}$ dan dinotasikan dengan $\hat{F}^*$.
- (5) Pendekatan estimasi bootstrap untuk mean dari distribusi $\hat{F}^*$ yaitu

$$\hat{\theta}^* = \sum_{i=1}^B \hat{\theta}_i^* \frac{1}{B}.$$

Algoritma *bootstrap* untuk mengestimasi *standar error* $\hat{\theta}$ sebagai berikut.

- Pilih B sampel variabel bebas *bootstrap* $x_1^*, x_2^*, \dots, x_B^*$ dengan tiap-tiap sampel beranggotakan n data yang ditarik dengan pengembalian dari X .
- Evaluasi replikasi bootstrap yang bersesuaian pada setiap sampel *bootstrap*

$$\hat{\theta}^*(b) = s(x_b^*); b = 1, 2, \dots, n.$$

- Mengestimasi *standar error* $se_F(\hat{\theta})$ dengan menggunakan *standar deviasi* sampel dari B replikasi

$$\hat{se}_B = \left(\sum_{b=1}^B \frac{[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right)^{1/2}; \hat{\theta}^*(\cdot) = \sum_{b=1}^B \frac{\hat{\theta}^*(b)}{B}.$$

Penduga *bootstrap* untuk bias didefinisikan sebagai $\widehat{bias}_B$, yaitu

$$\widehat{bias}_B = \widehat{\theta}^*(\cdot) - \widehat{\theta}.$$

2.3. Simulasi Monte Carlo

Simulasi Monte Carlo adalah suatu metode untuk mengevaluasi secara berulang suatu model matematika dengan membangkitkan data menggunakan himpunan bilangan acak sebagai masukan dengan tujuan untuk mencerminkan dengan baik keadaan sesungguhnya.

Tujuan menggunakan Simulasi Monte Carlo adalah untuk menentukan bagaimana variabel acak mempengaruhi sensitivitas, performa atau reliabilitas dari sistem yang sedang dimodelkan. Simulasi Monte Carlo digolongkan sebagai metode *sampling* karena input dibangkitkan secara acak dari suatu distribusi kepekatan peluang untuk proses *sampling* dari suatu populasi nyata.

3. Metode Penelitian

Data yang digunakan pada penelitian ini terdiri dari satu variabel bebas X dan satu variabel bebas Y . Data dibangkitkan dengan menggunakan paket program Minitab melalui Simulasi Monte Carlo. Tahap pertama adalah membangkitkan data populasi dengan prosedur sebagai berikut.

- (1) Tentukan nilai $\beta_0 = 3$ dan $\beta_1 = 2$.
- (2) Bangkitkan 500 nilai galat, ε , dari sebaran tertentu.
- (3) Tentukan 500 nilai X . Dalam penelitian ini, ditetapkan nilai variabel X dari 0 sampai 98 dengan kenaikan 2, masing-masing sebanyak 10 untuk setiap nilai.
- (4) Hitung nilai $Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i$.
- (5) Bilai $E(e_i) \neq 0$ maka $Y_i = Y_i^* - E(e_i)$.

Akan diperoleh data populasi yang terdiri dari 500 data berpasangan (X_i, Y_i) .

Pada tahap selanjutnya, dilakukan penarikan sampel berukuran n secara acak dari populasi ini. Data ini merupakan sampel asli dari mana pendugaan *bootstrap* akan dilakukan. Penelitian ini dilakukan dengan metode simulasi, dengan langkah-langkah sebagai berikut.

- (1) Lakukan langkah-langkah berikut sebanyak B kali.
 - Ambil sampel ulang berukuran n dari sampel asli. Pengambilan dilakukan secara acak dengan pengembalian.
 - Tentukan dugaan parameter regresi b_{0k}^* dan b_{1k}^* , $k = 1, 2, \dots, B$.
- (2) Hitung rata-rata b_0 dan b_1 yang diperoleh pada langkah 1.
- (3) Lakukan langkah 1 dan 2 sebanyak 500 kali, sehingga dapat diperoleh nilai harapan dan nilai ragam penduga *bootstrap* bagi parameter regresi.
- (4) Lakukan langkah 1-3 untuk berbagai kombinasi ukuran contoh, ulangan *bootstrap* dan sebaran galat.

Proses pada simulasi ini dilakukan untuk berbagai skenario yaitu :

- Bentuk sebaran ε , yaitu sebaran *Khi-kuadrat* dengan $db=1$, *eksponensial* (10), *gamma*(5,3), *t* dengan $db=10$, *F*(2,5), normal dengan ragam yang berbeda.
- Berbagai ukuran contoh (n), yaitu untuk $n_1 = 10, n_2 = 15, n_3 = 20, n_4 = 25, n_5 = 30, n_6 = 35, n_7 = 40, n_8 = 45, n_9 = 50$.
- Berbagai nilai B , yaitu untuk $B = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$.

4. Hasil dan Pembahasan

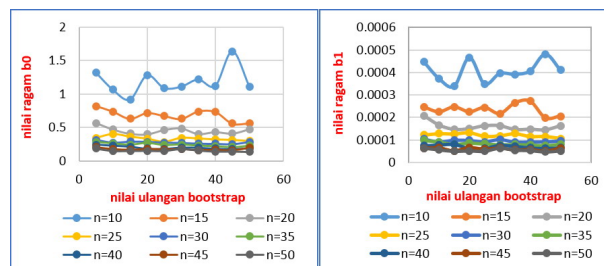
Penentuan banyaknya ulangan *bootstrap* (B) dilakukan dengan simulasi Monte Carlo dalam 500 kali ulangan. Pendugaan terbaik didasarkan pada kriteria ragam dari koefisien regresi yang akan diduga. Data populasi pada penelitian ini berukuran 500 data, berupa data berpasangan (X_i, Y_i) yang dibangkitkan dengan menggunakan paket program Minitab, dimana

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Untuk data populasi ini, ditetapkan nilai parameter $\beta_0 = 3$ dan $\beta_1 = 2$ sedangkan untuk nilai galat dibangkitkan dari berbagai sebaran, yaitu sebaran *Khi-kuadrat* dengan derajat bebas 10, sebaran *Eksponensial* (10), sebaran *Gamma* (5,3), sebaran-*t* dengan derajat bebas 10, sebaran-*F* (2,5), dan sebaran *Normal* dengan nilai ragam yang bervariasi. Dengan kata lain, simulasi ini dilakukan pada kondisi dimana asumsi mengenai sebaran galat terlanggar.

4.1. Simulasi dengan galat menyebar ($\varepsilon_i \sim \chi_{10}^2$)

Dalam kasus ini, terjadi pelanggaran asumsi, dimana galat yang seharusnya menyebar menurut sebaran *Normal* dengan nilai tengah nol dan ragam yang konstan, diatur sehingga menyebar menurut sebaran *Khi-kudrat* dengan derajat bebas 10 (sehingga memiliki nilai tengah 10).



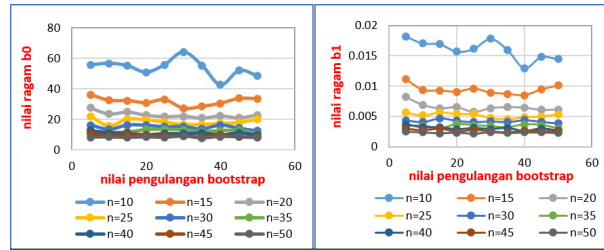
Gambar 1. Ragam Penduga Parameter Regresi ($\varepsilon_i \sim \chi_{10}^2$)

Dari Gambar 1 terlihat pada penduga parameter β_0 dan β_1 untuk $n \geq 40$, tidak perlu memperhatikan banyaknya ulangan *bootstrap* yang akan dilakukan, karena berapapun banyaknya ulangan *bootstrap* yang dilakukan sudah menghasilkan ragam yang kecil. Untuk sampel yang berukuran lebih kecil, dapat dilakukan 25

kali ulangan *bootstrap*. Namun untuk ukuran sampel sebesar 10 sebaiknya tidak dilakukan, karena masih menghasilkan ragam yang cukup besar.

4.2. Simulasi dengan galat menyebar ($\varepsilon_i \sim \exp_{(10)}$)

Sama halnya dengan kasus sebelumnya, terjadi pelanggaran asumsi dimana galat yang seharusnya menyebar menurut sebaran Normal dengan nilai tengah nol dan ragam yang konstan, ternyata menyebar menurut sebaran Eksponensial dengan $\lambda = 10$ sehingga memiliki nilai tengah 10.

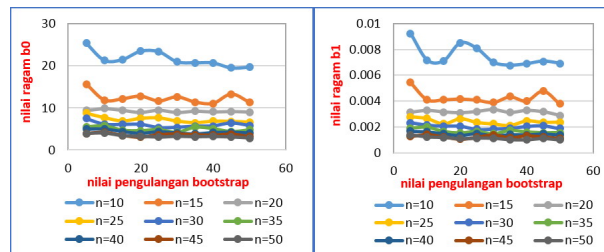


Gambar 2. Ragam Penduga Paramater Regresi ($\varepsilon_i \sim \exp_{(10)}$)

Dari Gambar 2 terlihat pada penduga parameter β_0 dan β_1 untuk $n > 40$, dapat dilakukan 25 kali ulangan *bootstrap*. Sedangkan untuk $n \geq 40$ cukup hanya melakukan 5x ulangan *bootstrap* saja karena sudah menghasilkan ragam yang kecil. Namun untuk ukuran sampel sebesar 10 sebaiknya tidak dilakukan, karena masih menghasilkan ragam yang cukup besar.

4.3. Simulasi dengan galat menyebar ($\varepsilon_i \sim GAM_{(5,3)}$)

Dalam kasus ini, juga terjadi pelanggaran asumsi yang sama dengan kedua kasus sebelumnya. Pada kasus ini, galat yang seharusnya menyebar menurut sebaran Normal dengan nilai tengah nol dan ragam yang konstan, diatur sehingga menyebar menurut sebaran Gamma (5,3) (nilai tengah=15 dan ragam=45).



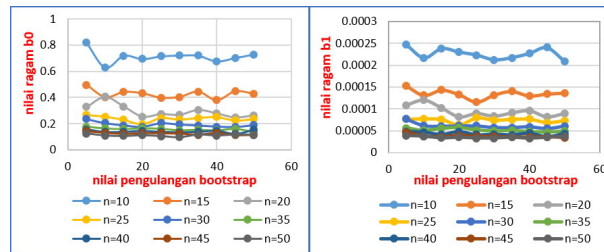
Gambar 3. Ragam Penduga Paramater Regresi ($\varepsilon_i \sim GAM_{(5,3)}$)

Dari Gambar 3 terlihat pada penduga parameter β_0 dan β_1 untuk $n > 30$, tidak perlu memperhatikan banyaknya ulangan *bootstrap* yang akan dilakukan,

karena berapapun banyaknya ulangan *bootstrap* yang dilakukan sudah menghasilkan ragam yang kecil. Untuk sampel yang berukuran lebih kecil, dapat dilakukan 20 kali ulangan *bootstrap*. Namun untuk $n \leq 10$ sebaiknya tidak dilakukan, karena masih menghasilkan ragam yang cukup besar.

4.4. Simulasi dengan galat menyebar ($\varepsilon_i \sim t_{(10)}$)

Dalam kasus ini, terjadi pelanggaran asumsi, dimana galat yang seharusnya menyebar menurut sebaran Normal dengan nilai tengah nol dan ragam yang konstan, diatur sehingga menyebar menurut sebaran-t dengan derajat bebas 10 (sehingga memiliki nilai tengah 10).

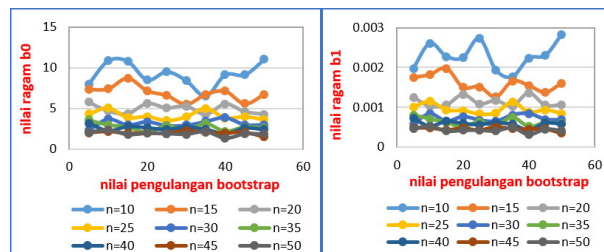


Gambar 4. Ragam Penduga Paramater Regresi ($\varepsilon_i \sim t_{(10)}$)

Dari Gambar 4 terlihat pada penduga parameter β_0 dan β_1 untuk $n \geq 40$, cukup dengan melakukan ulangan *bootstrap* sebanyak 5x sudah memperoleh ragam yang kecil. Untuk $n < 40$ dapat dilakukan 20 kali ulangan *bootstrap*. Namun untuk $n \geq 10$ sebaiknya tidak dilakukan, karena masih menghasilkan ragam yang cukup besar.

4.5. Simulasi dengan galat menyebar ($\varepsilon_i \sim F_{(2,5)}$)

Dalam kasus ini, terjadi pelanggaran asumsi, dimana galat yang seharusnya menyebar menurut sebaran Normal dengan nilai tengah nol dan ragam yang konstan, diatur sehingga menyebar menurut sebaran F (2,5).



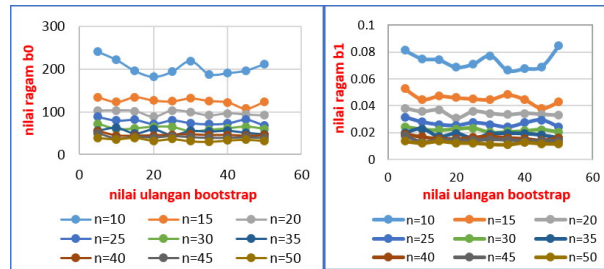
Gambar 5. Ragam Penduga Paramater Regresi ($\varepsilon_i \sim F_{(2,5)}$)

Dari Gambar 5 terlihat pada penduga parameter β_0 dan β_1 untuk $n \geq 40$, kita bisa saja melakukan ulangan *bootstrap* sebanyak 5x karena akan didapatkan nilai

penduga ragam yang sudah kecil. Untuk $n < 40$ dapat dilakukan 25 kali ulangan *bootstrap*. Namun untuk ukuran sampel sebesar 10 sebaiknya tidak dilakukan, karena masih menghasilkan ragam yang cukup besar.

4.6. Simulasi dengan galat menyebar ($\varepsilon_i \sim N$) dengan ragam yang berbeda

Berbeda dengan kasus-kasus sebelumnya, pada kasus ini asumsi yang terlanggar adalah asumsi kehomogenan ragam yang seharusnya konstan. Sehingga galat menyebar dengan sebaran Normal dengan nilai tengah 0 dan ragam 1 sebanyak 200 data, ragam 16 sebanyak 200 data dan ragam 25 sebanyak 100 data.



Gambar 6. Ragam Penduga Paramater Regresi ($\varepsilon_i \sim N$)

Dari Gambar 6 terlihat pada penduga parameter β_0 dan β_1 untuk $n \geq 40$, tidak perlu memperhatikan banyaknya ulangan *bootstrap* yang akan dilakukan, karena bagaimanapun banyaknya ulangan *bootstrap* yang dilakukan sudah menghasilkan ragam yang kecil. Untuk sampel yang berukuran lebih kecil, dapat dilakukan 25 kali ulangan *bootstrap*. Namun untuk $n \leq 10$ sebaiknya tidak dilakukan, karena masih menghasilkan ragam yang cukup besar.

5. Kesimpulan

Dari penelitian yang dilakukan diperoleh untuk semua ukuran contoh, pada ulangan *bootstrap* $B=5$ nilai ragam penduga *bootstrap* tersebut relatif lebih tinggi. Semakin banyak ulangan *bootstrap*, semakin kecil nilai ragam penduganya. Sampai pada suatu nilai B tertentu, ragam penduga tersebut kembali naik atau tetap turun namun dengan besarnya penurunan yang tidak terlalu berarti.

Dari keseluruhan hasil yang diperoleh dapat ditarik kesimpulan mengenai banyaknya ulangan *bootstrap* yang harus dilakukan untuk menduga parameter analisis regresi linier sederhana. Bila dimiliki sampel berukuran lebih atau sama dari 40, tidak perlu dilakukan banyak ulangan *bootstrap*, karena dengan melakukan ulangan yang sedikit penduga parameter regresi sederhana sudah memiliki ragam yang kecil dan melakukan ulangan yang lebih banyak tidak memperbaiki ragam penduga dengan signifikan. Untuk sampel yang berukuran lebih kecil, dapat dilakukan 25 kali ulangan *bootstrap*. Namun perlu juga dicatat bahwa ukuran sampel sebesar 10

sebaiknya tidak digunakan, karena menghasilkan penduga parameter regresi dengan ragam yang masih cukup besar.

6. Ucapan Terima kasih

Penulis mengucapkan terima kasih kepada Bapak Dr. Dodi Devianto, Bapak Dr. Admi Nazra, dan Ibu Maiyastri, M.Si yang telah memberikan masukan dan saran sehingga paper ini dapat diselesaikan dengan baik.

Daftar Pustaka

- [1] Bain, L.J and Max Engelhardt. 1992. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, California
- [2] Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2000. *Discrete-Event System Simulation*. Prentice-Hall
- [3] Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- [4] Kakiay, Thomas J. 2004. *Dasar Teori Antrian untuk Kehidupan Nyata*. Yogyakarta.
- [5] Rubinstein, R. Y. 1981. *Simulation and the Monte Carlo Method*. John Wiley and Sons, Inc., New York.
- [6] Supranto, J. 1994. *Statistik Teori dan Aplikasi Edisi Kelima Jilid 2*. Penerbit Erlangga, Jakarta.
- [7] Walpole, R.E dan Myers, R.H. 1995. *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan*. Penerbit ITB, Bandung.