Aplikasi Bootstrap Pada Analisis Regresi untuk Data Kecelakaan Kerja

Toto Hermawan
Pendidikan Matematika, Universitas Cokroaminoto Yogyakarta
Jl. Perintis Kemerdekaan, Gambiran, Umbulharjo, Kota Yogyakarta 55161
Email: toto.hermawan@mail.ugm.ac.id

ABSTRAK

Untuk mengetahui hubungan antara dua variable atau lebih dapat digunakan analisis regresi. Pengertian analisis regresi sendiri adalah metode analisis data yang memanfaatkan hubungan antara dua variable atau lebih. Hal yang menjadi perhatian dalam analisis regresi salah satunya adalah standar error dari estimasi koefisien regresi. Dalam regresi sudah terdapat formula untuk mengestimasi standar error. Selain itu, standar error juga dapat diestimasi dengan metode resampling, yaitu bootstrap. Bootstrap sangat berguna sebagai alternatif untuk estimasi parameter atau standar errornya ketika peneliti merasa ragu dapat memenuhi asumsi pada data mereka, misal data tidak berdistribusi normal. Selain itu bootstrap juga berguna ketika inferensi parametric memerlukan rumus yang sangat rumit untuk menghitung standar error (Widhiarso, 2012). Dalam tulisan ini akan dibandingkan estimasi standar error yang diperoleh melalui formula yang sudah ada dengan estimasi standar error yang diperoleh melalui resampling bootstrap.

Kata kunci: Analisis Regresi, Metode resampling, Standar Error , Estimasi Koefisien Regresi, Bootstrap, program R

ABSTRACT

To find out the relationship between two or more variables, regression analysis can be used. The definition of regression analysis itself is a data analysis method that utilizes the relationship between two or more variables. One concern in regression analysis is one of them is the standard error of estimation of the regression coefficient. In a regression there is already a formula for estimating standard errors. In addition, the standard error can also be estimated by the resampling method, which is bootstrap. Bootstrapping is very useful as an alternative to estimating parameters or standard errors when researchers feel hesitant to meet the assumptions in their data, for example the data are not normally distributed. In addition, bootstrapping is also useful when parametric inference requires a very complicated formula for calculating standard errors (Widhiarso, 2012). In this paper we will compare the standard error estimates obtained through existing formulas with the standard error estimates obtained through bootstrap resampling.

Keywords: Regression Analysis, Resampling Method, Error Standards, Regression Coefficient Estimation, Bootstrap, R program

PENDAHULUAN

Analisis regresi adalah metode analisis data yang memanfaatkan hubungan antara dua variable atau lebih. Secara umum, tujuan dari analisis regresi adalah:

a. Menyelidiki pola hubungan antara variabel prediktor dan variabel respon. Untuk melakukannya dapat dilakukan dengan membuat diagram pencar.
b. Mengestimasi nilai pada variabel respon berdasarkan nilai variabel prediktor yang dimiliki.
c. Menyelidiki variabel prediktor yang mana saja yang berpengaruh secara signifikan terhadap variabel respon.

Pada tahap estimasi koefisien parameter regresi, perhatian tertuju pada standar error dari estimator tersebut. Untuk mengestimasi standar error dari estimator parameter, dalam analisis regresi terdapat formula yang closed-form untuk menghitungnya. Pada beberapa kasus, seringkali tidak terdapat formula tersebut sehingga digunakan metode resampling bootstrap. Dalam paper ini akan digunakan formula biasa dan metode resampling bootstrap untuk menghitung standar error. Selanjutnya akan dibandingkan untuk mengetahui apakah keduanya memberikan hasil yang berbeda atau tidak. Data yang digunakan dalam studi kasus yang dilakukan adalah data tentang kecelakaan kerja.

PEMBAHASAN

1. Analisis Regresi Linear

Pembahasan akan dimulai dari model klasik regersi linier yang dibahas Legendre dan Gauss early pada tahun 1900an menurut (Efron, 1993). Data set x untuk model regresi linear dimana terdapat n buah observasi didefinisikan sebagai berikut:

$$\mathbf{x_i} = (\mathbf{c_i}, \mathbf{y_i});\ [illegible] = 1, 2, \cdots, [illegible] \qquad \textbf{(1)}$$

$c_i = \left(1, c_{i1}, c_{i2}, \cdots, c_{ip}\right)$ adalah vektor kovariat atau prediktor, sedangkan y_i adalah bilangan real yang menyatakan variabel responnya. Banyaknya variabel prediktor dinyatakan dengan p.

Model regresi linear dinyatakan sebagai berikut

$$[illegible], \cdots, \mathbf{n} \qquad \textbf{(2)}$$

Vektor parameter regresi $\beta = \left(\beta_0, \beta_1, \cdots, \beta_p\right)$ tidak diketahui nilainya dan akan diestimasi berdasarkan x. Error ε_i diasumsikan sebagai sampel random dari suatu distribusi, misal F, dengan $E(\varepsilon_i) = 0$.

$$[illegible] ([illegible] \cdots, [illegible] = [illegible] \qquad \textbf{(3)}$$

Berdasarkan persamaan (2) diperoleh harga harapan untuk y_i jika diketahui c_i adalah

$$\begin{aligned} \mu_i = E(y_i|c_i) &= E(c_i\beta + \varepsilon_i|c_i) \qquad (4) \\ &= E(c_i\beta|c_i) + E(\varepsilon_i|c_i) \\ &= c_i\beta \end{aligned}$$

Untuk n buah sampel random, model regresi (2) dapat ditulis sebagai berikut

$$\mathbf{y} [illegible] \qquad (5)$$

Dengan

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$c = \begin{bmatrix} 1 & c_{11} & \cdots & c_{p1} \\ 1 & c_{12} & \cdots & c_{p2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & c_{1n} & \cdots & c_{pn} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \text{ dan } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimator untuk $\underline{\beta}$ dapat diperoleh melalui metode kuadrat terkecil. Jika S menyatakan jumlah kuadrat dari error, maka

$$\begin{aligned} S &= \sum_{i=1}^{n} \varepsilon_i^2 = \underline{\varepsilon}'\underline{\varepsilon} = (y - c\beta)'(y - c\beta) \\ &= y'y - y'c\beta - (c\beta)'y + (c\beta)'(c\beta) \\ &= y'y - (c\beta)'y - (c\beta)'y + (c\beta)'(c\beta) \\ &= y'y - 2(c\beta)'y + (c\beta)'(c\beta) \end{aligned}$$

Jika S diturunkan terhadap $\underline{\beta}$ dan disamadengankan nol diperoleh

$$\frac{\partial S}{\partial \beta} = 0 \leftrightarrow -2c'y + 2c'c\beta = 0$$

$$2c'c\beta = 2c'y$$

$$\underline{\hat{\beta}} = (\quad) \qquad (6)$$

Untuk mengetahui keakuratan estimator $\underline{\hat{\beta}}$ di atas dapat dilakukan dengan menghitung standar errornya. Didefinisikan matriks G sebagai berikut

$$\mathbf{G} = \quad \mathbf{C} \qquad (7)$$

Variansi dari estimator $\underline{\hat{\beta}}$ adalah

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}((c'c)^{-1}c'y) \\ &= ((c'c)^{-1}c')\text{var}(y)((c'c)^{-1}c')' \\ &= ((c'c)^{-1}c')\text{var}(y)(c(c'c)^{-1}) \\ &= \sigma_F^2(c'c)^{-1} \end{aligned}$$

Karena $\text{var}(y) = \sigma_F^2 I$, dimana I adalah matriks identitas, maka

$$\quad = \sigma \qquad \textbf{(8)}$$

Sehingga standar error elemen ke-j dari vektor $\underline{\hat{\beta}}$ adalah

$$\quad \qquad \textbf{(9)}$$

dimana G^{jj} adalah elemen diagonal ke-j dari G^{-1} (G invers).

Dalam prakteknya, σ_F dapat diestimasi dengan

$$\hat{\sigma}_F = \left\{ \frac{\sum_{i=1}^{n} (y_i - \quad)^2}{n} \right\}^{1/2} \qquad \textbf{(10)}$$

(Efron, 1993)

Estimator di atas termasuk estimator yang bias. Oleh karena itu seringkali digunakan estimator berikut

$$\bar{\sigma}_F = \left\{ \frac{\sum_{i=1}^{n} (y_i - c_i\hat{\beta})^2}{n - p} \right\}^{1/2}$$

Jadi estimasi standar error dari $\hat{\beta}_j$ adalah

(11)

2. Bootstrap pada Analisis Regresi Linear

Model probabilitas $P \rightarrow x$ untuk regresi linear, sebagaimana pada persamaan (2) dan (3), mempunyai dua komponen, yaitu:

$$P = (\beta, F) \tag{12}$$

Keduanya adalah parameter yang perlu diestimasi. Estimasi untuk β telah diperoleh melalui metode kuadrat terkecil, yaitu $\hat{\beta}$. Jika $\hat{\beta}$ telah diketahui, maka bisa dihitung estimasi untuk error yaitu

(13)

Karena $\hat{\varepsilon}$ yang diestimasi adalah sejumlah n, maka distribusi empiris dari $\hat{\varepsilon}$ adalah

(14)

Untuk melakukan bootstrap pada analisis regresi, diambil sampel berjumlah n secara random dengan pengembalian dari error estimasi $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*)$. Dari ε^* yang diperoleh dihitung variabel respon bootstrap sebagai berikut

(15)

Jika data asli adalah $x_i = (c_i, y_i)$ (lihat persamaan (1)), maka data hasil bootstrap adalah $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$ dimana $x_i^* = (c_i, y_i^*)$.

Jadi model regresi bootstrap-nya adalah

(16)

dimana $E(\varepsilon_i^{**}) = 0$ dan $var(\varepsilon_i^{**}) = \sigma_F^2$. Jadi $E(y_i^*) = c_i\hat{\beta}^*$ dan $var(y_i^*) = \sigma_F^2$. Dari data hasil bootstrap tersebut dapat dihitung estimasi parameter regresi bootstrap, yaitu $\hat{\beta}^*$ sebagai berikut

(17)

Cara untuk memperoleh estimasi di atas analog dengan cara pada bagian 2. Variansi dari $\hat{\beta}^*$ adalah

$$\begin{aligned} var(\hat{\beta}^*) &= var((c'c)^{-1}c'y^*) \\ &= ((c'c)^{-1}c')var(y^*)((c'c)^{-1}c')' \\ &= ((c'c)^{-1}c')var(y^*)(c(c'c)^{-1}) \\ &= \sigma_F^2(c'c)^{-1} \end{aligned}$$

Karena $var(y^*) = \sigma_F^2 I$, dimana I adalah matriks identitas, maka

(18)

Sehingga standar error elemen ke-j dari vektor $\hat{\beta}^*$ adalah

(19)

Estimator untuk σ_F adalah

(20)

$$\hat{\sigma}_F = \left\{\frac{\sum_{i=1}^{n}(y_i^* - c_i\hat{\beta}^*)^2}{n}\right\} \text{ atau}$$

$$\bar{\sigma}_F = \left\{\frac{\sum_{i=1}^{n}(y_i^* - c_i\hat{\beta}^*)^2}{n-p}\right\}^{1/2}$$

Jika diperhatikan ternyata persamaan (19) sama dengan persamaan (9). Jadi dapat dikatakan bahwa untuk mengestimasi standar error dari bootstrap dapat dilakukan dengan formula untuk mencari standar error yang biasa. Yang membedakannya hanyalah nilai σ_F-nya saja.
Selain itu, standar error dari koefisien parameter regresi dapat pula diestimasi dengan langkah-langkah sebagai berikut:

a. Diambil sampel berjumlah n secara random dengan pengembalian dari error estimasi $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*)$.
b. Dari ε^* yang diperoleh dihitung variabel respon bootstrap, yaitu $y_i^* = c_i\hat{\beta} + \varepsilon_i^*$; $i = 1,2,\cdots,n$, sehingga dimiliki set data $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$ dimana $x_i^* = (c_i, y_i^*)$.
c. Berdasarkan set data bootstrap yang diperoleh, yaitu x^*, dihitung koefisien parameter regresi dengan rumus $\hat{\beta}^* = (c'c)^{-1}c'y^*$.
d. Langkah nomor a sampai c diulang sebanyak B kali sehingga dimiliki B buah nilai $\hat{\beta}^*$.
e. Dihitung standar error dari koefisien parameter regresi melalui hasil bootstrap pada nomor d, yaitu:

$$\hat{se}_B\left(\hat{\beta}_j^*\right) = \left\{\sum_{b=1}^{B}\left[\hat{\beta}_{jb}^* - \hat{\beta}_j^*(.)\right]^2 / (1400 - 1)\right\}^{1/2}$$

dengan $\hat{\beta}_j^*(.) = \frac{\sum_{b=1}^{B}\hat{\beta}_{jb}^*}{B}$; $j = 0,1,2,\cdots,p$

3. Studi Kasus

Dalam studi kasus ini digunakan data tentang jumlah kecelakaan kerja pada suatu perseroan (PT). Pengamatan dilakukan terhadap 43 orang karyawan PT tersebut. Untuk setiap karyawan dilakukan pencatatan jumlah jam kerja dalam satu tahun, divisi dimana karyawan tersebut ditempatkan, dan jumlah kecelakaan kerja yang dialami dalam satu tahun. Pada PT tersebut, terdapat empat buah divisi yaitu produksi (frame), weaving (penenunan), quality control, dan gudang/logistik. Berdasarkan data yang terkumpul, dapat dilakukan analisis regresi linear untuk menyelidiki pengaruh jam kerja dan divisi terhadap jumlah kecelakaan kerja.

Diperoleh model regresi estimasi **sebagai berikut:**

$$\text{kec. kerja} = 2{,}1386 + 0{,}01019 - 0{,}4953(d_{weaving}) - 0{,}1998(d_{q.control}) - 1{,}5804(d_{gudang.log}) \quad \textbf{(21)}$$

dimana

$$d_{weaving} = \begin{cases} 1 ; & \text{divisi weaving} \\ 0 ; & \text{divisi yang lain} \end{cases}$$

$$d_{q.control} = \begin{cases} 1 ; & \text{divisi q. control} \\ 0 ; & \text{divisi yang lain} \end{cases}$$

$$d_{gudang.log} = \begin{cases} 1 ; & \text{gudang } logistik \\ 0 ; & \text{divisi yang lain} \end{cases}$$

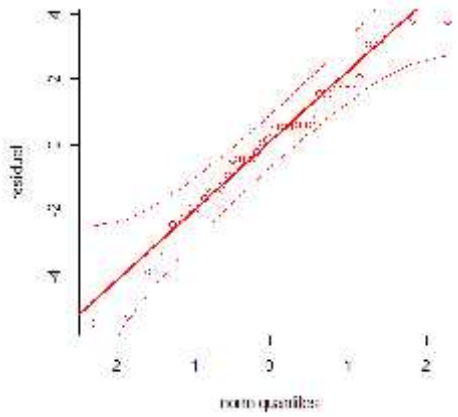
Berdasarkan persamaan di atas diketahui bahwa koefisien parameter regresi variabel jam bernilai positif walaupun nilainya cukup kecil. Artinya penambahan jam kerja yang banyak akan menambah jumlah kecelakaan kerja. Selain itu diperoleh koefisien regresi untuk ketiga variabel dummy bernilai negatif. Artinya divisi yang menjadi reference category, yaitu produksi (frame), memiliki angka kecelakaan kerja yang paling tinggi dibandingkan jumlah kecelakaan kerja pada divisi lain. Pertanyaan selanjutnya adalah seberapa akurat estimator koefisien parameter regresi di atas? Untuk menjawabnya maka dihitung standar error untuk setiap estimator.

Tabel 1. Estimasi Parameter Regresi dan Standar Error

Parameter	**Estimasi**		
	2,886	1,0068	1,0572
	0,001019	0,0004041	0,0004243
	-0,4953	0,8676	0,911
	-0,1998	0,8341	0,8758
	-1,5804	1,1784	1,2373

Pada kebanyakan paket program, estimasi untuk standar error yang biasa dipakai adalah $\overline{se}$, yaitu estimator yang tak bias. Berdasarkan estimasi parameter di atas dapat diperoleh qq plot residual sebagai berikut:

QQ Plot Residualnya



Dapat diketahui bahwa residual mendekati distribusi normal sehingga asumsi normalitas dalam analisis regresi terpenuhi.

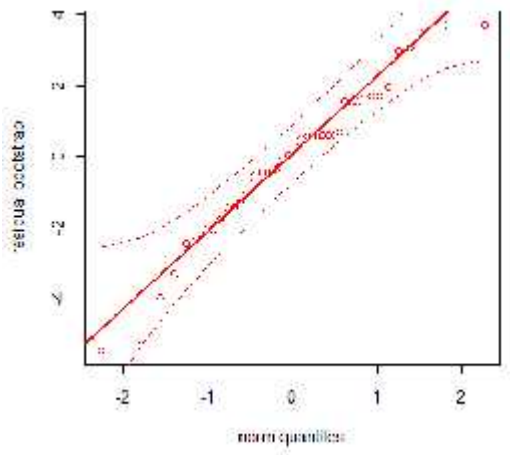
Untuk kasus regresi linear, terdapat formula yang closed-form untuk mengestimasi standar error. Berdasarkan tabel di atas, diperoleh nilai estimasi standar error yang cukup kecil. Jadi dapat dikatakan bahwa estimator koefisien regresi cukup akurat. Seandainya tidak terdapat formula yang closed-form dapat dilakukan bootstrap untuk mengestimasi standar errornya. Yang pertama dilakukan adalah menghitung residual berdasarkan nilai koefisien regresi yang diperoleh. Berdasarkan residual tersebut dilakukan pengambilan sampel secara random dengan pengembalian, namakan ε^{*}. Sampel yang terambil digunakan untuk menghitung $y_i^* = c_i\hat{\beta} + \varepsilon_i^*$. Berdasarkan nilai y_i^* yang diperoleh dihitung nilai estimasi parameter regresi yang baru. Jika dilakukan iterasi sebanyak 100 kali terhadap proses tersebut diperoleh estimasi untuk standar error adalah:

Tabel 2. Estimasi Parameter Regresi dan Standar Error Menggunakan Bootstrap

Parameter	Estimasi	[illegible]
[illegible]	2,8956	1,0133
	0,00102	0,0004225
	-0,4053	0,8124
	-0,1472	0,8461
	-1,48	1,1483

Diperoleh pula qq plot untuk residual bootstrap sebagai berikut:

QQ Plot Residualnya



Ternyata diperoleh bahwa asumsi normalitas residual tetap terpenuhi. Jika antara tabel 1 dan tabel 2 dibandingkan ternyata diperoleh hasil yang tidak jauh berbeda. Semakin banyak iterasi yang dilakukan maka hasil yang diperoleh akan mendekati hasil estimasi standar error yang diperoleh dengan cara biasa. Secara matematis dapat ditulis $\widehat{se}_{\infty}\left(\hat{\beta}_j\right) \approx \overline{se}\left(\hat{\beta}_j\right)$.

KESIMPULAN

Berdasarkan hasil pada bagian 4 diperoleh kesimpulan bahwa jumlah jam kerja dan jumlah kecelakaan kerja yang terjadi memiliki korelasi yang positif, artinya penambahan jam kerja diperkirakan akan menambah jumlah kecelakaan yang terjadi. Oleh karena itu, jika jumlah kecelakaan kerja yang terjadi berada pada kisaran yang mengkhawatirkan (cukup tinggi), maka salah satu solusi yang dapat dilakukan adalah dengan mengurangi jam kerja karyawan. Selain itu diperoleh pula kesimpulan bahwa angka kecelakaan kerja paling tinggi terjadi pada divisi produksi (frame). Oleh karena itu, divisi tersebut perlu mendapatkan perhatian khusus dalam hal penanggulangan kecelakaan kerja.

Berkaitan dengan analisis data, estimasi standar error menggunakan bootstrap mempunyai hasil yang tidak jauh berbeda dari estimasi menggunakan formula yang sudah ada. Seandainya tidak terdapat formula yang closed-form untuk mengestimasi standar error, maka bootstrap merupakan salah satu alternatif yang dapat dipilih. Penerapan bootstrap pada bagian 4 di atas merupakan salah satu dari sekian cara penerapan bootstrap untuk analisis regresi.

DAFTAR PUSTAKA

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

Krisnawardhani, Tanti. Salam, Nur. Anggraini, Dewi. 2010. Analisis Regresi Linear Berganda dengan Satu Variabel Boneka (Dummy Variable). Program Studi Matematika Universitas Lambung Mangkurat. Banjarbaru.

Seber. George A.F & Alanj.LEE. 2003. Linear Regression Analysis. Canada.

Widhiarso, Wahyu. 2012. Berkenalan dengan Bootstrap. Fakultas Psikologi UGM. Yogyakarta.