MODEL REGRESI LINIER BERGANDA MENGGUNAKAN PENAKSIR PARAMETER REGRESI *ROBUST* M-*ESTIMATOR* (Studi Kasus: Produksi Padi di Provinsi Jawa Barat Tahun 2009)

Rini Cahyandari, Nurul Hisani
Jurusan Matematika Fakultas Sains dan Teknologi
Universitas Islam Negeri Sunan Gunung Djati Bandung
Jl. AH. Nasution 105 Cibiru Bandung – 40614
e-mail: rcahyandari@yahoo.com

Abstrak

The least squares method is one method of parameter estimation in regression models. This method produces an unbiased estimator with consideration the assumptions are linearity, non-multicollinearity, non-autocorrelation, homoscedastic, and normally distributed error fulfilled. When the assumptions are not fulfilled, such as error distribution is not normal due to the existence of outliers, an estimation obtained are not exact.An alternative method that can overcome the problem of outliers is robust regression method using M-estimator (Iteratively Reweighted Least Squares). M-estimator is an iterative method using weighting function Huber and Tukey bisquare to estimate the parameters or coefficients in the regression model. The best model obtained by M-estimator robust method using Huber and Tukey bisquare is determined by the value $R^2_{adjusted}$ and standard error values.

Keywords : *Outliers, M-estimator robust regression method*

1. Pendahuluan

Analisis regresi merupakan analisis yang mempelajari bagaimana membentuk sebuah hubungan fungsional dari data untuk dapat menjelaskan atau meramalkan suatu fenomena alami atas dasar fenomena yang lain. Analisis regresi memiliki peranan yang penting dalam berbagai bidang ilmu pengetahuan. Kebanyakan analisis regresi bergantung pada metode kuadrat terkecil untuk mengestimasi parameter-parameternya dalam model regresi. Tetapi metode ini biasanya dibentuk dengan beberapa asumsi, seperti linearitas, tidak ada autokorelasi, tidak terjadi multikolinearitas, homoskedastisitas, dan *error* berdistribusi normal.

Pencilan (*outlier*) adalah pengamatan yang jauh dari pusat data yang mungkin berpengaruh besar terhadap koefisien regresi. Ketika terdapat pencilan (*outlier*) dalam data, maka dengan menggunakan estimasi metode kuadrat terkecil, hasil estimasi parameternya tidak akan memberikan informasi yang tepat bagi data yang ada, karena akan mengakibatkan nilai *error* menjadi besar (tidak normal). Sehingga telah dikembangkan analisis regresi *robust* sebagai perbaikan untuk estimasi kuadrat terkecil dalam model regresi linier berganda karena adanya suatu pencilan (*outlier*). Pada makalah ini, akan ditentukan suatu model regresi linier berganda dengan nilai estimasi parameter yang diperoleh dengan metode regresi *robust* M-*estimator* (*Iteratively Reweighted Least Squares*) menggunakan fungsi pembobot Huber dan Tukey *bisquare*. Sebagai studi kasus dipilih data produksi padi di Provinsi Jawa Barat

tahun 2009 yang dipengaruhi oleh luas panen (X_1) dan luas irigasi teknis (X_2).

Permasalahan dalam makalah ini dibatasi, diantaranya:

1. Metode regresi *robust* yang digunakan adalah M-*estimator*
2. Model yang digunakan adalah model regresi linier berganda
3. Beberapa asumsi seperti linearitas, non multikolinearitas, non autokorelasi dan homoskedastisitas, tetap terpenuhi

2. Model Regresi Linier Berganda

Model regresi yang menghubungkan dua atau lebih variabel bebas (X) dengan satu variabel terikat (Y) disebut regresi linier berganda, dan dinyatakan dalam model sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ \dots \ + \beta_p X_{pi} + \varepsilon_i, \quad i = 1, 2, \dots, n \tag{1}$$

Karena populasi jarang diamati secara langsung, maka digunakan persamaan regresi linier sampel sebagai estimasi persamaan regresi linier populasi. Misalkan estimasi parameter untuk $\beta_0, \beta_1, \dots, \beta_p$ pada persamaan (1) adalah $b_0, b_1, \dots, b_p$ dan e_i $(e_i = Y_i - \hat{Y}_i)$ merupakan *error* atau residual, maka model sampel untuk persamaan (1) adalah sebagai berikut:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi} + e_i \quad i = 1, 2, \dots, n \tag{2}$$

Model regresi linier berganda di atas dapat dinyatakan dalam bentuk matriks sebagai berikut:

$$Y_{(nx1)} = X_{(nx(p+1))} b_{((p+1)x1)} + e_{(nx1)} \tag{3}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & X_{11} & \dots & X_{p1} \\ 1 & X_{12} & \dots & X_{p2} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & X_{1n} & \dots & X_{pn} \end{bmatrix}, \beta = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_p \end{bmatrix},$$

$$e = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

dengan menurunkan jumlah kuadrat *error* secara parsial terhadap $\beta_0, \beta_1, \dots, \beta_p$ sama dengan nol sehingga diperoleh *estimator* kuadrat terkecil dari β sebagai berikut:

$$b = (X^T X)^{-1} X^T Y \tag{4}$$

3. Pencilan (*Outlier*)

Dalam suatu data seringkali ditemukan berbagai hal yang menyebabkan tidak terpenuhinya asumsi klasik regresi. Ketika asumsi ada yang tidak terpenuhi, maka penggunaan metode kuadrat terkecil akan memberikan kesimpulan yang kurang baik atau nilai estimasi bersifat bias dan interpretasi hasil yang diberikan menjadi tidak valid. Salah satu penyebab tidak terpenuhinya asumsi klasik regresi adalah adanya pencilan (*outlier*) pada pengamatan yang menyebabkan distribusi *error* tidak normal.

Pencilan (*outlier*) adalah pengamatan yang jauh dari pusat data yang mungkin berpengaruh besar terhadap koefisien regresi atau pengamatan yang menyimpang dari sekumpulan pengamatan yang lain. Pencilan (*outlier*)

dapat muncul karena kesalahan dalam memasukkan data, kesalahan pengukuran, analisis, atau kesalahan-kesalahan lain. Keberadaan dari pencilan (*outlier*) akan menyebabkan kesulitan dalam analisis data, sehingga perlu untuk dihindari. Permasalahan yang muncul akibat adanya pencilan (*outlier*) antara lain sebagai berikut:

1. *Error* yang besar dari model yang terbentuk $E(e_i) \neq 0$.
2. Variansi dari data akan menjadi lebih besar.
3. Taksiran interval akan memiliki rentang yang lebih besar.

Pencilan (*outlier)* ini dapat diketahui secara visual atau eksak dengan melakukan diagnosis *error* atau residual dari model regresi yang terbentuk. Secara eksak, pencilan (*outlier*) dapat dideteksi melalui perhitungan metode diagnosis *error* seperti *Studentized* Residual, nilai DFFITS dan *Cook's Distance*. Adapun ketentuan dalam pengambilan keputusan ada atau tidak adanya pengamatan pencilan (*outlier*) adalah sebagai berikut:

$$\text{Jika} \begin{cases} Leverage\ Values > \frac{2p}{n} \\ \text{DFFITS} > 2\sqrt{\frac{p}{n}} \\ Cook's\ Distance > \frac{4}{n} \end{cases} \text{Pencilan (}outlier\text{)}$$

Ket: n = jumlah pengamatan (sampel) ; p = jumlah parameter

4. Metode Regresi *Robust* M-*estimator*

Metode regresi *robust* adalah metode untuk mengestimasi koefisien regresi yang tidak peka terhadap adanya penyimpangan asumsi yang mendasarinya. M-*estimator*merupakan metode regresi *robust* yang sering digunakan. M-*estimator*dipandang dengan baik untuk mengestimasi parameter yang disebabkan oleh pencilan (*outlier*).Pada prinsipnya, M-*estimator* merupakan estimasi yang meminimumkan jumlah fungsi *error*ρ:

$$min \sum_{i=1}^{n} \rho(e_i)$$
$$= min \sum_{i=1}^{n} \rho\left(-\sum_{p=0}^{k} X_{pi}\beta_p\right) \quad (5)$$

Fungsi ρ dipilih sebagai representasi fungsi pembobot dari *error*. Untuk memperoleh suatu versi skala *invariant* dari *estimator* ini dilakukan dengan menyelesaikan persamaan:

$$min \sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}}\right)$$
$$= min \sum_{i=1}^{n} \rho\left(\frac{Y_i - \sum_{p=0}^{k} X_{pi}\beta_p}{\hat{\sigma}}\right)$$

$b_0, b_1, \ldots, b_p$ merupakan nilai *estimator* dari $\beta_0, \beta_1, \ldots, \beta_p$ dengan:

$$min \sum \rho(e_i^*)$$
$$= min \sum \rho\left(\frac{e_i}{\hat{\sigma}}\right)$$

estimasi untuk $\hat{\sigma}$ adalah:

$$\hat{\sigma} = \frac{MAD}{0.6745}$$
$$= \frac{median\{|e_i - median(e_i, \ldots, e_n)|\}}{0.6745}$$

Dengan meminimumkan persamaan (6), turunan parsial pertama dari ρ terhadap $\beta_p$sama dengan nol, yaitu:

$$\sum_{i=1}^{n} X_{pi}\psi\left(\frac{Y_i - \sum_{p=0}^{k} X_{pi}\beta_p}{\hat{\sigma}}\right) = 0, \quad p = 0, 1, \dots, k$$

Tukey *bisquare* (9)

$$w(e_i^*) = \begin{cases} \left[1 - \left(\frac{e_i^*}{c}\right)^2\right]^2 & |e_i^*| \le c \\ 0 & |e_i^*| > c \end{cases} \qquad c = 4.685$$

Dengan $\psi = \rho'$, mendefinisikan $w(e_i^*) = \frac{\psi(e_i^*)}{e_i^*}$ diperoleh:

$$\sum_{i=1}^{n} X_{pi} w_{i0}\left(Y_i - \sum_{p=0}^{k} X_{pi} b_{p0}\right) = 0, \quad p = 0, 1, \dots, k \qquad (10)$$

Dalam notasi matriks, persamaan (10) dapat dituliskan menjadi:

$$X^T W_0 X b = X^T W_0 Y \qquad (11)$$

W_0 adalah matriks diagonal *n x n* dari bobot dengan elemen-elemen diagonal $w_{1,0}, w_{2,0}, \dots, w_{n,0}$. Xadalah matriks variabel bebas berukuran $(n\ x\ (p+1))$ dan Y adalah matriks variabel terikat berukuran$(n\ x\ 1)$.
Sehingga estimator regresi *robust* dengan M-*estimator* (IRLS) untuk β adalah:

$$b_{l+1} = (X^T W_l X)^{-1}(X^T W_l Y) \qquad (12)$$

Fungsi pembobot untuk metode M-*estimator* dengan fungsi Huber dan Tukey *bisquare* adalah sebagai berikut:

Metode	Fungsi Pembobot	Interval
Huber	$w(e_i^*) = \begin{cases} 1 & \lvert e_i^*\rvert \le c \\ \frac{c}{\lvert e_i^*\rvert} & \lvert e_i^*\rvert > c \end{cases}$	c = 1.345

5. Tahapan Analisis Data

Pada makalah ini terdapat enam tahapan analisis data, yaitu:

1. Menentukan nilai estimasi parameter menggunakan metode kuadrat terkecil.
2. Uji asumsi klasik regresi yaitu uji asusmsi linearitas, normalitas, non multikolinearitas, non autokorelasi dan homoskedastisitas, untuk mengetahui adanya penyimpangan asumsi klasik regresi atau tidak.
3. Jika terdapat penyimpangan asumsi normalitas, lakukan pendeteksian pengamatan pencilan (outlier) dengan nilai *Studentized* Residual, nilai DFFITS dan *Cook's Distance.*
4. Menentukan nilai estimasi parameter menggunakan metode regresi *robust* M-*estimator*. Langkah-langkah dalam tahap ini adalah:
   1) Mengestimasi parameter atau koefisien regresi b menggunakan metode kuadrat terkecil, sehingga di dapatkan $\hat{Y}_{i,0}$ dan $e_{i,0} = Y_i - \hat{Y}_{i,0}$, $(i = 1, 2, \dots, n)$.
   2) Menentukan $\hat{\sigma}_0$ dan pembobot awal $w(e_i^*)$ sesuai dengan fungsi Huber dan Tukey *bisquare*.
   3) Disusun matriks pembobot berupa matriks diagonal W_0 dengan elemen $w_{1,0}, w_{2,0}, \dots, w_{n,0}$.
   4) Dihitung *estimator* koefisien regresi yaitu, $b_{Robust\ ke\ 1} = (X^T W_0 X)^{-1}(X^T W_0 Y)$.
   5) Dengan menggunakan $b_{\text{Robust ke 1}}$*dihitung pula* $\sum_{i=1}^{n}|Y_i - \hat{Y}_{i,1}|$ *atau* $\sum_{i=1}^{n}|e_{i,1}|$

6) Langkah 2 sampai 5 diulang sampai didapatkan $\sum_{i=1}^{n}|e_{i,n}|$ atau b_{Robust} konvergen, yaitu perubahan antara $\sum_{i=1}^{n}|e_{l+1}|$dan $\sum_{i=1}^{n}|e_{l}|$ atau perubahan $b_{Robust\ ke-(l+1)}$ dan $b_{Robust\ ke-l}$ lebih kecil dari 0.1%.

5. Lakukan pengujian signifikansi parameter dengan menggunakan uji statistik t dan uji statistik F untuk mengetahui tingkat keberartian atau signifikansi masing-masing parameter baik secara parsial maupun keseluruhan terhadap model regresi yang diperoleh.
6. Menentukan model terbaik antara fungsi Huber dan fungsi Tukey *bisquare* berdasarkan nilai $R^2_{adjusted}$ yang terbesar dan nilai *standard error* yang terkecil.

6. Studi Kasus

Sebagai studi kasus dipilih data produksi padi di Provinsi Jawa Barat tahun 2009 di setiap kota atau kabupaten (Badan Pusat Statistik). Persamaan yang digunakan dalam kasus ini adalah:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

dimana,
Y = jumlah produksi padi (ton)
X_1 = jumlah luas panen (hektar)
X_2 = jumlah luas irigasi teknik (hektar)
i = unit *coss section*, yaitu kabupaten atau kota di Provinsi Jawa Barat

- Hasil estimasi parameter metode kuadrat terkecil (*Ordinary Least Squares*) dengan menggunakan *software* R.2.14.1 fasilitas R-*Commander* adalah sebagai berikut:

```
>Model.OLS<-lm(Jumlah.Produksi.Padi ~ Luas.Panen + Luas.Irigasi.Teknis, data=ProduksiPadi2009)
> summary(Model.OLS)

Call:
lm(formula = Jumlah.Produksi.Padi ~ Luas.Panen + Luas.Irigasi.Teknis,data = ProduksiPadi2009)

Residuals:
   Min      1Q  Median      3Q     Max
-60762   -3300   -1541    4201   54546

Coefficients:
                  Estimate   Std. Error  t value Pr(>|t|)
(Intercept)       1570.5832  6466.6317   0.243    0.8103
Luas.Panen        5.6805     0.0971      58.503   <2e-16 ***
Luas.Irigasi.Teknis  0.5443  0.2855      1.907    0.0691
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 21240 on 23 degrees of freedom
Multiple R-squared: 0.9973,     Adjusted R-squared: 0.9969
F-statistic:  4251 on 3 and 23 DF,  p-value: < 2.2e-16
```

 ISSN 1979-8911

Berdasarkan hasil estimasi parameter luas panen dan luas irigasi teknis di atas, maka model regresi linier yang dapat dibentuk adalah sebagai berikut:

$$\hat{Y}_i = 1570.5832 + 5.6805\, X_{1i} + 0.5443\, X_{2i} \qquad (14)$$

- Hasil estimasi parameter metode regresi M-*estimator* fungsi Huber dengan menggunakan *software* R.2.14.1 fasilitas R-*Commander* adalah sebagai berikut:

```
>library(MASS)
>model.huber<-
rlm(Jumlah.Produksi.Padi~Luas.Panen+Luas.Irigasi.Teknis,data=ProduksiPadi2009)
>summary(model.huber)

Call:    rlm(formula=Jumlah.Produksi.Padi~Luas.Panen+Luas.Irigasi.Teknis,data    =
ProduksiPadi2009)

Residuals:
    Min          1Q        Median       3Q        Max
-67291.4     -4365.2      401.4       3872.2     51726.6


Coefficients:
                      Value          Std. Error    t value
(Intercept)         -439.1787      2818.4734     -0.1558
Luas.Panen              5.7641        0.0423     136.2004
Luas.Irigasi.Teknis     0.2761        0.1244       2.2194

Residual standard error: 3370 on 23 degrees of freedom
```

Berdasarkan hasil estimasi parameter luas panen dan luas irigasi teknis di atas, maka model regresi linier dengan menggunakan M-*estimator* fungsi Huber yang dapat dibentuk adalah sebagai berikut:

$$\hat{Y}_i = -439.1787 + 5.7641\, X_{1i} + 0.2761\, X_{2i} \qquad (15)$$

- :

- Hasil estimasi parameter metode regresi M-*estimator* fungsi Tukey *bisquare*dengan menggunakan *software* R.2.14.1 fasilitas R-*Commander* adalah sebagai berikut

```
>library(MASS)
>model.bisquare<-
rlm(Jumlah.Produksi.Padi~Luas.Pane
n+Luas.Irigasi.Teknis,
data=ProduksiPadi2009,psi=psi.bisqu
are)
>summary(model.bisquare,cor=F)


Call:rlm(formula=Jumlah.Produksi.P
adi~Luas.Panen+Luas.Irigasi.Teknis,
```

data=ProduksiPadi2009, psi = psi.bisquare)
Residuals:
Min 1Q Median 3Q Max
-71932.4 -2291.7 -261.5 2078.4 49072.9

Coefficients:
Value Std.Error t value
(Intercept) 110.6351 1398.9123 0.0791
Luas.Panen 5.8074 0.0210 276.4723
Luas.Irigasi.Teknis 0.1114 0.0618 1.8039

Residual standard error: 5976 on 23 degrees of freedom

Berdasarkan hasil estimasi parameter luas panen dan luas irigasi teknis di atas, maka model regresi linier dengan menggunakan M-*estimator* fungsi Tukey *bisquare* yang dapat dibentuk adalah sebagai berikut:

$$\hat{Y}_i = 110.6351 + 5.8074\ X_{1i} + 0.1114\ X_{2i} \quad (16)$$

Model dari masing masing metode telah diperoleh, yaitu model hasil dari metode kuadrat terkecil (14), model hasil dari metode M-*estimator* dengan fungsi pembobot Huber (15), dan model hasil dari metode M-*estimator* dengan fungsi pembobot Tukey *bisquare* (16). Selanjutnya dari ketiga model regresi linier tersebut akan ditentukan model regresi linier terbaik untuk kasus jumlah produksi padi di Provinsi Jawa Barat tahun 2009. Kriteria yang dapat digunakan untuk menentukan model regresi terbaik, yaitu dengan menggunakan nilai *standarderror* (*se*) dan nilai $R^2_{adjusted}$. Model terbaik memiliki nilai *standarderror* (se) terkecil dan $R^2_{adjusted}$ terbesar. Dengan menggunakan *software* R.2.14.1 fasilitas R-GUI dan R-*Commander* diperoleh hasil output nilai untuk *standarderror*(*se*) dan nilai $R^2_{adjusted}$ pada masing-masing metode adalah sebagai berikut:

Metode	Variabel	Nilai	*Standard Error*	$R^2_{adjusted}$
	Intersep	1570.5832		
OLS	X_1	5.6805	21241.24	0.9969
	X_2	0.5443		
	Intersep	-439.1787		
M-*estimator*	X_1	5.7641	3370	0.9970
(Huber)	X_2	0.2761		
	Intersep	110.6351		
M-*estimator*	X_1	5.8074	5976	0.9967
(Tukey bisquares)	X_2	0.1114		

Tabel 1: Perbandingan Nilai *Standard Error* dan $R^2_{adjusted}$

Dari tabel 1 di atas dapat dilihat bahwa nilai *standarderror*fungsi Huber lebih kecil dari nilai *standard error* fungsi Tukey *bisquare* yang berarti bahwa variansi yang dapat dijelaskan oleh model hasil metode M-*estimator* dengan fungsi Huber lebih kecil dibanding model hasil M-*estimator* dengan fungsi Tukey *bisquare*. Dan nilai $R^2_{adjusted}$fungsi Huber lebih besar dari $R^2_{adjusted}$ fungsi Tukey *bisquare*, dengan $R^2_{adjusted}$ = 0.9970 = 99.70% menunjukkan bahwa variansi total Y sebesar 99.70% dapat dijelaskan oleh luas panen (X_1) dan luas irigasi teknis (X_2) pada model produksi padi di Provinsi Jawa Barat tahun 2009.

Dengan demikian, metode yang paling baik dalam mengestimasi parameter atau koefisien regresi terhadap faktor-faktor yang mempengaruhi jumlah produksi padi di Provinsi Jawa Barat tahun 2009 adalah metode regresi *robust* dengan M-*estimator* menggunakan fungsi pembobot Huberyang menghasilkan model regresi linier setelah dilakukan uji – t dan uji F adalah sebagai berikut:

$$\hat{Y}_i = 5.7641\ X_{1i} + 0.2761\ X_{2i} \quad (17)$$

7. Kesimpulan

Model produksi padi di Provinsi Jawa Barat tahun 2009 yang paling baik adalah model dengan menggunakan metode regresi *robust* M-*estimator* atau *Iteratively Reweighted Least Squares* dengan menggunakan fungsi pembobot Huberkarena memiliki nilai *standarderror* (*se*) yang lebih kecil dan nilai $R^2_{adjusted}$ yang lebih besar. Dengan model regresi linier berganda untuk produksi padi di Provinsi Jawa Barat tahun 2009 sebagai berikut:

$$\hat{Y}_i = 5.7641\ X_{1i} + 0.2761\ X_{2i}$$

yang artinya: setiap peningkatan 1% luas panen akan meningkatkan jumlah produksi padi sebesar 5.7641%, dan setiap peningkatan 1% luas irigasi teknis akan meningkatkan jumlah produksi padi sebesar 0.2761%.

Daftar Pustaka

[1] Badan Pusat Statistika. 2010. Jabar Dalam Angka 2010.

[2] Chen, Colin. 2002. *Robust Regression and Outlier Detection with the ROBUSTREG Procedure*. Paper 265-267. SAS Institute: Cary, NC

[3] Fox, John. 2002. *Robuts Regression Appendix to An R and S-PLUS Companion to Applied Regression*.

[4] Gujarati, Damodar N. 2004. *Basic Econometrics*. The McGraw Hill-Companies.

[5] Hisani, Nurul. 2012. Model Regresi Linier Berganda Menggunakan Penaksir Parameter Regresi Robust M-*estimator* (*Iteratively Reweihgted Least Squares*) Melalui *Software* R (Studi Kasus: Produksi Padi di Provinsi Jawa Barat Tahun 2009). Bandung: Skirpsi S1 Matematika UIN SGD.

[6] J. Rousseeuw, Peter and M. Leroy Annick. 1987. *Robust Regression and Outlier Detection.* Canada: John Wiley & Sons Inc.

[7] Norman R. Draper, Harry Smith. 1998. *Applied Regression Analysis*. USA: John Wiley & Sons Inc.