SELANG KEPERCAYAAN UNTUK KOEFISIEN GARIS REGRESI LINEAR DENGAN METODE *LEAST MEDIAN SQUARES*[1]

Harmi Sugiarti
Jurusan Statistika FMIPA Universitas Terbuka

email: harmi@ut.ac.id

ABSTRAK

Adanya penyimpangan terhadap asumsi dasar, khususnya jika terdapat pengamatan pencilan *(outlier)*, penggunaan metode kuadrat terkecil (OLS) dalam menduga selang kepercayaan dapat mengakibatkan berkurangnya ketelitian dalam pendugaan selang bagi koefisien garis regresi. Tulisan ini bertujuan untuk membandingkan antara lebar selang kepercayaan koefisien garis regresi yang diperoleh dengan menggunakan metode OLS dan metode *least median square*s (LMS) jika terdapat pencilan (outlier). Penelitian ini dilakukan dengan menggunakan data simulasi dan data terapan berupa nilai Tugas Tutorial Online (Tuton), Nilai Partisipasi Tuton, dan nilai UAS mata kuliah Metode Statistik I (SATS4121). Hasil analisis menunjukkan Selang kepercayaan untuk koefisien regresi yang diperoleh dengan metode LMS lebih sempit dibanding dengan metode OLS jika data tidak mengandung pencilan, Sedangkan untuk data yang mengandung pencilan, metode LMS memberikan selang kepercayaan yang lebih sempit dibanding metode OLS, Pada data terapan yang sangat kecil sekali mengandung pencilan, metode OLS memberikan selang kepercayaan yang lebih sempit dibanding metode LMS.

Keywords**:** selang kepercayaan, metode *least median square*s, pencilan

PENDAHULUAN

Hubungan linear antara satu peubah respons dengan p peubah bebas dapat dimodelkan sebagai: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$, $i = 1, 2, ..., n$ dimana Y_i adalah nilai peubah respons pada pengamatan ke-*i*, X_i adalah nilai peubah bebas pada pengamatan ke-i, β_0, β_1, $\beta_2, ..., \beta_p$ adalah koefisien regresi yang tidak diketahui nilainya dan akan dicari nilai taksirannya.

Taksiran parameter dapat berupa taksiran titik atau taksiran selang *(interval)*. Selang kepercayaan adalah suatu kisaran nilai yang dianggap mengandung nilai parameter populasi yang sebenarnya. Batas bawah (B) dan batas atas (A) selang tersebut dihitung dari suatu sampel acak yang ditarik dari populasi bersangkutan. Oleh karena itu sebelum penarikan sampel dilakukan, B dan A merupakan besaran acak. Untuk setiap pilihan yang wajar atas kedua batas itu selalu ada peluang positif bahwa selang kepercayaannya akan gagal mencakup nilai parameter yang sebenarnya. Sebelum percobaan dilakukan, terlebih dahulu ditetapkan nilai koefisien kepercayaan *(confidence coefficient)*. Koefisien ini menetapkan peluang bahwa selang kepercayaan akan mencakup nilai parameter yang sebenarnya. Oleh karena itu kita menginginkan peluang tersebut sedekat mungkin dengan 1.

[1] Disampaikan dalam Seminar Nasional FMIPA Universitas Terbuka 2011 pada tanggal 11 Juli 2011

Misalkan kita memilih koefisien kepercayaan $(1-\alpha)$, maka selang kepercayaan yang dihasilkan akan dinamakan selang kepercayaan $(1-\alpha)100\%$ bagi suatu parameter. Nilai B dan A dikatakan menentukan selang kepercayaan $(1-\alpha)100\%$ bagi suatu parameter apabila memenuhi kriteria sebagai berikut:

a) $\mathrm{P}\left(\mathrm{B} \le \text{nilai parameter yang sebenarnya } \le A\right) \ge (1-\alpha)$

b) nilai-nilai B dan A dapat dihitung dari sampel yang telah diambil dari populasi.

Tingkat kepercayaan $(1-\alpha)100\%$ mempunyai arti apabila percobaan pengambilan sampel acak dengan ukuran tertentu yang sama dari suatu populasi dan perhitungan nilai B dan A diulang berkali-kali, maka $(1-\alpha)100\%$ dari selang kepercayaan yang dihasilkan akan mengandung nilai parameter yang sebenarnya $\left(B \le \mu \le A\right)$. Selang kepercayaan yang cukup baik adalah selang kepercayaan yang mempunyai lebar selang yang sempit dan persentase selang yang memuat parameter cukup besar (Koopmans, 1987).

Model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i,\ i = 1, 2, \ldots, n.$ atau dengan notasi matriks dapat ditulis sebagai: $\underset{nx1}{\boldsymbol{Y}} = \underset{nxp}{\boldsymbol{X}}\ \underset{px1}{\boldsymbol{\beta}} + \underset{nx1}{\boldsymbol{\varepsilon}}$ dapat ditaksir dengan metode kuadrat terkecil (OLS). Penaksir parameter yang diperoleh dengan metode OLS akan bersifat tak bias linear terbaik *(best linear unbiased estimator*) jika asumsi yang mendasari metode OLS dipenuhi. Asumsi regresi linier klasik tersebut antara lain adalah: (a) model regresi dispesifikasikan dengan benar, (b) faktor galat *(error)* menyebar normal dengan mean nol dan variansi tertentu, (c) tidak terjadi heteroskedastisitas pada ragam galat, (d) tidak terjadi multikolinieritas antara peubah bebas, (e) tidak ada autokorelasi dalam galat, dan (f) tidak ada pencilan *(outlier)*. Pada dasarnya, metode ini meminimumkan jumlah kuadrat simpangan Y dari nilai harapannya $\boldsymbol{E(Y)}$ yaitu meminimumkan $\boldsymbol{\varepsilon'\varepsilon} = \boldsymbol{(Y - X\beta)'(Y - X\beta)}$ sehingga dengan menyelesaikan persamaan normal $\left[\boldsymbol{X'X}\right]\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X'Y}\right]$ diperoleh penaksir OLS bagi $\boldsymbol{\beta}$ adalah: $\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X'X}\right]^{-1}\left[\boldsymbol{X'Y}\right]$,

$\boldsymbol{s}^2\left(\hat{\boldsymbol{\beta}}\right) = \hat{\boldsymbol{\sigma}}^2\left[\boldsymbol{X'X}\right]^{-1}$, dan $\hat{\boldsymbol{\sigma}}^2 = \dfrac{\left(\boldsymbol{Y'Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'Y}\right)}{n\text{ - }p\text{ -}1}$ (Draper & Smith,1992).

Apabila ada penyimpangan terhadap asumsi dasar, khususnya jika model regresi dibangun dari data yang mengandung pengamatan pencilan yang berpotensi sebagai pengamatan berpengaruh, maka penggunaan metode kuadrat terkecil (OLS) tidak dapat memberikan penaksir yang bersifat *best linear unbiased estimator*. Pengamatan pencilan adalah pengamatan dengan sisaan yang cukup besar, sedangkan pengamatan berpengaruh adalah pengamatan yang dapat mempengaruhi hasil pendugaan koefisien regresi, sehingga tindakan membuang pengamatan yang berpengaruh akan mengubah secara signifikan persamaan regresi serta kesimpulannya. Selain itu, penggunaan metode OLS dapat mengakibatkan berkurang-nya ketelitian dalam pendugaan selang bagi koefisien garis regresi, sementara tindakan membuang atau mengabaikan pengamatan pencilan yang berpotensi sebagai pengamatan berpengaruh bukanlah prosedur yang bijaksana. Pengamatan pencilan adakalanya memberikan informasi yang cukup berarti, misalnya karena pencilan timbul dari kombinasi keadaan yang tidak biasa yang mungkin saja sangat penting dan perlu diselidiki lebih lanjut (Draper & Smith, 1992).

Untuk mendeteksi adanya pengamatan pencilan terhadap nilai-nilai $\boldsymbol{X}$ nya, dapat dilakukan dengan melihat matriks dugaan *(hat matrix)* yang didefinisikan sebagai:

$$H = X\ (X'X)^{-1} X'$$

Unsur ke-i pada diagonal utama matriks dugaan yakni h_{ii} biasanya dinamakan pengaruh *(leverage)* kasus ke-i yang dapat diperoleh dari $h_{ii} = \underline{x}_i' \left[\boldsymbol{X'X}\right]^{-1} \underline{x}_i$, dimana $\underline{x}_i'$ adalah vektor baris ke-i dari matriks $\boldsymbol{X}$. Nilai h_{ii} terletak antara 0 dan 1 yang jumlahnya sama dengan p , yaitu banyaknya parameter regresi. Nilai *leverage* h_{ii} yang besar menunjukkan bahwa pengamatan ke-i berada jauh dari pusat semua pengamatan $\boldsymbol{X}$. Suatu nilai leverage h_{ii} biasanya dianggap besar apabila nilainya lebih dari dua kali rataan semua leverage $(2p/n)$. Pada dasarnya nilai h_{ii} yang semakin besar menunjukkan semakin besar potensinya berpengaruh dalam pendugaan parameter regresi (Myers, 1990).

Dalam Myers (1990), guna mendeteksi adanya pengamatan yang berpengaruh, dapat digunakan nilai perbedaan dugaan peubah tak bebas terbakukan (*DFFITS*) yang dirumuskan sebagai:

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\ \sqrt{h_{ii}}}$$

dimana: $\hat{y}_i =$ nilai pendugaan y_i , $\hat{y}_{i,-i} =$ nilai pendugaan y_i tanpa pengamatan ke-i, $s_{-i} =$ dugaan simpangan baku tanpa pengamatan ke-i dan $h_{ii} =$ unsur ke-i dari diagonal matriks dugaan. Jika p menyatakan banyaknya parameter dan n menyatakan banyaknya pengamatan, maka suatu pengamatan akan merupakan pengamatan berpengaruh dalam persamaan regresi apabila mempunyai nilai $\left|DFFITS\right|_i > 2\sqrt{(p/n)}$.

Guna mengatasi kelemahan dari metode OLS, perlu dicoba metode lain yang bersifat tidak sensitif terhadap pelanggaran asumsi-asumsi, yaitu metode regresi *robust (robust regression)*. Ada beberapa metode pendugaan/ penaksiran koefisien garis regresi yang bersifat *robust* telah dikembangkan, diantaranya adalah metode pendugaan parameter regresi berdasarkan pada penduga *least median of square* (LMS). Metode LMS mempunyai keuntungan untuk mengurangi pengaruh dari sisaan *(residual)*. Menurut Rousseeuw dan Leroy (2003), penduga LMS diperoleh dengan mencari model regresi yang meminimumkan median dari *h* kuadrat sisaan (e_i^2) atau didefinisikan sebagai:

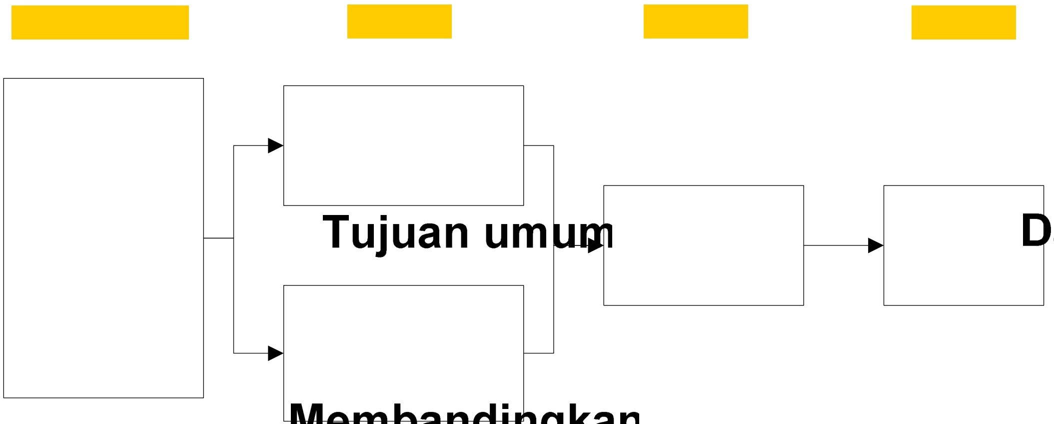
$$\hat{\beta}_{LMS} = \arg\min_{\beta}\ \underset{i}{median}\ e_i^2 \quad \text{dengan} \quad e_i^2 = \left(y_i - x_i^T\theta\right)^2; i = 1,2,...,n$$

Ukuran sebaran dari galat dapat ditaksir dengan cara menentukan terlebih dahulu nilai awal $s^0 = 1,4826\left[1 + 5/(n-p)\right]\sqrt{\underset{i}{median}\ e_i^2(\hat{\beta})}$. Faktor $1,4826 = \dfrac{1}{\Phi^{-1}(0,75)}$ diusulkan karena $\dfrac{median_i\ |z_i|}{\Phi^{-1}(0,75)}$ merupakan penaksir konsisten untuk σ jika z_i berdistribusi $N\left(0,\sigma^2\right)$. Selanjutnya nilai awal s^0 digunakan untuk menentukan pembobot w_i untuk setiap pengamatan, yaitu $w_i = \begin{cases} 1 & \text{jika } |e_i/\hat{\sigma}| \le 2,5 \\ 0 & \text{jika } |e_i/\hat{\sigma}| > 2,5 \end{cases}$ Berdasarkan pembobot w_i , nilai akhir taksiran σ untuk regresi LMS dihitung oleh $\sigma^* = \sqrt{\left(\sum_{i=1}^{n} w_i\ e_i^2\right) \Big/ \left(\sum_{i=1}^{n} w_i - p\right)}$.

Tulisan ini bertujuan untuk membandingkan lebar selang kepercayaan koefisien garis regresi yang diperoleh dengan menggunakan metode OLS dan metode LMS jika terdapat pencilan.

METODE

Data yang digunakan dalam kajian ini adalah data simulasi dan data terapan berupa nilai Tugas Tutorial Online (Tuton), Nilai Partisipasi Tuton, dan nilai UAS mata kuliah Metode Statistik I (SATS4121) masa registrasi 2007.1–2010.1. Desain penelitian disajikan pada Gambar 1 sebagai berikut.



Gambar 1. Desain I

Adapun langkah-langkah yang akan dilakukan dalam penelitian ini adalah:

1. Membangkitkan data berpasangan $(X_1, X_2, \varepsilon, Y)$ dimana (X_1, X_2) sebagai peubah bebas dan Y sebagai peubah tak bebas dengan $Y_i = 1 + X_{1i} + X_{2i} + \varepsilon_i$ serta $\varepsilon \sim NIID(0, \sigma^2)$.
2. Mendapatkan pengamatan pencilan dengan mengganti sejumlah tertentu pengamatan Y dengan nilai ekstrim sedemikian sehingga diperoleh pengamatan pencilan yang berpengaruh.
3. Menghitung lebar selang kepercayaan untuk koefisien garis regresi menggunakan metode OLS untuk data simulasi dengan atau tanpa pencilan .
4. Menghitung lebar selang kepercayaan untuk koefisien garis regresi menggunakan metode LMS untuk data simulasi dengan atau tanpa pencilan .
5. Menghitung lebar selang kepercayaan untuk koefisien garis regresi menggunakan metode OLS untuk data terapan.
6. Menghitung lebar selang kepercayaan untuk koefisien garis regresi menggunakan metode LMS untuk data terapan.

HASIL DAN PEMBAHASAN

Berdasarkan data simulasi tanpa pencilan, taksiran koefisien garis regresi, simpangan baku dari koefisien garis regresi, dan lebar selang 95% koefisien garis regresi yang diperoleh dengan metode OLS dan metode LMS masing-masing ditunjukkan oleh Tabel 1.

Tabel 1. Lebar Selang Kepercayaan 95% untuk Koefisien Garis Regresi untuk Data Simulasi Tanpa Pencilan

Koefisien Regresi	OLS			LMS		
	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**
$\hat{\beta}_0$	0,510	0,344	0,673	1,107	0,545	1,067
$\hat{\beta}_1$	0,973	0,049	0,097	0,969	0,124	0,243
$\hat{\beta}_2$	1,130	0,051	0,099	0,969	0,093	0,182

Metode OLS memberikan selang kepercayaan yang lebih sempit dibanding metode LMS, namun jika dilihat dari nilai taksiran untuk koefisien garis regresi, metode LMS memberikan taksiran yang lebih sesuai dibanding metode OLS, dimana nilai $\hat{\boldsymbol{\beta}}$ mendekati 1.

Pada data simulasi dengan 5% pencilan, metode OLS dan metode LMS masing-masing memberikan taksiran koefisien garis regresi, simpangan baku dari koefisien garis regresi, dan lebar selang 95% koefisien garis regresi seperti tampak pada Tabel 2; dimana metode LMS memberikan selang kepercayaan untuk koefisien garis regresi yang lebih sempit dibanding metode OLS. Selain itu, metode LMS juga memberikan taksiran koefisien garis regresi yang mendekati nilai 1 sebagaimana telah didesain pada metode simulasi sebelumnya, yaitu $\beta_0 = \beta_1 = \beta_2 = 1$ atau $\boldsymbol{\beta} = 1$.

Tabel 2. Lebar Selang Kepercayaan 95% untuk Koefisien Garis Regresi untuk Data Simulasi dengan 5% Pencilan

Koefisien Regresi	OLS			LMS		
	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**
$\hat{\beta}_0$	2,659	1,139	2,232	0,994	0,332	0,650
$\hat{\beta}_1$	0,805	0,164	0,321	0,925	0,046	0,091
$\hat{\beta}_2$	0,969	0,168	0,329	1,017	0,048	0,094

Tidak jauh dari hasil simulasi dengan 5% pencilan, pada data simulasi dengan 10% pencilan, metode LMS memberikan selang kepercayaaan yang lebih sempit dibanding metode OLS, serta metode LMS juga memberikan taksiran koefisien garis regresi yang mendekati nilai 1 sebagaimana telah didesain pada metode simulasi sebelumnya, yaitu $\beta_0 = \beta_1 = \beta_2 = 1$ atau $\boldsymbol{\beta} = 1$. Secara rinci, metode OLS dan metode

LMS masing-masing memberikan taksiran koefisien garis regresi, simpangan baku dari koefisien garis regresi, dan lebar selang 95% koefisien garis regresi untuk data simulasi dengan 10% pencilan dapat dilihat pada Tabel 3.

Tabel 3. Lebar Selang Kepercayaan 95% untuk Koefisien Garis Regresi untuk Data Simulasi dengan 10% Pencilan

Koefisien Regresi	**OLS**			**LMS**		
	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**
$\hat{\beta}_0$	4,015	1,454	2,850	2,137	0,335	0,657
$\hat{\beta}_1$	0,774	0,209	0,409	0,866	0,044	0,086
$\hat{\beta}_2$	0,830	0,215	0,420	0,885	0,046	0,091

Pada data terapan yang mengandung sekitar 1% pencilan (pengamatan yang berpengaruh), metode OLS memberikan selang kepercayaan untuk koefisien garis regresi yang lebih sempit dibanding metode LMS. Hal ini sesuai dengan hasil analisis untuk data simulasi tanpa pencilan. Secara rinci, metode OLS dan metode LMS masing-masing memberikan taksiran koefisien garis regresi, simpangan baku dari koefisien garis regresi, dan lebar selang 95% koefisien garis regresi untuk data terapan dapat dilihat pada Tabel 4.

Tabel 4. Lebar Selang Kepercayaan 95% untuk Koefisien Garis Regresi untuk Data Terapan

Koefisien Regresi	**OLS**			**LMS**		
	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**	$\hat{\beta}$	$s(\hat{\beta})$	**Lebar Selang**
Constant	35,510	2,104	4,124	30,109	1,655	3,244
Tugas 1	0,005	0,037	0,073	0,016	0,028	0,054
Tugas 2	0,003	0,051	0,101	0,007	0,037	0,073
Tugas 3	-0,057	0,051	0,099	-0,079	0,037	0,072
Aktivasi	0,092	0,068	0,134	0,184	0,052	0,102

KESIMPULAN

Selang kepercayaan untuk koefisien regresi yang diperoleh dengan metode LMS lebih sempit dibanding dengan metode OLS jika data tidak mengandung pencilan, Sedangkan untuk data yang mengandung pencilan, metode LMS memberikan selang kepercayaan yang lebih sempit dibanding metode OLS, Pada data terapan yang sangat kecil sekali mengandung pencilan, metode OLS memberikan selang kepercayaan yang lebih sempit dibanding metode LMS.

DAFTAR PUSTAKA

- Draper, N,R, & Smith, H, 1992, *Applied regression analysis, 2nd ed*, Wiley, New York,
- Koopmans, L,H, 1987, *Introduction to Contemporary Statistical Methods*, 2nd ed, Boston: PWS,
- Myers, R,H, 1990, *Classical and modern regression with applications*, 2nd ed, PWS-Kent, Boston,
- Rousseeuw,P,J, & Leroy,A,M, 2003, *Robust regression and outlier detection*, Wiley, New York,

KEMBALI KE DAFTAR ISI