PENGOPTIMALAN SOFTWARE S-PLUS GUNA ESTIMASI MODEL REGRESI UNTUK DATA DENGAN KESALAHAN PENGUKURAN MENGGUNAKAN METODE BAYES

Hartatik,M.Si
Program Studi DIII Teknik Informatika Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Sebelas Maret Surakarta
Jl. Ir. Sutami 36 A Surakarta 57126
Telp. (0271) 663450
E-mail: d3ilkom@uns.ac.id, hartatik@uns.ac.id

ABSTRAK

Misal diberikan suatu data (X_i,Y_i), maka model regresinya adalah

$$Y_i = g(X_i) + \varepsilon_i$$

*di mana Xi adalah elemen ke- i dari variabel prediktor **X** dan Yi adalah elemen ke- i dari variabel respon **Y**. Variabel **X** yang merupakan variabel prediktor dari hasil pengamatan biasanya merupakan konstanta tertentu, namun terkadang juga dijumpai **X** yang merupakan variabel random atau variable dimana nilainya bukan konstanta tetap . Untuk itulah dalam hal ini model regresinya disebut dengan model regresi dengan kesalahan pengukuran. Ada dua metode Pendekatan yaitu parametrik dan Nonparametrik. Dalam penelitian ini untuk pendekatan parametrik digunakan Ordinary Least Square (OLS), dan untuk Nonparametrik bila kesalahan pengukuran diabaikan digunakan metode B-spline dan bila kesalahan pengukuran tidak diabaikan digunakan Metode Iterative Conditional Modes (ICM)..Dan pemanfaatan Software yang tepat serta pengembangannya dapat memberikan hasil estimasi yang bagus, dan dalam penelitian kali ini dengan menggunakan S-Plus.*

Kata Kunci: Bayes, ICM, kesalahan pengukuran, regresi, Software S-Plus.

1. LATAR BELAKANG

Dalam kehidupan sehari-hari banyak sekali kejadian yang bisa dijelaskan dalam suatu kurva regresi. Kurva regresi adalah kurva yang menjelaskan hubungan antara suatu variabel prediktor, **X**, dan variabel respon, **Y**. Misal diberikan suatu data (**X**,**Y**), model regresinya dapat dituliskan sebagai

$$Y = g(\mathrm{X}) + \varepsilon \quad (1)$$

di mana **X** adalah variabel prediktor, **Y** variabel respon, **g** adalah suatu fungsi tertentu, dan **ε** adalah sesatan random independent dengan mean nol dan variansi σ_ε^2.

Variabel **X** yang merupakan variabel prediktor dari hasil pengamatan biasanya diasumsikan sebagai variabel tetap (*fixed variable*). Namun kenyataannya, sering dijumpai **X** yang bukan *fixed variable* tetapi variabel random atau variabel **X** diukur dengan kesalahan (*error in variable*). Namun hal ini sering diabaikan untuk alasan praktis dan kemudahan perhitungan. Sebagai contoh, ingin diamati masalah pendapatan. Jika responden yang diwawancara tidak bisa menyebutkan pendapatannya secara tepat, tentunya hasil catatan penelitian akan lebih tinggi atau lebih rendah dari nilai yang sebenarnya. Hal ini dikenal dengan kesalahan pengukuran. Kasus ini bisa dijumpai diantaranya dalam masalah epidemiologi, geologi, dan survival.

Kesalahan pengukuran (*measurement error*) adalah kesalahan yang muncul manakala suatu nilai dicatat tidak persis sama dengan nilai sebenarnya dalam kaitan dengan suatu proses pengukuran. Sehingga berkaitan dengan definisi ini, ada 3 variabel di dalam model kesalahan pengukuran, yaitu variabel yang menyatakan data hasil pengamatan, variabel yang menyatakan data sesungguhnya yang tidak terukur, dan variabel kesalahan pengukuran. Secara matematis, model kesalahan pengukuran dapat dituliskan sebagai berikut

$$W = X + U \quad (2)$$

dengan **W** adalah variabel yang menyatakan hasil pengamatan yang disebut dengan variabel pengganti (*surrogate*), **X** adalah variabel prediktor yang tidak teramati (*latent variable*), dan **U** adalah variabel kesalahan pengukuran yang diasumsikan normal independen dengan mean 0 dan variansi $\sigma_{\mathbf{u}}^2$.

Berdasarkan model regresi (1) dan (2), ada dua pendekatan digunakan untuk menduga kurva regresi yaitu metode regresi parametrik dan nonparametrik. Metode regresi parametrik merupakan metode yang sering digunakan untuk menduga kurva regresi. Namun metode regresi parametrik memiliki keterbatasan untuk menduga pergerakan data yang tidak diharapkan. Jika salah satu asumsi dari metode regresi parametrik tidak dipenuhi, maka kurva regresi dapat diduga dengan menggunakan metode regresi nonparametrik.

Tujuan dari penelitian ini adalah untuk menduga kurva regresi bila ada kesalahan pengukuran dalam data. Metode yang digunakan adalah untuk pendekatan parametrik dengan OLS, sedangkan untuk pendekatan Nonparametrik digunakan *Naïve Method* dan Metode ICM.

2. MODEL KESALAHAN PENGUKURAN

Menurut Carrol, et al (1995), kesalahan yang muncul manakala suatu nilai dicatat tidak persis sama dengan nilai sebenarnya dalam kaitan dengan suatu kekurangan di dalam proses pengukuran disebut dengan kesalahan pengukuran. Kesalahan pengukuran erat kaitannya dengan *realibility* dari suatu alat ukur.

Setiap alat ukur harus memiliki kemampuan untuk memberikan hasil pengukuran yang konsisten. Pada alat ukur fenomena fisik seperti berat badan, tinggi badan, konsistensi hasil pengukuran bukanlah hal yang sulit untuk dicapai, namun perlu juga untuk mempertimbangkan kesalahan pengukuran akibat peralatan yang digunakan. Khususnya alat ukur alternatif yang digunakan di dalam dunia kesehatan, seperti PET Scanner yang merupakan alat alternatif untuk mendeteksi adanya stroke dengan jalan mengukur aliran darah dalam otak dengan lebih aman dibandingkan dengan alat ukur angiogram yang beresiko kematian. Untuk itulah, perlu kiranya untuk hati-hati di dalam melakukan pengukuran karena peralatan pengukuran yang sudah baku itu bisa merupakan sumber dari kesalahan pengukuran.

Tidak seperti di dalam pengukuran fisik yang mungkin sudah ada peralatan baku yang bisa digunakan, dalam pengukuran fenomena sosial seperti sikap, opini, dan persepsi pengukuran yang konsisten agak sulit dicapai, karena tidak ada peralatan yang baku yang bisa mengukur variabel itu. Bisa saja orang yang sama akan memberikan jawaban yang berbeda dengan alat ukur (pernyataan) yang sama. Hal ini merupakan salah satu sumber adanya kesalahan pengukuran.

Setiap hasil pengukuran khususnya fenomena sosial merupakan kombinasi antara hasil pengukuran dengan kesalahan pengukuran. Model kesalahan pengukuran dituliskan dalam bentuk umum sebagai

$$\mathbf{W} = \gamma_0 + \gamma_1 \mathbf{X} + \gamma_2 \mathbf{Z} + \mathbf{U},$$

dengan Z adalah variabel instrumen.

Selain itu, model kesalahan pengukuran yang lebih sederhana (model kesalahan pengukuran klasik) dapat dituliskan seperti pada (2) dengan elemen-elemennya sebagai berikut

$$W_{ij} = X_i + U_{ij},\ i=1,\ldots,n \text{ dan } j=1,\ldots,m_i.$$

dengan

W_{ij} : nilai yang diperoleh dari hasil pengamatan (*Surrogate Variable*)

X_i : nilai sebenarnya yang tidak teramatai (*Latent Variable*)

Uij : kesalahan pengukuran.

Makin kecil kesalahan pengukuran, makin reliabel suatu alat ukur. Sebaliknya makin besar kesalahan pengukuran makin tidak reliabel suatu alat ukur. Besar kecilnya pengukuran dapat dilihat dari korelasi antara pengukuran pertama dan kedua. Bila angka korelasi dikuadratkan maka didapatkan koefisien determinasi yang merupakan petunjuk besarnya hasil pengukuran yang sebenarnya. Sebagai contoh jika, r =0.9 maka r^2 =0.81, ini berarti bahwa 81% merupakan hasil pengukuran sebenarnya dan 19% menunjukkan besarnya kesalahan pengukuran. Sehingga sangatlah perlu untuk mempertimbangkan adanya kesalahan pengukuran dalam analisis statistik sehingga didapatkan proses inferensi yang lebih baik.

3. EFEK KESALAHAN PENGUKURAN

Misalkan suatu model regresi $Y = g(X) + \varepsilon$, dengan $\mathbf{g(X)} = \mathbf{X\beta}$, $\beta = (\beta_0, \ldots, \beta_k)$.

Berikut pengaruh dari adanya kesalahan pengukuran.

1. Terhadap mean :

$$E(W) = E(X + U) = E(X)$$

Kesalahan pengukuran tidak menyebabkan bias di dalam nilai harapan.

2. Terhadap variansi:

$$\mathrm{Var}(W) = \mathrm{Var}(X + U) = \mathrm{Var}(X) + \mathrm{Var}(U) - 2\mathrm{Cov}(X, U)$$

.

Kesalahan pengukuran menyebabkan bias di dalam nilai variansi.

1. Terhadap kovariansi

$$\mathrm{Cov}(W, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(U, Y) = \mathrm{Cov}(X, Y).$$

4. Terhadap slope kurva regresi

Dalam hal ini diambil regresi linear sederhana yang dituliskan sebagai berikut

$$\begin{aligned} Y &= \mathrm{bo} + \mathrm{b}_1 X + \varepsilon \\ Y &= \mathrm{bo} + \mathrm{b}_1 (W - U) + \varepsilon \\ Y &= \mathrm{bo} + \mathrm{b}_1 W + v \end{aligned} \qquad (3)$$

dengan $v = \varepsilon - U$.

Estimasi b_1 untuk persamaan (3) adalah sebagai berikut

$$\begin{aligned} \hat{\mathrm{b}}_1 &= \frac{\mathrm{Cov}(W, Y)}{\mathrm{Var}(W)} = \frac{\mathrm{Cov}(W, (\mathrm{bo} + \mathrm{b}_1 X + \varepsilon))}{\mathrm{Var}(W)} \\ &= \frac{\mathrm{Cov}(W, (\mathrm{bo} + \mathrm{b}_1 (W - U) + \varepsilon))}{\mathrm{Var}(X)} \\ &= \frac{\mathrm{Cov}(W, (\mathrm{bo} + \mathrm{b}_1 (W) + v))}{\mathrm{Var}(W)} \\ &= \mathrm{b}_1 + \frac{\mathrm{Cov}(W, v)}{\mathrm{Var}(W)} \end{aligned} \qquad (4)$$

Dalam permasalahan model regresi dengan kesalahan pengukuran, maka nilai

$$\begin{aligned}\mathrm{Cov}(W,v) &= \mathrm{Cov}(X+U, \varepsilon - \mathrm{b}_1 U) \\ &= \mathrm{Cov}(X,\varepsilon) + \mathrm{Cov}(X,\mathrm{b}_1 U) + \mathrm{Cov}(U,\varepsilon) + \mathrm{Cov}(U,-\mathrm{b}_1 U)\end{aligned} \quad (5)$$

Diasumsikan *u* dan ε independen satu dengan yang lainnya, sehingga

$$\mathrm{Cov}(X,\mathrm{b}_1 U) + \mathrm{Cov}(U,\varepsilon) + \mathrm{Cov}(U,-\mathrm{b}_1 U) = 0$$

dan

$$\mathrm{Cov}(X,v) = \mathrm{Cov}(U,-\mathrm{b}_1 U) = -b_1 \mathrm{Var}(U) \quad (6)$$

Karena $\mathrm{Cov}(X,v) \neq 0$, maka persamaan (4) menjadi

$$\begin{aligned}\hat{\mathrm{b}}_1 &= \mathrm{b}_1 + \frac{\mathrm{Cov}(X,v)}{\mathrm{Var}(X)} \\ &= \mathrm{b}_1 + \frac{-\mathrm{b}_1 \mathrm{Var}(U)}{\mathrm{Var}(X)} \neq \mathrm{b}_1\end{aligned} \quad (7).$$

Dan berdasarkan persamaan (7) maka $E(\hat{b}_1) \neq b_1$. Jadi estimator b_1 tidak memenuhi **sifat unbias**.

Selanjutnya, akan dikaji estimator yang lebih baik bila ada kesalahan pengukuran dalam data:

a. Asumsi-asumsi:

a.1 vektor error (ε_i, U_i) , i=1,2,3,…,n dimana U_i adalah elemen baris ke-i dari **U**, variabel random yang berdistribusi Normal independen dengan mean nol dan mempunyai matrik variansi sebagai berikut

$$\begin{pmatrix} \sigma_\varepsilon^2 & \Sigma_{\varepsilon \mathrm{U}} \\ \Sigma_{\mathrm{U}\varepsilon} & \Sigma_{\mathrm{UU}} \end{pmatrix} = \mathrm{diag}(\sigma_{\varepsilon\varepsilon}, \sigma_{UU_{11}}, ..., \sigma_{UU_{kk}}).$$

a.2 Distribusi dari (ε_i, U_i) adalah independen dari X, untuk semua *i* dan *j* , dimana X_i adalah baris ke-*i* dari **X**.

a.3 X*i* adalah variabel normal independen berdistribusi normal independen dengan mean nol dan matrik kovariansi singular $\boldsymbol{\Sigma}_{\mathrm{xx}}$.

Jika **X** dan **U** berdistribusi normal, maka **W** variabel random berdistribusi normal independen dengan mean nol dan matrik kovariansi

$$\boldsymbol{\Sigma}_{\mathbf{WW}} = \boldsymbol{\Sigma}_{\mathbf{XX}} + \boldsymbol{\Sigma}_{\mathbf{UU}}.$$

Misalkan rasio variansi *error* dengan variansi total dinotasikan dengan λ_{jj} dimana

$$\lambda_{jj} = \sigma_{\mathbf{WW}}^{-1} \sigma_{\mathbf{UU}}$$

dan $\sigma_{\mathbf{WW}}$ adalah varians dari **W**. dan 1- λ_{jj} adalah reliabilitas variabel ke-*j*.

a.4 Diasumsikan bahwa λ_{jj} diketahui. Matrik diagonal dari rasio dituliskan dengan

$$\boldsymbol{\Lambda}_{\mathbf{UU}} = \mathrm{diag}(\lambda_{11}, \lambda_{22}, \lambda_{33} ..., \lambda_{\kappa\kappa}),$$

b. Didefinisikan estimator $\hat{\boldsymbol{\beta}}$ setelah dikalikan dengan faktor koreksi, yaitu

$$\hat{\beta} = H^{-1}(\mathrm{n}^{-1} W^{\mathrm{T}} Y),$$

dengan $H = \mathrm{n}^{-1}(W^{\mathrm{T}} W) - D\Lambda_{UU} D$

Selanjutnya akan dikaji tentang sifat sampel besar untuk estimasi $\hat{\boldsymbol{\beta}}$.

$$\hat{\beta} = \beta + (\mathrm{n}^{-1} W^{\mathrm{T}} W - D\Lambda_{UU} D)^{-1} [\mathrm{n}^{-1}(W^{\mathrm{T}} v) + D\Lambda_{UU} D\beta]$$

$$\mathrm{n}^{1/2}(\hat{\beta} - \beta) = (\mathrm{n}^{-1} W^{\mathrm{T}} W - D\Lambda_{UU} D)^{-1} \left[\mathrm{n}^{-1/2}(W^{\mathrm{T}} v) + \mathrm{n}^{1/2} D\Lambda_{UU} D\beta \right] \quad (8)$$

4. B-SPLINE

Jika terdapat spline dengan orde *m* dan kumpulan knots yang memenuhi $a < \xi_1 < ... < \xi_k < b$. Dapat didefinisikan sejumlah *2m* knot tambahan $\xi_{-(m-1)}, ..., \xi_1, \xi_0, \xi_{k+1}, ..., \xi_{k+m}$ dimana

$$\xi_{-(m-1)}, ..., \xi_0 = a \quad \text{dan} \quad \xi_{k+1}, ..., \xi_{k+m} = b.$$

B-spline orde m yang sesuai untuk knot $\xi_1, ..., \xi_k$ dinyatakan dengan

$$\mathrm{B}_{j,m}(\mathrm{X}) = \frac{\mathrm{X} - \xi_j}{\xi_{i+m-1} - \xi_j} \mathrm{B}_{j,m-1}(\mathrm{X}) + \frac{\xi_{j+m} - \mathrm{X}}{\xi_{j+m} - \xi_{j+1}} \mathrm{B}_{j+1,m-1}(\mathrm{X})$$

,

$$\text{dengan } \mathrm{B}_{j,1}(X) = \begin{cases} 1, X \in [\xi_j, \xi_{j+1}) \\ 0, \text{untuk yang lain.} \end{cases} \quad (16)$$

Misal diasumsikan bahwa

$$\hat{\mathrm{g}}(\mathrm{X}) = \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{B}_{\mathrm{j,m}}(\mathrm{X}) \beta_{\mathrm{j}},$$

dengan $B_1(X),...,B_p(X)$, adalah basis untuk vektor spline berderajat *m* dengan titik knot $\xi_1,...,\xi_k$ dan $p=m+k$. Maka

$$S(g)=\sum_{i=1}^{n}\{Y_i-\hat{g}(X_i)\}^2+\alpha\int_a^b\{g^{(2)}(x)\}^2dx$$

$$S(g)=\sum_{i=1}^{n}\left\{Y_i-\sum_{j=1}^{p}B_{j,m}(X_i)\beta_j\right\}^2+\alpha\int_a^b\left\{\sum_{i=1}^{p}\beta_i B_i''(x)\right\}^2dx$$

Sehingga estimator *Penalized Least Square* di atas dapat dituliskan dengan

$$S(g)=(Y-B(X)\beta)^T(Y-B(X)\beta)+\alpha\beta^T D\beta\,, \quad (17)$$

dengan elemen ke-*ij* dari **D** adalah

$$D=\int_a^b\{B_i''(x)B_j''(x)\}\,dx\,.$$

Maka penyelesaian untuk (17) adalah

$$\frac{\partial S(g)}{\partial\beta}=\frac{\partial\{Y-B(X)\beta\}^T\{Y-B(X)\beta\}+\alpha\beta^T D\beta}{\partial\beta}$$

$$0=-2(B(X))^T Y+2(B(X))^T B(X)\beta+2\alpha D\beta$$

$$0=-(B(X))^T Y+\left((B(X))^T B(X)+\alpha D\right)\beta\quad.$$

$$\left((B(X))^T B(X)+\alpha D\right)\beta=(B(X))^T Y$$

$$\beta=\left((B(X))^T B(X)+\alpha D\right)^{-1}(B(X))^T Y$$

Sehingga penduga fungsi g adalah

$$\hat{g}(X)=\sum_{j=1}^{p}B_{j,m}(X)\beta_j\,,p=m+k \qquad (18)$$

dengan

$$\beta=\left((B(X))^T B(X)+\alpha D\right)^{-1}(B(X))^T Y\,.$$

B-spline dalam penelitian ini, diaplikasikan untuk data bila kesalahan pengukuran diabaikan, dan juga bila kesalahan pengukuran diperhitungkan dalam model. Estimasi kurva regresi bila kesalahan pengukuran diabaikan yaitu meregresikan antara **W** dan **Y**. Di samping itu juga akan digunakan spline untuk estimasi kurva regresi dengan *Naïve Method*, yaitu dengan meregresikan antara rata-rata **W** dan **Y**.

5. METODE ICM UNTUK MODEL DENGAN KESALAHAN PENGUKURAN

Untuk mengestimasi fungsi **g** dalam model kesalahan pengukuran dengan Metode ICM diperlukan informasi prior sebelum menentukan distribusi posterior yang selanjutnya digunakan untuk iterasi dalam Metode ICM.

Misalkan data observasi diasumsikan berdistribusi normal,

$$Y\sim N(g(X_i),\sigma_Y{}^2)$$

$$W\sim N(X_i,\sigma_W{}^2)$$

Distribusi Prior untuk **X** dan **g** yang digunakan untuk model kesalahan pengukuran (1) adalah sebagai berikut

$$X\sim N(\mu_X,\sigma_X{}^2)$$

dimana μ_X dan σ_X^2 adalah suatu konstanta yang ditentukan.

Sedangkan untuk distribusi *prior* **g** dilakukan pendekatan "*partially improper*" (Green and Silverman, 1995) yaitu:

$$p(g)\propto\exp-\frac{\alpha}{2}(g^T Kg),$$

Prior untuk varians-varians yang digunakan dalam Metode ICM merupakan hasil dari estimasinya, yaitu sebagai berikut:

$$\hat{\sigma}_U^2=\frac{\sum_{i=1}^{n}(m_i-1)s_i^2}{\sum_{i=1}^{n}(m_i-1)}\,,$$

dan menurut Green and Silverman (1995), estimasi dari varians *error* adalah

$$\sigma_\varepsilon^2=\frac{\sum_{i=1}^{n}\{Y_i-g(X_i)\}^2}{tr\{I-A(\hat{\alpha})\}}\,,$$

dengan $\hat{\alpha}$ diestimasi dengan metode GCV.

Pemilihan prior dimaksudkan guna membentuk distribusi posterior bersama. Parameter dalam model (1) yaitu $X, g, \sigma_X^2\sigma_\varepsilon^2$, dan σ_u^2. Misalkan, $\theta=(X,g,\sigma_\varepsilon^2,\sigma_X^2,\sigma_u^2)$ maka densitas posterior dari θ adalah

$$p(\theta|Y,W)\propto p(\theta)\,p(Y|\theta)\,p(W|\theta)$$

$$\propto p(X,g,\sigma_X^2,\sigma_\varepsilon^2,\sigma_U^2)\,p\left(Y\middle|g,X,\sigma_\varepsilon^2\right)p\left(W\middle|X,\sigma_u^2\right)$$

$$\propto p\left(Y\middle|g,X,\sigma_\varepsilon^2\right)p\left(W\middle|X,\sigma_u^2\right)p(X)\,p(g)\,p(\sigma_X^2)\,p(\sigma_U^2)\,p(\sigma_\varepsilon^2)$$

Distribusi posterior bersama dengan Metode ICM bersyarat pada $\hat{\sigma}_{\mathbf{X}}^2,\hat{\sigma}_\varepsilon^2,dan\,\hat{\sigma}_{\mathbf{u}}^2$ adalah

$$p(\theta|Y,W)\propto p\left(Y\middle|g,X,\sigma_\varepsilon^2\right)p\left(W\middle|X,\sigma_u^2\right)p(X)\,p(g)$$

$$p(\sigma_X^2)\,p(\sigma_U^2)\,p(\sigma_\varepsilon^2)\,p(\alpha)$$

$$\propto\exp\left[-\frac{1}{2\sigma_\varepsilon^2}(Y_i-g(X_i))^2-\frac{1}{2\sigma_W^2}(W_i-X_i)^2\right.$$

$$\left.-\frac{1}{2\sigma_X^2}(X_i-\mu_X)^2+-\frac{\hat{\alpha}}{2\sigma_\varepsilon^2}g^T Kg\right] \qquad (22)$$

Pendekatan ICM seperti yang dijelaskan oleh Besag (1986), yaitu dengan menentukan mode posterior dari (22). Dalam Metode ICM parameter diperbarui (*update*) dalam setiap iterasinya.

Berikut ini algoritma dari Metode ICM.

1. **Menentukan nilai awal** $\hat{\sigma}_{\mathbf{X}}^2, \hat{\sigma}_{\varepsilon}^2, \text{dan } \hat{\sigma}_{\mathbf{u}}^2$**, dan** $\mathbf{g}^{(0)}$

   $\mathbf{g}^{(0)}$ merupakan hasil estimasi dengan menggunakan *Naïve Method*, yaitu meregresikan antara nilai Y dengan rata-rata dari W. Nilai awal dari $\hat{\sigma}_{\mathbf{X}}^2, \hat{\sigma}_{\varepsilon}^2, \text{dan } \hat{\sigma}_{\mathbf{u}}^2$ diturunkan dari (19), (20), (21).
2. **Berdasarkan pada kondisi 1, menentukan data** $\mathbf{X}^{(i)}$ **bersyarat** $\mathbf{g}^{(i-1)}$ **yang memaksimalkan** (22)

   Nilai $\mathbf{X}^{(i)}$ yang memberikan nilai yang maksimum pada (22) tidak dapat diselesaikan secara analitik. Untuk itu, dalam hal ini, digunakan metode grid uniform untuk memaksimumkan (22).
3. **Menentukan vektor** $\mathbf{g}^{(i)}$ **bersyarat pada** $\mathbf{X}^{(i)}$**, dengan menggunakan Regresi Spline.**

   Berdasarkan pada (22), dengan bersyarat pada $\mathbf{X}^{(i)}$, memaksimumkan densitas posterior (22) sama artinya dengan meminimumkan

$$\left[\frac{1}{2\sigma_{\varepsilon}^2} \sum_{i=1}^{n} \left(Y_i - g(X_i)\right)^2 + \frac{\hat{\alpha}}{2\sigma_{\varepsilon}^2} \mathrm{g}^T \mathrm{Kg} \right]. \qquad (23)$$

   Bentuk (23) merupakan persamaan *Smoothing Spline*. Untuk itu, mengestimasi **g** dapat digunakan B-Spline
4. **Iterasi *i*=*i*+1diulang sampai *i* tertentu.**

6. PROGRAM S-PLUS

Pemberdayaan software SPSS untuk membuat model regresi dari data dengan kesalahan pengukuran adalah sbb:

Program ICM

```
icm
function(w,y,regress=F,ak=T,iter=10,
spar=0,grid=500,sigw=0){
n _ length(y)
k _ dim(w)[2]
wbar _ apply(w,1,mean)

# initial values of x:
x _ wbar

# initial estimate of sigw:
if (sigw ==0){
ssw _ apply(w,1,var)
sigw _ sqrt(mean(ssw))
}

#  xp  is  the  grid  values  over
x...keep track of the spline values
xp
seq(min(wbar),max(wbar),length=grid)

# Initial smoothing
fit
smooth.spline(x,y,all.knots=ak,spar=
spar,cv=T)
spar _ fit$spar
fit2 _ predict(fit,xp)
res <- (fit$yin - fit$y)/(1-fit$lev)

# initial estimate of sigy:
sigy _ sqrt(var(res))

for (i in 1:iter){

# find the condition post. value for
each x_i
    for (ii in 1:n){
        val
1/(2*sigy^2))*((y[ii] - fit2$y)^2)
val
(k/(2*sigw^2))*((mean(w[ii,])
xp)^2)
if (regress){
val
(1/(2*var(x)^2))*((mean(xp) - xp)^2)
}

# Set x_i to its max
x[ii] _ xp[val==max(val)]
    }

    fit
smooth.spline(x,y,spar=spar,all.knot
s=ak)
#  this  saves  the  predicted  values
for  each  grid  point  (the  height  of
y)
    fit2 _ predict(fit,xp)


fit21 _ predict(fit,xp)
res1  <-  (fit$yin  -  fit$y)/(1-
fit$lev)

#  estimate of sigy icm:
sigy1 _ sqrt(var(res1))


out
list(x=xp,y=fit2$y,s,xfit=x,sigy=sig
y,sigy1=sigy,sigw=sigw,spar=spart)
out
}
```

Simulasi Data Untuk:

Kasus 1:

```
#######here     is     an     example
simulation: (uncomment and run)
ff1 _ function(x)
{
  d  <-  sin(0.5  *  pi  *  x)/(  1  +
(sign(x) + 1)*(2*x^2))
 d
}

##### generate the x-y-w
n _ 100
x _ rnorm(n,0.5,0.25)
y _ ff1(x) + rnorm(n,0,.015)
repeats _ 2
w _ matrix(0,nrow=n,ncol=repeats)
for (j in 1:repeats){
 w[,j]          _          x         +
rnorm(n,0,(sqrt(3/7)*0.25) )}
wbar _ apply(w,1,mean)
xp                                    _
seq(min(wbar),max(wbar),length=500)
yp _ ff1(xp)

###### neat pictures (prints (x,y)
points, (wbar,y) , true curve,
###### naive, and icm spline
plot(x,y,pch=5,xlim=c(min(wbar),max(
wbar)),ylim=c(min(y),max(y)),
 xlab="",ylab="")
points(wbar,y,pch=18,col=2)
lines(xp,yp,lty=1,lwd=2)
ddd _ smooth.spline(wbar,y)
lines(ddd$x,ddd$y,lty=1,col=2,lwd=2)
im5 _ icm(w,y,iter=5)
lines(im5$x,im5$y,lty=1,lwd=2,col=3)
```

kasus 2:

```
#######here     is     an     example
simulation: (uncomment and run)
ff1 _ function(x)
{
  d  <-  sin(0.5  *  pi  *  x)/(  1  +
(sign(x) + 1)*(2*x^2))
 d
}
##### generate the x-y-w
n _ 100
x _ rnorm(n,0.5,0.25)
y _ ff1(x) + rnorm(n,0,.015)
repeats _ 2
w _ matrix(0,nrow=n,ncol=repeats)
for (j in 1:repeats){
 w[,j]          _          x         +
rnorm(n,0,(sqrt(3/7)*0.25) )}
wbar _ apply(w,1,mean)
xp                                    _
seq(min(wbar),max(wbar),length=500)
yp _ ff1(xp)

###### neat pictures (prints (x,y)
points, (wbar,y) , true curve,
###### naive, and icm spline
plot(x,y,pch=5,xlim=c(min(wbar),max(
wbar)),ylim=c(min(y),max(y)),
 xlab="",ylab="")
points(wbar,y,pch=18,col=2)
lines(xp,yp,lty=1,lwd=2)
ddd _ smooth.spline(wbar,y)
lines(ddd$x,ddd$y,lty=1,col=2,lwd=2)
im5 _ icm(w,y,iter=5)
lines(im5$x,im5$y,lty=1,lwd=2,col=3)
```

Kasus 3:

```
#######here     is     an     example
simulation: (uncomment and run)
ff1 _ function(x)
{
 d <- 10*sin(4*pi*x)
 d
}
##### generate the x-y-w
n _ 100
x _ rnorm(n,0.5,0.25)
y _ ff1(x) + rnorm(n,0,.015)
repeats _ 2
w _ matrix(0,nrow=n,ncol=repeats)
for (j in 1:repeats){
 w[,j]          _          x         +
rnorm(n,0,(sqrt(3/7)*0.25) )}
wbar _ apply(w,1,mean)
xp                                    _
seq(min(wbar),max(wbar),length=500)
yp _ ff1(xp)

###### neat pictures (prints (x,y)
points, (wbar,y) , true curve,
###### naive, and icm spline
plot(x,y,pch=5,xlim=c(min(wbar),max(
wbar)),ylim=c(min(y),max(y)),
 xlab="",ylab="")
points(wbar,y,pch=18,col=2)
lines(xp,yp,lty=1,lwd=2)
ddd _ smooth.spline(wbar,y)
lines(ddd$x,ddd$y,lty=1,col=2,lwd=2)
im5 _ icm(w,y,iter=5)
lines(im5$x,im5$y,lty=1,lwd=2,col=3)
```

Kasus 4:

```
#######here     is     an     example
simulation: (uncomment and run)
ff1 _ function(x)
{
 d <- 10*sin(4*pi*x)
 d
}

##### generate the x-y-w
n _ 500
x _ rnorm(n,0.5,0.25)
y _ ff1(x) + rnorm(n,0,.015)
repeats _ 2
w _ matrix(0,nrow=n,ncol=repeats)
for (j in 1:repeats){
 w[,j]          _          x         +
rnorm(n,0,(sqrt(3/7)*0.25) )}
wbar _ apply(w,1,mean)
xp                                    _
seq(min(wbar),max(wbar),length=500)
```

```
yp _ ff1(xp)

###### neat pictures (prints (x,y)
points, (wbar,y) , true curve,
###### naive, and icm spline
plot(x,y,pch=5,xlim=c(min(wbar),max(
wbar)),ylim=c(min(y),max(y)),
 xlab="",ylab="")
points(wbar,y,pch=18,col=2)
lines(xp,yp,lty=1,lwd=2)
ddd _ smooth.spline(wbar,y)
lines(ddd$x,ddd$y,lty=1,col=2,lwd=2)
im5 _ icm(w,y,iter=5)
lines(im5$x,im5$y,lty=1,lwd=2,col=3)
```

Kasus 5:

```
#######here     is     an     example
simulation: (uncomment and run)
ff1 _ function(x)
{
 d <- x^4
 d
}

##### generate the x-y-w
n _ 100
x _ rnorm(n,0,1)
y _ ff1(x) + rnorm(n,0,.1)
repeats _ 2
w _ matrix(0,nrow=n,ncol=repeats)
for (j in 1:repeats){
 w[,j] _ x + rnorm(n,0,.01) }
wbar _ apply(w,1,mean)
xp
seq(min(wbar),max(wbar),length=500)
yp _ ff1(xp)

###### neat pictures (prints (x,y)
points, (wbar,y) , true curve,
###### naive, and icm spline
plot(x,y,pch=5,xlim=c(min(wbar),max(
wbar)),ylim=c(min(y),max(y)),
 xlab="",ylab="")
points(wbar,y,pch=18,col=2)
lines(xp,yp,lty=1,lwd=2)
ddd _ smooth.spline(wbar,y)
lines(ddd$x,ddd$y,lty=1,col=2,lwd=2)
im5 _ icm(w,y,iter=5)
lines(im5$x,im5$y,lty=1,lwd=2,col=3)
```

Kasus 6:

```
#######here     is     an     example
simulation: (uncomment and run)
ff1 _ function(x)
{
 d <- x^4
 d
}

##### generate the x-y-w
n _ 500
x _ rnorm(n,0,1)
y _ ff1(x) + rnorm(n,0,.1)
repeats _ 2
w _ matrix(0,nrow=n,ncol=repeats)
for (j in 1:repeats){
 w[,j] _ x + rnorm(n,0,1) }
wbar _ apply(w,1,mean)
xp
seq(min(wbar),max(wbar),length=500)
yp _ ff1(xp)

###### neat pictures (prints (x,y)
points, (wbar,y) , true curve,
###### naive, and icm spline
plot(x,y,pch=5,xlim=c(min(wbar),max(
wbar)),ylim=c(min(y),max(y)),
 xlab="",ylab="")
points(wbar,y,pch=18,col=2)
lines(xp,yp,lty=1,lwd=2)
ddd _ smooth.spline(wbar,y)
lines(ddd$x,ddd$y,lty=1,col=2,lwd=2)
im5 _ icm(w,y,iter=5)
lines(im5$x,im5$y,lty=1,lwd=2,col=3)
```

7. SIMULASI

Model : Fungsi Eksponen

$$g(x) = exp(x),$$

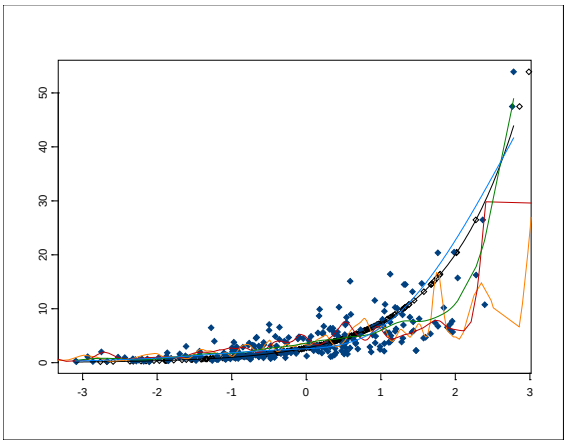
Dengan n=250, $\sigma_{\varepsilon}^2 = 0.01^2$,

$\sigma_{\mathbf{u}}^2 = 1^2$, $\mu_{\mathbf{x}} = 0$, dan $\sigma_{\mathrm{x}}^2 = 1$.

Dan dari hasil output S-plus, didapatkan kurva regresi spline dan nilai MSE dari simulasi di atas seperti terlihat dalam Tabel 1 dan gambar 1.

Tabel 1. Nilai MSE Fungsi Eksponens

No	*Metode*	*Nilai MSE*
1	Non ME[1]	19.90525
2	Non ME[2]	25.0184
3	Naïve	12.80861
4	ICM	4.231545



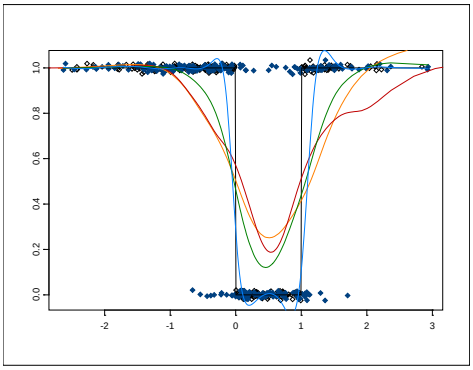
Gambar 1. Estimasi Kurva untuk Fungsi Eksponens.

ISSN: 1907-5022

Model 2: Fungsi *Truncated*: $g(x) = x_+ + (1-x)_+$.

dengan n=250, $\sigma_\varepsilon^2 = 0.01^2$, $\sigma_U^2 = 0.5^2$, $\mu_x = 0$, dan $\sigma_x^2 = 1$.

Dan dari hasil output S-plus, didapatkan kurva regresi spline dan nilai MSE seperti dalam Tabel 2 dan gambar 2.



Gambar 2. Estimasi Kurva untuk Fungsi *Truncated*.

Tabel 2. Nilai MSE Fungsi *Truncated*.

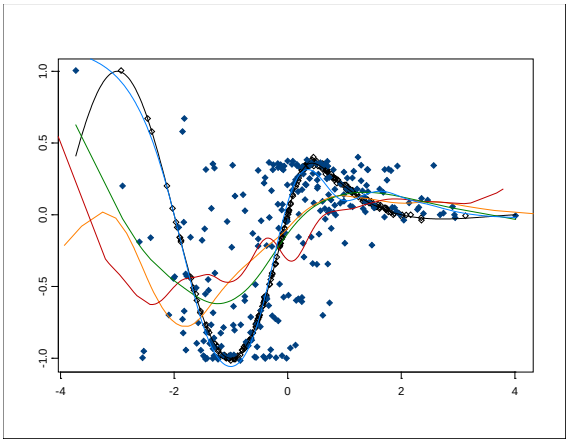
No	*Metode*	*Nilai MSE*
1	Non ME[1]	0.157338
2	Non ME[2]	0.139422
3	Naïve	0.102883
4	ICM	0.00143

Model 3 : Fungsi Trigonometri

$$g(x) = \frac{sin(\pi x/2)}{1+2x^2(sin\, g(x)+1)},$$

n=250, $\sigma_\varepsilon^2 = 0.01^2$, $\sigma_u^2 = 0.1^2$, $\mu_x = 0$, dan $\sigma_x^2 = 1$.

Dan dari hasil output S-plus, didapatkan kurva regresi spline dari simulasi di atas adalah sebagai berikut:



Gambar 3 Estimasi Kurva untuk Fungsi Trigonometri

Tabel 3 Nilai MSE untuk Fungsi Trigonometri

No	*Metode*	*Nilai MSE*
1	Non ME[1]	0.158436
2	Non ME[2]	0.195413
3	Naïve	0.147683
4	ICM	0.030937

Gambar 1, 2, dan 3 di atas menunjukkan kurva regresi dari 3 metode, yaitu kurva regresi dari data dengan mengabaikan adanya kesalahan pengukuran, dengan *naïve Method*, dan juga dengan ICM. Dilihat dari ketiga gambar , pengabaian kesalahan pengukuran, berpengaruh terhadap kurva regresi. Terlihat bahwa kurva regresi bila kesalahan pengukuran diabaikan jauh dari kurva sebenarnya. Nampak juga dari Gambar 1, 2, dan 3, kurva regresi dengan ICM yang paling mendekati dengan kurva sebenarnya,g(x). Selain itu, dengan *Naïve Method*, juga memberikan estimasi kurva regresi yang lebih mendekati kurva g(x) dibandingkan dengan kurva bila kesalahan pengukuran diabaikan.

Hal ini membuktikan bahwa adanya kesalahan pengukuran dalam variabel prediktor, X, berpengaruh terhadap kurva regresi.

Selanjutnya dilakukan simulasi untuk masing-masing model 1, model 2, model 3, dengan variasi nilai n = 25, 50, 100, 250, $\sigma_\varepsilon^2 = 0.01^2, 0.1^2, 1^2$, , $\sigma_U^2 = 0.01^2, 0.1^2, 0.5^2$ dan 1^2, $\mu_x = 0$, dan $\sigma_x^2 = 1$.

8. KESIMPULAN

Berdasarkan pembahasan di atas, maka dapat disimpulkan bahwa:

1. Adanya kesalahan pengukuran di dalam analisis regresi, khususnya linear sederhana, menyebabkan estimator koefisien regresi tidak memenuhi sifat unbias, yaitu

   $$\mathbf{E}(b_1) \neq b_1.$$

   Dan secara umum didapatkan bentuk estimator yang telah dikoreksi untuk $\boldsymbol{\beta}$, yaitu

   $$\hat{\beta} = H^{-1}\left(n^{-1} W^T Y\right)$$

   merupakan estimator yang secara asimtotik berdistribusi Normal,

   $$n^{1/2}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \Sigma_{XX}^{-1} \Gamma \left(\Sigma_{XX}^{-1}\right)\right).$$

2. Estimasi kurva regresi dengan metode ICM yaitu estimasi g yang ditentukan secara *iterative* dan memberikan nilai maksimum pada densitas posterior bersama sebagai berikut

$$p(\theta|Y,W) \propto \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(Y_i - g(X_i))^2 - \frac{1}{2\sigma_W^2}(W_i - X_i)^2 - \frac{1}{2\sigma_X^2}(X_i - \mu_X)^2 + -\frac{\hat{\alpha}}{2\sigma_\varepsilon^2} g^T K g \right]$$

3. Berdasarkan hasil simulasi, fungsi Eksponens, *Truncated* dan Trigonometri, didapatkan nilai MSE dari masing-masing metode. Hasil simulasi menunjukkan bahwa:
   a. Semakin besar σ_U^2, menunjukkan bahwa perbedaan MSE antara non ME dan ICM semakin besar. Berdasarkan dari hasil simulasi, MSE non ME (pengabaian kesalahan pengukuran} bisa mencapai 4 kali lebih besar dari model bila ada kesalahan pengukuran diestimasi dengan ICM, baik utuk *n*=25, 50,100,dan 250. Sedangkan untuk *Naïve Method* juga menunjukkan nilai MSE yang lebih baik dibandingkan bila adanya kesalahan pengukuran diabaikan.
   b. Berdasarkan hasil simulasi juga menunjukkan bahwa metode ICM memberikan nilai yang lebih baik dibandingkan dengan *Naïve Method*.

DAFTAR PUSTAKA

Amemiya, Y and Fuller, W. A (1984), "Estimation for Multivariate Errors in Variable Model with Estimated Error Covariance Matrix, " , *Annals of Statistics*, *Annals of Statistics*, 12, 497-509.

Berry, S. A., Carroll, R. J. and Ruppe (2002), "Bayesian smoothing regression splines for measurement problems", *Journal of the Am Statistical Association*, 9, 160–169.

Besag, J. (1986). " On the Satistical Analysıs oı Dirty Pictures" (with Discussion), *Journal of the Royal Statistical Society*, Series B, 48, 259-279.

Box, G. E. P. and Tiao, G. (1973), *Bayesian Inference in Statistical Analysis*, Addison–Wesley, London.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999), "Nonparametric regression with errors in covariates", *Biometrika*, 86, 541–554.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, Chapman and Hall, New York.

Cook, J. R. and Stefanski, L. A. (1994), "Simulation–extrapolation estimation in parametric measurement error models", *Journal of the American Statistical Association*, **89**, 1314–1328.

Fan, J. and Truong, Y. K. (1993), "Nonparametric regression with errors in variables", *Annals of Statistics*, 21, 1900–25.

Fuller, W. A and Hidiroglou, M. A (1976), " Regression Estimation After Correcting for Attenuation", *Journal of the American Statistical Association*, **73**, 99-169.

Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Hardle,W. (1990), *Applied Nonparametrics Regression*, Australia,: Cambridge Univercity Press.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall: New York.

Ruppet, D. (2002), "Selecting the Number of Knots for Penalized Splines", *Journal of Computation and graphical Statistics*, **11**, 735-757.

Wahba, G. (1978), "Bayesian Confidence Interval" for Cross-Validated Smoothing Spline", *Journal of Royal Statistical Society*, Ser. B. 45. 133-150.