PEMILIHAN MODEL REGRESI LINIER BERGANDA DENGAN KRITERIA AIC

Ira Ernawati
Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Halu Oleo, Kendari, Indonesia,
Email:ernawatiira883@gmail.com

Wayan Somayasa[1], Arman[2] dan Alfian[3]
Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Halu Oleo, Kendari, Indonesia,
Email:[1]wayan.somayasa@uho.ac.id [2]arman.mtmk@uho.ac.id, [3]Alfian@uho.ac.id

ABSTRAK

Tujuan dari penelitian ini adalah untuk mengetahui pemilihan model terbaik dalam regresi linier berganda dengan kriteria AIC dan pengaplikasiannya pada data real. AIC (*Akaike Information Criterion*) merupakan pengukuran untuk kualitas relatif pada model statistik dari data yang diberikan untuk pemilihan model terbaikdari beberapa model yang ada. Menurut metode AIC, model regresi terbaik adalah model regresi yang mempunyai nilai AIC terkecil. Jenis penelitian yang digunakan adalah penelitian terapan, yaitu penelitian yang memberikan solusi atas permasalahan tertentu secara praktis. Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari siswa setelah melakukan kegiatan pembelajaran selama tiga tahun dijenjang SMKN 1 Kulisusu Barat. Data yang digunakan terdiri dari variabel respon yang merupakan rata-rata nilai UN, dan variabel prediktor yang terdiri dari rata-rata nilai tryout, nilai ujian kompetensi, dan rata-rata nilai ujian sekolah. Penerapan kriteria AIC pada studi kasus faktor-faktor yang mempengaruhi rata-rata nilai UN di SMKN 1 Kulisusu Barat, menghasilkan model $\hat{Y}_i = 0{,}72 + 0{,}41X_{1i} + 0{,}29X_{2i} + 0{,}21X_{3i}; i = 1, \ldots, 35$ sebagai model regresi terbaik dengan kriteria AIC terendah, yaitu sebesar $-266{,}79$.

Kata kunci: Kriteria AIC, model regresi linear berganda, variable respons, variable predictor, model terbaik

ABSTRACT

The purpose of this research is to determine the selection of the best model in multiple linear regression with AIC criteria and its application to real data. AIC (Akaike Information Criterion) is a measurement for the relative quality of a statistical model from the data provided for the selection of the best model from several existing models. According to the AIC method, the best regression model is the regression model that has the smallest AIC value. The type of research used is applied research, which is research that provides solutions to certain problems in a practical way. The data used in this study are secondary data obtained from students after carrying out learning activities for three years at the SMKN 1 Kulisusu Barat. The data used consists of a response variable which is the average of the UN values, and a predictor variable which consists of an average tryout score, competency test scores, and average school test scores. The application of the AIC criteria to the case study of factors affecting the average UN score at SMKN 1 Kulisusu Barat, resulted in the $\hat{Y}_i = 0{,}72 + 0{,}41X_{1i} + 0{,}29X_{2i} + 0{,}21X_{3i}; i = 1, \ldots, 35$ as the best regression model with the lowest AIC criteria, which was -266.79.

Keywords : ***AIC criterion, multiple linear regression, response variable, predictor variable, the best model.***

1. Pendahuluan

Dalam analisis regresi, diperlukan suatu model yang digunakan untuk mengetahui hubungan antara variabel tak prediktor (respon) dengan satu atau lebih variabel prediktor (prediktor). Model regresi dapat diperoleh dengan melakukan estimasi terhadap parameter-parameternya menggunakan metode tertentu. Adapun metode yang digunakan untuk mengestimasi parameter model regresi, khususn ya parameter model regresi linier berganda adalah dengan metode kuadrat terkecil (*ordinari least square*) dan metode kemungkinan maksimum (*maximum likelihood*) (Kuetner, dkk, 2004).

Selain metode kuadrat terkecil dan metode maksimum likelihood dalam pemilihan model, analisis regresi juga bisa menggunakan kriteria AIC untuk mendapatkan model terbaik.

AIC (*Akaike Information Criterion*) merupakan pengukuran untuk kualitas relatif pada model statistik dari data yang diberikan untuk pemilihan model terbaik dari beberapa model yang ada. Jika $f(y)$ adalah densitas dari distribusi sebenarnya dan $g(y)$ adalah yang ditentukan oleh model, ukuran perbedaan yang terkenal adalah perbedaan *Kullback-Leibler*.

$$KL(f, g) = \int \ln \frac{g(X)}{f(X)} f(y)\, dy \qquad (1)$$

Perhatikan bahwa ini bukan jarak sebenarnya, karena $KL(f, g) \neq KL(g, f)$. Sebaliknya, ini adalah ukuran perbedaan antara f tetap yang sebenarnya dan berbagai model yang bersaing g.

Model yang mungkin digunakan $g(y\,;\,\boldsymbol{\theta})$ di mana $\boldsymbol{\theta}$ adalah ruang parameter dan menggunakan notasi $KL(f, g\,;\,\boldsymbol{\theta})$ untuk mencerminkan ini. Model f mayor yang sebenarnya mungkin tidak berbentuk $g(y\,;\,\boldsymbol{\theta})$ untuk beberapa $\boldsymbol{\theta}$ tertentu.

Model yang sesuai dengan nilai $\boldsymbol{\theta}$ yang meminimalkan $KL(f, g\,;\,\boldsymbol{\theta})$ atau secara ekuivalen yang meminimalkan

$$-\int \ln g(y\,;\,\boldsymbol{\theta}) f(y)\, dy = -E[\ln g(\boldsymbol{Y}\,;\,\boldsymbol{\theta})] \qquad (2)$$

Pada Persamaan (1.2) terdapat dua yang tidak diketahui yaitu parameter $\boldsymbol{\theta}$ dan model benar yang tidak diketahui pada f.

Untuk mengestimasi $\boldsymbol{\theta}$, misalkan sampel $\boldsymbol{Y}$ dan mengestimasi $\boldsymbol{\theta}$ menggunakan estimasi maksimum *likelihood* $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ yang memaksimalkan $g(\boldsymbol{Y}\,;\,\boldsymbol{\theta})$ sebagai fungsi dari $\boldsymbol{\theta}$. Menjadi

$$\Delta = -\int \ln g\big(x\,;\,\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\big) f(x) dx \qquad (3)$$

Dengan nilai yang diterapkan

$$E[\Delta] = \iint \ln g\left(x\,;\,\widehat{\boldsymbol{\theta}}(y)\right) f(x)\, f(y) dx\, dy$$

Kriteria Δ, mengukur perbedaan antara model f benar yang tidak diketahui dan model yang paling sesuai dengan bentuk $g(y\,;\,\boldsymbol{\theta})$. Untuk mendapatkan kriteria operasional, dibutuhkan perkiraan Δ. Perkiraan standar Δ (sebenarnya diperkirakan 2Δ) adalah *kriteria informasi Akaike* (AIC) (Akaike, 1973).

$$AIC = -2 \ln g\big(\boldsymbol{Y}\,;\,\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\big) + 2r \qquad (4)$$

di mana :

$r =$ dimensi dari vektor parameter $\boldsymbol{\theta}$.

Menurut metode AIC, model regresi terbaik adalah model regresi yang mempunyai nilai AIC

Pada bagian 2 dari paper ini akan dibahas tentang metode penelitian. Bagian 3 mendiskusikan tentang pemilihan model terbaik. Selanjutnya pada bagian 4 dibahas tentang aplikasi kriteria AIC pada data reel dan bagian 5 tentang kesimpulan.

2. Metode

Jenis penelitian yang digunakan adalah penelitian. Dalam penelitian ini akan dilakukan beberapa pemodelan regresi linier berganda dengan metode kuadrat terkecil dan akan dilakukan seleksi umum dengan menghitung nilai AIC dari masing-masing model regresi.

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari siswa setelah melakukan kegiatan pembelajaran selama tiga tahun dijenjang SMKN 1 KULISUSU BARAT yang diselenggarakan secara nasional yang meliputi tiga mata pelajaran, yaitu Matematika, Bahasa Inggris dan Bahasa Indonesia

Adapun prosedur yang dijalankan dalam mencapai tujuan dalam penelitian ini adalah sebagai berikut :

1) Menurunkan prosedur kriteria AIC pada regresi linier berganda
2) Melakukan pemodelan regresi linier berganda dengan 3 variabel prediktor.
3) Melakukan uji kenormalan residual yang diperoleh dari model regresi linier berganda
4) Melakukan uji signifikansi parameter regresi linier berganda, untuk melihat variabel prediktor yang berpengaruh signifikan terhadap variabel respon
5) Melakukan pemilihan model terbaik dengan kriteria AIC
6) Mendapatkan model regresi terbaik.
7) Menghitung koefisien determinasi (R^2) dari model yang didapatkan
8) Menarik kesimpulan

3. Pemilihan Model Terbaik

3.1 *Akaike's Information Criterion* (AIC)

Akaike's Information Criterion (AIC) dapat diterapkan pada situasi pemodelan apa pun. Dalam kasus model linier, model sebenarnya $f(y)$ dari Y dapat diambil sebagai normal multivariat dengan vektor rata-rata μ, dan varian matriks $\sigma_0^2 \boldsymbol{I}_n$. Model kandidat tipikal $g(y\,;\,\boldsymbol{\theta})$ juga normal multivariat, dengan vektor rata-rata $\boldsymbol{X}\beta$ varian matriks $\sigma^2 \boldsymbol{I}_n$. Matrix regresi $\boldsymbol{X}$diasumsikan $n\ x\ p$ dari rank p, dan vektor parameter dalam hal ini adalah $\boldsymbol{\theta} = (\beta', \sigma^2)'$.

$$-\ln\ g(y\,;\,\boldsymbol{\theta}) = \frac{n}{2} \ln\,(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - \boldsymbol{X}\beta)'(y - \boldsymbol{X}\beta) \qquad (5)$$

Dalam kasus model linier, distribusi sebenarnya f berbentuk $g(y\,,\boldsymbol{\theta})$ untuk beberapa $\boldsymbol{\theta}$ jika dan hanya jika μ dalam bentuk $C(\boldsymbol{X})$, yang ekivalen dengan $\lambda = 0$. Ini berlaku, $n\hat{\sigma}^2/\sigma_0^2 \sim \chi^2_{n-p}$, sehingga menggunakan

$$E\left[\frac{1}{X}\right] = \frac{1}{v-2} \qquad (6)$$

Untuk χ_v^2 variabel acak X, didapatkan

$$E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] = \frac{n}{n-p-2} \qquad (7)$$

dan

$$E[2\Delta] = nE[\ln\,(2\hat{\sigma}^2)] + \frac{n(n+p)}{n-p-2} \qquad (8)$$

Baik 2Δ dan AIC memiliki ekspektasi yang sebanding dengan $O(n)$. Bias dalam penduga $E[2\Delta]$ oleh AIC adalah

$$\begin{aligned} E[AIC - 2\Delta] &= n + 2(p+1) - \frac{n(n+p)}{n-p-2} \\ &= 2(p+1) - \frac{2n(p+1)}{n-p-2} \end{aligned} \qquad (9)$$

Meskipun biasnya berorde lebih kecil dari $E[2\Delta]$, diperoleh penduga $E[2\Delta]$ yang tepat tak bias dengan menggunakan kriteria yang dimodifikasi

$$AIC_c = n \ln\ 2\pi\hat{\sigma}^2 + \frac{n(n+p)}{n-p-2} \quad (10)$$

Hurvich dan Tsai (1991) memberikan beberapa contoh di mana $\mathrm{AIC_c}$ merupakan penduga $E[2\Delta]$ yang jauh lebih baik dari pada AIC.

Jika μ tidak ada di $C(\boldsymbol{X})$, maka dari Persamaan (12.17) dalam buku Seber dan Lee (2007), $nE[\hat{\sigma}^2] = \sigma_0^2(\lambda + n - p)$ dan

$$E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] \approx \frac{\sigma_0^2}{E[\hat{\sigma}^2]} = \frac{n}{\lambda + n - p} \quad (11)$$

Mengarah ke

$$E[AIC - 2\Delta] \approx n + 2(p+1) - \frac{(n+p+\lambda)n}{\lambda + n - p}$$
$$= 2\left(p + 1 - \frac{np}{\lambda + n - p}\right) \quad (12)$$

Bentuk AIC yang sedikit berbeda muncul jika berasumsi bahwa σ^2 diketahui. Jika ini masalahnya, hanya ada p parameter dan

$$-2\ln\ g\left(y, \hat{\theta}(y)\right)$$
$$= n \ln(2\pi\sigma^2) + \frac{\left(y - \boldsymbol{X}\hat{\beta}\right)'\left(y - \boldsymbol{X}\hat{\beta}\right)}{\sigma^2} \quad (13)$$

Sehingga konstanta $n \ln(2\pi\sigma^2)$ tidak tergantung pada model

$$AIC = \frac{RSS_p}{\sigma^2} + 2p \quad (14)$$

Jadi, dalam hal ini AIC sangat mirip dengan C_p. jika σ^2 diganti dengan perkiraan $\hat{\sigma}^2$, maka berbeda dengan konstanta n.

Generalisasi yang jelas dari versi AIC ini adalah dengan mempertimbangkan kriteria

$$\frac{RSS_p}{\sigma^2} + a_n p \quad (15)$$

di mana a_n diperoleh bergantung pada n. Pilihan $a_n = \ln n$ mengarah ke kriteria yang dikenal sebagai *Bayesian Information Criteria* (BIC).

3.2 Metode Stepwise Seleksi Maju

Misalkan pada Persamaan (12.3) dalam buku Seber dan Lee (2007), diketahui model regresi linier berganda

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \quad (16)$$

di mana $\boldsymbol{X}$ berordo $n\,x\,(K+1)$, dengan K menyatakan banyak variabel. Untuk mengidentifikasi variabel yang signifikan dan memiliki koefisien regresi tidak sama dengan nol. Misalkan K variabel tersebut dibagi menjadi dua himpunan. Himpunan pertama terdiri dari variabel $p-1$ yang dianggap penting. Sedangkan himpunan kedua terdiri dari sisanya, yaitu variabel $K-p+1$, dan diduga memiliki koefisien regresi sama dengan nol. Untuk mengetahui himpunan kedua tidak mengandung variabel signifikan, kita dapat mengujinya dengan menggunakan uji F statistik :

$$F = \frac{RSS_p - RSS_{K+1}}{RSS_{K+1}} \frac{n-K-1}{K-p+1} \quad (17)$$

Pengujian ini merupakan pembeda antara dua model, masing-masing memiliki variabel K dan $p-1$. Jika $K = p$ maka

$$F = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}}(n-p-1) \quad (18)$$

Misalkan dalam sebuah model memiliki dua variabel penjelas yang mungkin yaitu x_1 dan x_2. Sehingga hanya ada empat model yang mungkin, yaitu model $\{0\}$ yang hanya terdiri dari suku konstan saja, model $\{x_1\}$, model $\{x_2\}$, dan model $\{x_1, x_2\}$. Model regresi terstandarkan dengan dua variabel penjelas ini adalah:

$$Y_i = \alpha_0 + \gamma_1 x_{i1}^* + \gamma_2 x_{i2}^* + \varepsilon_i \quad (19)$$

Untuk penyederhanaan, asumsikan bahwa varian σ^2 dari ε_i. Diketahui $\sigma^2 = 1$.

Taksiran kuadrat terkecil dari γ_1 adalah :

$$\hat{\gamma}_1 = \frac{r_1 - rr_2}{1 - r^2} \quad (20)$$

di mana :

$$r_1 = \sum_i x_{i1}^* Y_i$$
$$r_2 = \sum_i x_{i2}^* Y_i$$
$$r = \sum_i x_{i1}^* x_{i2}^*$$

Misalkan criteria AIC digunakan sebagai kriteria pemilihan model. Karena $\sigma^2 = 1$. Maka Persamaan (4.16) dapat ditulis sebagaiberikut :

$$AIC = RSS_P + 2p \quad (21)$$

RSSuntuk model tiga parameter penuh adalah :

$$RSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \boldsymbol{Y}'\boldsymbol{X}^*\hat{\boldsymbol{\gamma}}$$
$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{1}{1-r^2}(r_1, r_2)\begin{pmatrix}1 & -r\\ -r & 1\end{pmatrix}\begin{pmatrix}r_1\\ r_2\end{pmatrix}$$
$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{r_1^2 - 2rr_1r_2 + r_2^2}{1 - r^2} \quad (22)$$

Jadi , penulisan $SSY = \sum_{i=1}^n (Y_i - \bar{Y})^2$, AIC untuk model ini adalah

$$AIC = SSY - \frac{r_1^2 - 2rr_1r_2 + r_2^2}{1-r^2} + 6$$
$$= SSY - r^* + 6 \quad (23)$$

di mana $r_* = (r_1^2 - 2rr_1r_2 + r_2^2)/(1 - r^2)$.

Demikian pula, AIC_suntuk model lainnya adalah

$AIC = SSY - r_1^2 + 4$, untuk model $\{x_1\}$ (24)

$AIC = SSY - r_2^2 + 4$, untuk model $\{x_2\}$, (25)

$AIC = SSY + 2$, untuk model $\{x_0\}$, (26)

Pemilihan model dengankriteria AIC terkecil setara dengan algoritma berikut :

1. Memilih $\{x_1, x_2\}$ jika $r_* \geq 4$, $|rr_1 - r_2| \geq \sqrt{2(1-r^2)}$ dan $|rr_2 - r_1| \geq \sqrt{2(1-r^2)}$

2. Memilih $\{x_1\}$ jika $r_1^2 \geq 2$, $r_1^2 \geq r_2^2$ dan $|rr_1 - r_2| < \sqrt{2(1-r^2)}$
3. Memilih $\{x_2\}$ jika $r_2^2 \geq 2$, $r_2^2 \geq r_1^2$dan $|rr_2 - r_1| < \sqrt{2(1-r^2)}$
4. Memilih $\{x_0\}$ jika $r_1^2 < 2, r_2^2 < 2$ dan $r_* < 4$.

3.1.1 Penskalaan Variabel penjelas

Misalakan kita juga menskalakan kolom dari $\boldsymbol{X}$ sehingga mereka memiliki panjang satuan. Jika $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2$ dan $X_{ij}^* = \left(X_{ij} - \bar{X}_j\right)/s_j$.
Maka model persamaan regresi menjadi :

$$Y_i = \alpha_0 + \gamma_1 X_{i1}^* + \gamma_2 X_{i2}^* + \cdots + \gamma_p X_{ip}^* + \varepsilon_i, i = 1, 2, \ldots, n \qquad (27)$$

di mana :
$\gamma_j = \beta_j s_j, j = 1,2,\ldots,p$.
Karena transformasi masih satu-ke-satu, $\hat{\gamma}_j = \hat{\beta}_j s_j$ dan $\hat{\alpha}_0 = \bar{Y}$. Jika $\boldsymbol{X}^* = X_{ij}^*$ dan $\boldsymbol{\gamma} = [\gamma_1 \gamma_2 \ \ldots \ \gamma_p]^T$, maka penduga bagi $\boldsymbol{\gamma}$ dapat dinyatakan dalam bentuk matriks yaitu sebagai berikut :

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y} = \boldsymbol{R}_{xx}^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y} \qquad (28)$$

di mana :
$\boldsymbol{R}_{xx}$sekarang adalah matriks korelasi

$$\boldsymbol{R}_{xx} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} \qquad (29)$$

dan

$$r_{jk} = \sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)/s_j s_k \qquad (30)$$
$$j, k = 1,2, \ldots, p$$

adalah koefisien korelasi sampel untuk variabel penjelas ke-j dan ke-k. Jika kita menuliskan notaasi baru $\boldsymbol{X}^* = \left(\boldsymbol{X}^{*(1)}, \ldots, \boldsymbol{X}^{*(2)}\right)$ untuk kolom $\boldsymbol{X}^*$, diketahui bahwa $r_{jk} = \boldsymbol{X}^{*(j)T}\boldsymbol{X}^{*(k)}$.

4. Aplikasi dengan Kriteria AIC

Dalam penelitian ini, digunakan data berupa sampel mengenai kegiatan pembelajaran selama tiga tahun di SMKN 1 Kulisusu Barat, Kabupaten Buton. Variabel terikat (Y) adalah nilai ujian nasional, dengan tiga variabel penjelas, yaitu X_1 (rata-rata nilai ujian nasional), X_2 (nilai ujian kompetensi), dan X_3 (rata-rata nilai ujian sekolah). Misalkan model yang cocok adalah :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, \qquad (31)$$

$i = 1,2, \ldots, 35$
maka didapatkan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 8{,}59 & 8{,}18 & 8{,}04 \\ 1 & 8{,}60 & 8{,}65 & 8{,}53 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 7{,}82 & 7{,}73 & 7{,}58 \end{bmatrix}; \quad \boldsymbol{Y} = \begin{bmatrix} 8{,}17 \\ 8{,}40 \\ \vdots \\ 7{,}38 \end{bmatrix}$$

Selanjutnya dengan menggunakan Persamaan (1.4), maka diperoleh penduga bagi $\boldsymbol{\beta} = [\beta_0 \beta_1 \beta_2 \beta_3]^T$adalah sebagai berikut :

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0{,}72 \\ 0{,}41 \\ 0{,}29 \\ 0{,}21 \end{bmatrix}$$

Selanjutnya jika model lengkap (32) ditransformasi kedalam model regresi terpusat dan berskala, yaitu:

$$Y_i = \alpha_0 + \gamma_1 X_{i1}^* + \gamma_2 X_{i2}^* + \cdots + \gamma_k X_{ik}^* + \varepsilon_i$$

dengan

$$X_{ij}^* = \frac{\left(X_{ij} - \bar{X}_j\right)}{s_j}$$

$j = 1,2,3$
$i = 1, 2, \ldots, 35$
Penduga bagi $\boldsymbol{\gamma} = [\gamma_1 \gamma_2 \ \ldots \ \gamma_p]^T$ dihitung dengan menggunakan Persamaan (4.32) yaitu :

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y} = \boldsymbol{R}_{xx}^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}$$

$$= \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}^{-1} \begin{bmatrix} X_{1\,1}^* & \cdots & X_{1\,35}^* \\ X_{2\,1}^* & \cdots & X_{2\,35}^* \\ X_{3\,1}^* & \cdots & X_{3\,35}^* \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_{35} \end{bmatrix}$$

$\boldsymbol{R}_{xx}$ adalah matrix korelasi yang elemen-elemenya dihitung dengan menggunakan Persamaan (4.34) sehingga

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y} = \boldsymbol{R}_{xx}^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}$$
$$= \begin{bmatrix} 26{,}85 \\ -7{,}67 \\ -8{,}21 \end{bmatrix}$$

Kita dapat merumuskan nilai AIC untuk model lengkap dengan menggunakan Persamaan (4.25) berikut:

$$AIC = RSS_P + 2p$$

dengan

$$RSS_P = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \boldsymbol{Y}^T\boldsymbol{X}^*\hat{\boldsymbol{\gamma}}$$

diketahui

$$\bar{Y} = 8{,}07$$

sehingga

$$RSS_P = 11{,}80 - 286.59 = -274{,}79$$

Sehingga nilai AIC untuk model lengkap adalah :

$$AIC = -274{,}79 + 2(4) = -266{,}79$$

Selanjutnya kita dapat menggunakan Persamaan (4.27), (4.28), (4.29) dan (4.30) untuk menentukan nilai AIC pada pilihan model yang lainnya.

Untuk keseluruhan model yang mungkin, nilai AIC yang didapatkan adalah

$$AIC_{\{x_1,x_2,x_3\}} = -274{,}79 + 2(4) = -266{,}79,$$
$$AIC_{\{x_1,x_2\}} = 11{,}80 - 252{,}02 + 6 = -234{,}22,$$
$$AIC_{\{x_1,x_3\}} = 11{,}80 - 267 + 6 = -249{,}2\ ,$$
$$AIC_{\{x_2,x_3\}} = 11{,}80 - 81{,}46 + 6 = -63{,}66,$$
$$AIC_{\{x_1\}} = 11{,}80 - 14{,}95^2 + 4 = -207{,}70,$$
$$AIC_{\{x_2\}} = 11{,}80 - 0{,}89^2 + 4 = -65{,}56,$$
$$AIC_{\{x_3\}} = 11{,}80 - 1{,}06^2 + 4 = -14{,}89,$$
$$AIC_{\{x_0\}} = SSY + 2 = 11{,}80 + 2 = 13{,}8.$$

Model regresi terbaik adalah model yang memiliki nilai AIC terkecil. Berdasarkan hasil yang diperoleh model dengan AIC terkecil terdapat pada model lengkap yang memiliki 3 variabel penjelas. Sehingga model regresi terbaik berdasarkan kriteria AIC adalah

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, i = 1, 2, \cdots, 35$$

Model estimasinya adalah

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$

atau

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

dengan

$$\boldsymbol{Y} = \begin{bmatrix} 8.17 \\ 8.40 \\ \vdots \\ 7.38 \end{bmatrix};$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 8,59 & 8,18 & 8,04 \\ 1 & 8,60 & 8,65 & 8,53 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 7,82 & 7,73 & 7,58 \end{bmatrix}$$

dan

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 0,72 \\ 0,41 \\ 0,29 \\ 0,21 \end{bmatrix}$$

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa:

1. Prosedur kriteria AIC dalam model regresi linier berganda dapat dihitung dengan menggunakan persamaan:
$$AIC = \frac{RSS_p}{\sigma^2} + 2p$$
di mana p adalah banyak variabel prediktor (X), dan $RSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \frac{r_1^2 - 2rr_1r_2 + r_2^2}{1-r^2}$
2. Penerapan kriteria AIC pada kasus faktor-faktor yang mempengaruhi rata-rata nilai UN di SMKN 1 Kulisusu Barat, menghasilkan model $\hat{Y}_i = 0,72 + 0,41X_{1i} + 0,29X_{2i} + 0,21X_{3i}; i = 1, \dots, 35$ sebagai model regresi terbaik dengan kriteria AIC terendah, yaitu sebesar $-266,79$. Sehingga didapatkan faktor-faktor yang mempengaruhi nilai UN di SMKN 1 Kulisusu Barat yaitu, nilai tryout (X_1), nilai ujian kompetensi (X_2), dan rata-rata nilai ujian sekolah (X_3).

5.2 Saran

Penelitian ini menggunakan kriteria AIC dalam pemilihan model regresi linier berganda, untuk penelitian selanjutnya dapat menggunakan kriteria pemilihan yang lain yaitu *Bayesian Information Criterion (BIC)*, sebagai pembandin dalam pemilihan model terbaik pada regresi linier beganda.

Ucapan Terimaksih. Penulis mengucapkan terimakasih yang sebesar-besarnya atas segala curahan perhatian yang diberikan oleh dosen pembimbing sehingga penelitian ini dapat diselesaikan dengan baik. Ucapan terimakasih juga penulis sampaikan kepada segenap sivitas akademika FMIPA UHO atas segala keramahan pelayanan yang diberikan sehingga penulis dapat menyelesaikan penelitian ini tepat waktu.

Daftar Pustaka

D. Cahyawati, H. Tanuji dan R. Abdiati. (2009). Efektivitas Metode Regresi Robust Penduga Welsch Dalam Mengatasi Pencilan Pada Pemodelan Regresi Linear Berganda. *Penelitian Sains 1*, 1-7.

J. Cavanaugh. (2005). *The Schwarz Information Criterion (SIC).* The University of Lowa: Departement Biostatistics.

N. Draper, dan H. Smith. (1992). *Analisis Regresi Terapan* (Kedua ed.). Jakarta: Gramedia Pustaka Utama.

M. Faturahman. (2009). Pemilihan Model Regresi Terbaik Menggunakan Metode Akaike's Information Criterion. *Informatika Mulawarman, 14*, 37-41.

D. Gujarati. (2007). *Essentials of Econometrics.* (A. M. Julius, Penerj.) Jakarta: Erlangga.

E.T. Herdianti. (2007). Pemilihan Model Terbaik : Suatu Kajian Pustaka. *Matematika, Statistik & Komputasi, 14*, 43-53.

M. Kuetner, C. Nachtsheim dan J. Neter. (2004). *Applied Linear Regression Models.* New York : McGraw-Hill/ Irwin.

D. Kurniawan. (2008). Diambil kembali dari http://www.google.com.id/2008/regresi linear.html

W. Somayasa. (2019). *Statistika Matematika (Teori Peluang dan Dasar-dasar Inferensi Parameter*). Kendari : Universitas Halu Oleo.

Tarno. (2007). Metode Kuadrat Terkecil : Suatu Kajian Pustaka. *Sains dan Matematika, 5*, 69-72.

Diterima pada tanggal 2 Januari 2023.
Terbit online pada tanggal 29 April 2023

.