

Deteksi Pencilan pada Model Regresi Beta dalam Kasus Kemiskinan Perdesaan di Indonesia Tahun 2022

Neng Marlina*, Nusar Hajarisman

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*marlin.ylnt@gmail.com, nusarhajarisman@unisba.ac.id

Abstract. In a certain dataset, there may be one or a group of data points that are far from the rest of the data in a regression model, which should be suspected as outliers. These outliers have the potential to influence the significance of regression coefficients. Furthermore, the presence of outliers in the data can cause the regression model to fail to meet its assumptions. Outliers can be detected using Pearson residuals, deviance residuals, and leverage values, among others. In this thesis, outlier detection will be conducted in the case of rural poverty in Indonesia in 2022. The rural poverty data used is in percentage form, making it suitable for beta regression, which is useful for modeling data with response variables in the standard unit interval (0,1). Moreover, beta regression is also very useful when the dependent variable's distribution does not follow a normal distribution and when traditional linear models cannot be used due to unmet classical assumptions. Parameter estimation for the model uses the Maximum Likelihood Estimator. Based on the available data, using a 5% significance level, it can be concluded that in Indonesia in 2022, rural poverty is only influenced by the rural Gini ratio. Subsequently, outlier detection was conducted through the calculation of Pearson residuals, deviance residuals, and leverage values, which indicated that the provinces of Aceh, Riau, Bengkulu, Maluku, West Papua, and Papua are considered outliers. Therefore, an appropriate analysis method is needed when there are outliers in the data.

Keywords: *beta distribution, beta regression, outliers, deviance residuals, leverage values, rural poverty.*

Abstrak. Dalam set data tertentu, mungkin akan terdapat satu buah atau sekelompok data yang jauh dari sekumpulan data pada model regresi dan data tersebut perlu dicurigai sebagai pencilan. Pencilan ini berpotensi sebagai penentu signifikansi koefisien regresi. Selain itu, kehadiran pencilan dalam data dapat menyebabkan model regresi tidak memenuhi asumsinya. Pencilan dapat dideteksi diantaranya menggunakan residu pearson, residu devians, dan nilai *leverage*. Dalam skripsi ini, akan dilakukan deteksi pencilan pada kasus kemiskinan perdesaan di Indonesia tahun 2022. Data kemiskinan perdesaan yang digunakan berbentuk persentase, sehingga akan cocok dengan regresi beta yang berguna untuk memodelkan data dengan variabel respon yang berupa interval satuan standar (0,1). Selain itu, regresi beta sangat berguna pula ketika distribusi variabel dependen tidak mengikuti distribusi normal serta ketika model linier tradisional tidak bisa digunakan karena asumsi klasik tidak terpenuhi. Penaksiran parameter model menggunakan *Maximum Likelihood Estimator*. Berdasarkan data yang ada dengan menggunakan taraf signifikan 5% dapat disimpulkan bahwa di Indonesia tahun 2022, kemiskinan perdesaan hanya dipengaruhi oleh gini rasio perdesaan. Kemudian, dilanjutkan deteksi pencilan melalui perhitungan residu pearson, residu devians, dan nilai *leverage* yang menunjukkan Provinsi Aceh, Riau, Bengkulu, Maluku, Papua Barat, dan Papua dianggap sebagai pencilan. Sehingga, diperlukan metode analisis yang cocok ketika terdapat adanya data pencilan.

Kata Kunci: *distribusi beta, regresi beta, pencilan, residu devians, nilai leverage, kemiskinan perdesaan.*

A. Pendahuluan

Analisis regresi klasik merupakan alat statistika untuk mengevaluasi hubungan antara satu variabel dengan variabel lainnya atau satu variabel dengan beberapa variabel lainnya. Analisis regresi klasik ini cukup aplikatif karena bisa diterapkan diberbagai bidang. Namun, terdapat beberapa hal lain yang harus dipertimbangkan diantaranya mengenai situasi respon.

Tidak sedikit bidang studi yang melibatkan respon dalam bentuk persentase, rasio, ataupun proporsi yang artinya diukur secara kontinu pada skala interval 0 sampai 1. Model regresi klasik tidak akan cocok dalam memodelkan data tersebut, dikarenakan responnya dengan rentang yang terbatas [1]. Selain itu, varians tidak akan konstan dan distribusi dari respon yang tidak simetris. Respon dengan bentuk rasio akan berdistribusi beta karena memiliki interval satuan standar (0,1). Salah satu metode analisis regresi yang dapat digunakan untuk memodelkan respon interval (0,1) yaitu regresi beta. Model regresi beta akan memberikan estimasi parameter yang lebih akurat dan efisien dibandingkan dengan regresi klasik saat variabel respon yang diobservasi memiliki distribusi yang tidak simetris atau ketika muncul masalah heteroskedastisitas [2]. Untuk menaksir parameter model regresi beta dapat menggunakan metode *Maximum Likelihood Estimator* (MLE) dengan hasil turunan fungsi *log-likelihood* (Satyahadewi et al., 2023).

Setelah mendapatkan model, penting untuk mengidentifikasi pengamatan yang tidak lazim pada model dan data tersebut perlu dicurigai sebagai pencilan. Pencilan adalah ukuran kesesuaian antara data dan model yang digunakan. Keberadaan outlier dalam data dapat mengganggu proses analisis data, sehingga mengakibatkan residual dan varians pada data menjadi lebih besar [3]. Artinya, dengan adanya pencilan berakibat pada pengujian hipotesis yang hasilnya keliru pada akhirnya model terpengaruh dan tidak bisa digunakan (Dima et al., 2023).

Berdasarkan data BPS, masih ada sekitar 29 juta warga Indonesia yang hidup di bawah garis kemiskinan, dan dari jumlah tersebut, sekitar 18 juta orang tinggal di daerah perdesaan. Variabel indikator kemiskinan menjadi fokus penting dalam memengaruhi kemiskinan perdesaan. Nilai tukar petani berpengaruh negatif terhadap tingkat kemiskinan di perdesaan. Hal ini disebabkan karena nilai tukar petani mencerminkan harga jual produk pertanian dan harga pokok barang atau jasa yang digunakan oleh petani di pedesaan [4]. Faktor lain yang mempengaruhi tingkat kemiskinan adalah gini rasio. Distribusi pendapatan yang tidak merata menimbulkan ketimpangan pendapatan yang merupakan awal dari permasalahan kemiskinan. Selain nilai tukar petani dan gini rasio perdesaan, terdapat faktor lain yang mempengaruhi kemiskinan yaitu tenaga kerja sektor pertanian [5]. Jika sektor pertanian mengalami kesulitan atau tidak berkembang, hal ini dapat mempengaruhi pendapatan dan taraf hidup pekerja di sektor tersebut sehingga dapat berakibat pada kemiskinan (Hananti et al., 2023).

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini yaitu “Bagaimana hasil deteksi pencilan yang dilakukan pada regresi beta dalam memodelkan faktor-faktor yang mempengaruhi kemiskinan perdesaan di Indonesia tahun 2022?”. Penelitian ini ditujukan untuk mengetahui hasil deteksi pencilan yang dilakukan pada regresi beta dalam memodelkan faktor-faktor yang mempengaruhi kemiskinan perdesaan di Indonesia tahun 2022.

B. Metodologi Penelitian

Distribusi Beta

Distribusi beta sangat fleksibel [6]. Distribusi beta sering digunakan untuk data proporsi, rasio, ataupun persentase yang nilainya antara (0,1), ini dikarenakan fungsi densitas beta adalah fungsi densitas dua parameter yang didefinisikan pada interval tertutup $0 < y < 1$. Misalkan variabel acak Y dikatakan mempunyai distribusi peluang beta dengan parameter $a > 0$ dan $b > 0$ jika dan hanya jika fungsi densitas Y adalah

$$f(y; a, b) = \frac{1}{B(a, b)} y^{a-1} (1 - y)^{b-1}, \quad 0 < y < 1 \quad (1)$$

dengan $B(a, b)$ merupakan fungsi beta. Adapun hubungan antara fungsi beta dan fungsi gamma yaitu:

$$B(a, b) = \int_0^1 y^{a-1}(1-y)^{b-1} dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{dan} \quad \Gamma(a) = (a-1)\Gamma(a-1)$$

Maka untuk penulisan lainnya untuk fungsi densitas distribusi beta adalah:

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad 0 < y < 1 \quad (2)$$

$\Gamma(\cdot)$ adalah fungsi gamma. Rata-rata dan varians dari y masing-masing adalah sebagai berikut:

$$E(y) = \frac{a}{(a+b)} \quad \text{dan} \quad \text{var}(y) = \frac{ab}{(a+b+1)(a+b)^2}$$

Pendekatan yang umum digunakan untuk memodelkan respons adalah dengan menggunakan nilai rata-rata dari data respons sebagai perkiraan atau prediksi. Selain itu, dalam mendefinisikan suatu model, model tersebut berisi parameter dispersi. Untuk mencapai struktur regresi beta yang memodelkan rata-rata respon dan parameter dispersi, perlu menggunakan parameterisasi yang berbeda dari fungsi densitas beta.

Misalnya parameter baru yang dibuat yaitu $\mu = \frac{a}{(a+b)}$ dan $\phi = a + b$, di mana $a > 0$ dan $b > 0$.

Maka fungsi densitas peluang dari distribusi beta yang awalnya ditulis $B(a, b)$, kini diwakili oleh $B(\mu, \phi)$. Sehingga, fungsi densitas peluang distribusi beta yang telah diparameterisasi adalah:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad (3)$$

di mana $0 < y < 1$, $0 < \mu < 1$ dan $\phi > 0$

Rata-rata dan varians dari y masing-masing adalah sebagai berikut:

$$E(y) = \mu \quad \text{dan} \quad \text{var}(y) = \frac{(1-\mu)\mu}{(\phi+1)}$$

Model Regresi Beta

Fungsi densitas peluang distribusi beta yang telah di-parameterisasi melibatkan dua parameter yaitu μ dan ϕ . Oleh karena itu, dalam membentuk model regresi beta, perlu menggunakan dua fungsi hubung. Model linear umum untuk parameter lokasi (μ) adalah :

$$g(\mu_i) = x_i^T \beta \quad (4)$$

di mana $g(\cdot)$ merupakan fungsi monoton suatu fungsi hubung yang memiliki turunan.

Ada beberapa pilihan yang memungkinkan untuk fungsi hubung $g(\cdot)$. Misalkan menggunakan fungsi hubung logit $g(\mu) = \log\left\{\frac{\mu}{(1-\mu)}\right\}$, dan fungsi hubung probit $g(\mu) = \Phi^{-1}(\mu)$, dimana $\Phi(\cdot)$ adalah fungsi distribusi kumulatif dari variabel acak normal standar (baku), serta fungsi hubungan log-log komplementer $g(\mu) = \log\{-\log(1-\mu)\}$. Fungsi hubung logit merupakan fungsi hubung kanonik dan akan mengembalikan penaksir parameter ke dalam bentuk log odds rasio. Untuk fungsi hubung logit $g(\mu)$ adalah:

$$g(\mu) = \text{logit}(\mu) = \ln\left(\frac{\mu}{(1-\mu)}\right) = x_i^T \beta \quad (5)$$

$$\text{di mana } \mu = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

Estimasi Parameter Regresi Beta

Untuk menduga parameter model regresi beta salah satunya yaitu dengan menggunakan *Maximum Likelihood Estimator* (MLE). Prinsip dari maksimum *likelihood* adalah mencari β dengan memaksimalkan fungsi *likelihood*. Berikut ini merupakan *Maximum Likelihood Estimator* (MLE) :

$$L(\beta, \phi) = \prod_{i=1}^n f(\mu_i, \phi) = \prod_{i=1}^n \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} y^{\mu_i\phi-1}(1-y)^{(1-\mu_i)\phi-1} \quad (6)$$

Agar lebih mudah dalam mencari β , fungsi *likelihood* terlebih dahulu diubah menjadi logaritma natural sebagai berikut:

$$\ell(\beta, \phi) = \ln L(\beta, \phi) = \sum_{i=1}^n \log(f(\mu_i, \phi)) \tag{7}$$

dengan

$$\log(f(\mu_i, \phi)) = \log(\Gamma(\phi)) - \log(\Gamma(\mu_i\phi)) - \log\{\Gamma((1 - \mu_i)\phi)\} + (\mu_i\phi - 1) \log(y_i) + \{(1 - \mu_i)\phi - 1\} \log(1 - y_i)$$

Misalkan $y_i^* = \log\left\{\frac{y_i}{(1-y_i)}\right\}$ dan $\mu^* = \psi(\mu_i\phi) - \psi((1 - \mu_i)\phi)$.

Langkah selanjutnya, mencari matriks informasi Fisher. Matriks informasi Fisher ini digunakan untuk menghitung varian *Maximum Likelihood Estimator*. Untuk langkahnya yaitu:

Pertama, fungsi skor diperoleh dari turunan pertama dari fungsi *log-likelihood* terhadap β dan ϕ .

- Fungsi skor untuk β diberikan:

$$U_\beta(\beta, \phi) = \frac{\partial \ell(\beta, \phi)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i(\mu_i, \phi)}{\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \phi X^T T (y^* - \mu^*) \tag{8}$$

dengan X adalah sebuah matriks $n \times k$ pada baris ke- i yaitu x_i^T , $T = \text{diag}\left\{\frac{1}{g'(\mu_1)}, \dots, \frac{1}{g'(\mu_n)}\right\}$, $y^* = (y_1^*, \dots, y_n^*)$ dan $\mu^* = (\mu_1^*, \dots, \mu_n^*)$.

- Fungsi skor untuk ϕ diberikan:

$$U_\phi(\beta, \phi) = \frac{\partial \ell(\beta, \phi)}{\partial \phi} = \sum_{i=1}^n \{\mu_i(y_i^* - \mu_i^*) + \log(1 - y_i) - \psi((1 - \mu_i)\phi) + \psi(\phi)\} \tag{9}$$

Kedua, menurunkan kembali fungsi *log-likelihood* terhadap (β, ϕ) , yang disebut dengan matriks informasi Fisher.

Misalkan $W = \text{diag}\{w_1, \dots, w_n\}$ dengan elemen-elemen

$$w_i = \phi \{\psi'(\mu_i\phi) + \psi'((1 - \mu_i)\phi)\} \frac{1}{\{g'(\mu_i)\}^2}$$

dan $c_i = \phi \{\psi'(\mu_i\phi)\mu_i - \psi'((1 - \mu_i)\phi)(1 - \mu_i)\}$ dengan $c_i = (c_1, \dots, c_n)^T$. Dimana $\psi'(\cdot)$ adalah fungsi trigamma dan misalkan pula $D = \text{diag}\{d_1, \dots, d_n\}$ dengan $d_i = \psi'(\mu_i\phi)\mu_i^2 + \psi'((1 - \mu_i)\phi)(1 - \mu_i)^2 - \psi'(\phi)$. Kemudian, matriks informasi Fisher akan ditemukan setelah mendapatkan turunan kedua dari fungsi *log-likelihood* terhadap (β, ϕ) yaitu

$$\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n \frac{\partial}{\mu_i} \left(\frac{\partial \ell_i(\mu_i, \phi)}{\mu_i} \frac{d\mu_i}{d\eta_i} \right) \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} x_{ij} \tag{10}$$

Sehingga didapat matrix informasi Fishier untuk (β, ϕ) adalah:

$$K = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 \ell(\beta, \phi)}{\partial \beta^2} & \frac{\partial^2 \ell(\beta, \phi)}{\partial \beta \partial \phi} \\ \frac{\partial^2 \ell(\beta, \phi)}{\partial \beta \partial \phi} & \frac{\partial^2 \ell(\beta, \phi)}{\partial \phi^2} \end{pmatrix} = \begin{pmatrix} \phi X^T W X & X^T T c \\ c^T T^T X & \text{tr}(D) \end{pmatrix}$$

Persamaan-persamaan tersebut diperoleh dengan memaksimumkan fungsi *log-likelihood* secara numerik menggunakan algoritma optimasi nonlinier, seperti algoritma Newton, algoritma quasi-Newton, atau Fisher Scoring. Proses iterasi untuk estimasi parameter akan berhenti ketika kondisi konvergensi terpenuhi, yakni ketika $\|\beta^{(t+1)} - \beta^{(t)}\| \leq \epsilon$, di mana ϵ adalah bilangan yang sangat

kecil yang ditentukan oleh peneliti. Hasil estimasi parameter akhir diperoleh pada iterasi terakhir yakni $\beta^{(t+1)}$ [7].

Kemiskinan Perdesaan

Kemiskinan di daerah perdesaan dapat disebabkan oleh rendahnya produktivitas pertanian, akses terbatas terhadap sumber daya dan teknologi, serta ketidaksetaraan dalam distribusi pendapatan. Adapun persamaan rumus perhitungan persentase penduduk miskin yaitu sebagai berikut:

$$P_\alpha = \frac{1}{n} \sum_{i=1}^q \left[\frac{z - y_i}{z} \right]^\alpha \quad (11)$$

di mana:

- α = 0
- z = Garis kemiskinan
- y_i = Rata-rata pendapatan per kapita bulanan bagi penduduk yang hidup dalam kemiskinan ($i = 1, 2, 3, \dots, q$), $y_i < z$
- q = Penduduk yang berada di bawah batas kemiskinan dalam jumlah tertentu
- n = Jumlah penduduk

Penduduk yang memiliki tingkat pengeluaran per kapita per bulan di bawah Garis Kemiskinan (GK) dianggap sebagai penduduk yang hidup dalam kemiskinan. Garis Kemiskinan didapat dari penjumlahan garis kemiskinan makanan dan garis kemiskinan non makanan.

Faktor-faktor yang mempengaruhi kemiskinan Perdesaan

Variabel indikator kemiskinan menjadi fokus penting dalam memengaruhi kemiskinan perdesaan. Telah disebutkan sebelumnya, menurut penelitian sebelumnya faktor-faktor yang berpotensi mempengaruhi kemiskinan perdesaan diantaranya nilai tukar petani, gini rasio, dan tenaga kerja sektor pertanian. Pada penelitian ini data yang digunakan yaitu data sekunder yang bersumber dari website resmi Badan Pusat Statistik. Banyaknya data yang digunakan yaitu sebanyak unit penelitian yang berupa provinsi yang memiliki perdesaan yakni ada sebanyak 33 data. Variabel penelitian yang digunakan yaitu kemiskinan perdesaan (%), nilai tukar petani (poin), gini rasio perdesaan (poin), tenaga kerja informal sektor pertanian (%).

Uji Signifikansi Parameter Secara Simultan

Uji simultan digunakan untuk memeriksa ada atau tidaknya pengaruh variabel independen terhadap variabel dependen yang dilakukan secara bersama-sama atau serentak. Adapun hipotesis yang digunakan adalah sebagai berikut:

$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$; Tidak ada variabel independen yang mempengaruhi variabel dependen.

H_1 : Minimal ada satu $\beta_j \neq 0$ dimana $j = 1, 2, \dots, k$; Minimal terdapat satu variabel independen berpengaruh terhadap variabel yang dependen

Statistik uji yang digunakan adalah uji G yang merupakan uji rasio kemungkinan (*likelihood ratio test*) [8]:

$$G = -2 \ln \left[\frac{L_0}{L_1} \right] = -2 \ln [L_0 - L_1] \quad (12)$$

dimana:

L_0 = nilai *likelihood* untuk model tanpa variabel independen

L_1 = nilai *likelihood* untuk model dengan variabel independen

Statistik G mengikuti distribusi chi-kuadrat dengan derajat bebas 1. Kriteria pengujianya adalah tolak H_0 jika nilai statistik uji G melebihi $\chi^2_{(\alpha, k)}$ atau p -value lebih kecil daripada α . Sehingga dapat disimpulkan bahwa minimal terdapat satu variabel independen berpengaruh terhadap variabel dependen.

Uji Signifikansi Parameter Secara Parsial

Untuk melihat tingkat signifikansi parameter regresi ke- j maka dilakukan uji parsial yaitu pengujian dari setiap variabel secara individual. Hipotesis yang digunakan yaitu:

$H_0 : \beta_j = 0$, artinya variabel independen tidak berpengaruh signifikan secara parsial terhadap variabel dependen.

$H_1 : \beta_j \neq 0$, artinya variabel independen berpengaruh signifikan secara parsial terhadap variabel dependen.

untuk $j = 0, 1, 2, \dots, k$

Statistik uji yang digunakan adalah uji Wald :

$$|W| = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{13}$$

di mana:

β_j = nilai koefisien regresi

$se(\hat{\beta}_j)$ = standar deviasi dari β_j

Statistik uji Wald ini mengikuti distribusi normal dengan kriteria pengujianya adalah tolak H_0 jika nilai statistik uji Wald melebihi nilai kritis $Z_{(\alpha/2)}$ atau p -value lebih kecil daripada α , sehingga dapat disimpulkan bahwa variabel independen memiliki dampak yang signifikan secara parsial terhadap variabel dependen [9].

Mendeteksi Data Pencilan

Residu Pearson

Residu pearson dapat didefinisikan sebagai *studentized residual* atau residu yang dibakukan. Adapun perhitungan residu pearson dibakukan dapat didefinisikan sebagai berikut:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}} \tag{14}$$

di mana $\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$ dan $\widehat{var}(y_i) = \frac{(1-\hat{\mu}_i)\hat{\mu}_i}{(\hat{\phi}+1)} = \frac{[1-g^{-1}(x_i^T \hat{\beta})][g^{-1}(x_i^T \hat{\beta})]}{(\hat{\phi}+1)}$. Grafik dari residu terhadap indeks observasi ke- i seharusnya tidak menunjukkan pola. Jika nilai $|r_i|$ lebih dari 2, maka observasi ke- i tersebut merupakan data pencilan.

Residu Deavins

Devians merupakan statistik yang sering digunakan untuk statistik kecocokan model (*goodness-of-fit statistic*). Statistik ini mencerminkan perbedaan antara model penuh dan model yang diestimasi. Dalam konteks rasio kemungkinan, devians setara dengan dua kali perbedaan absolut antara *log-likelihood* untuk model yang diestimasi dengan *log-likelihood* untuk model penuh. Misalkan dengan itu diberikan

$$D(y; \mu, \phi) = \sum_{i=1}^n 2(\ell_i(\tilde{\mu}_i, \phi) - \ell_i(\mu_i, \phi)) \tag{15}$$

di mana $\tilde{\mu}_i$ merupakan nilai dari μ_i dengan menyelesaikan $\frac{\partial \ell_i}{\partial \mu_i} = 0$, yaitu $\phi(y_i^* - \mu_i^*) = 0$. Perlu

diperhatikan $D(y; \hat{\mu}, \hat{\phi}) = \sum_{i=1}^n (r_i^d)^2$. r_i^d sebagai penyimpangan residu ke- i . Maka residu devians dapat dituliskan sebagai berikut:

$$r_i^d = sign(y_i - \hat{\mu}_i) \left\{ 2 \left(\ell_i(\tilde{\mu}_i, \hat{\phi}) - \ell_i(\hat{\mu}_i, \hat{\phi}) \right) \right\}^{1/2} \tag{16}$$

di mana $\tilde{\mu}_i \approx y_i$. Observasi ke- i memberikan suatu jumlah $(r_i^d)^2$ pada penyimpangan, dan dengan demikian, observasi dengan nilai absolut r_i^d yang besar dapat dikatakan sebagai pencilan.

Nilai Leverage

Metode yang digunakan untuk mengidentifikasi pencilan terhadap variabel X adalah dengan menggunakan nilai pengaruh (*Leverage Point*) [10]. *Leverage* digunakan untuk menggambarkan

permasalahan yang terjadi pada variabel independen. Nilai *leverage* memberikan arti ukuran jarak dari titik observasi x_i ke \bar{x} . Nilai *leverage* diperoleh berdasarkan unsur-unsur diagonal dalam matriks HAT dengan menggantikan X dengan $W^{1/2}X$ yang dihitung berdasarkan persamaan berikut:

$$H = W^{1/2}X(X^T W X)^{-1}X^T W^{1/2} \quad (17)$$

dengan $W = \text{diag}(\widehat{\text{var}}(y_i)) = \text{diag}\left(\frac{(1-\hat{\rho}_i)\hat{\rho}_i}{(\hat{\phi}+1)}\right)$

Nilai *leverage* berkisar antara 0 dan 1. Kombinasi yang dapat menyebabkan adanya pengaruh yang besar terhadap model adalah nilai *leverage* yang besar yang diikuti oleh residu yang relatif besar pula. Nilai h_{ii} yang lebih besar daripada $2p/n$, dipandang mempunyai *leverage* yang besar dan mengindikasikan bahwa data ke- i merupakan data pencilan (*outlier*).

C. Hasil Penelitian dan Pembahasan

Evaluasi Model Regresi Beta

Pada penerapannya, dalam rangka mengetahui ada atau tidaknya data pencilan pada model analisis pengaruh kemiskinan perdesaan di Indonesia tahun 2022, dilakukan terlebih dahulu pemodelan dengan menggunakan regresi beta. Dari data kemiskinan perdesaan, nilai tukar petani, gini rasio perdesaan, dan tenaga kerja sektor pertanian, adapun model yang diperoleh dengan menggunakan software *Rstudio* adalah sebagai berikut:

Tabel 1. Model Analisis Regresi Beta

Parameter	Koefisien	Std. Error	W_{hitung}	P -value
Konstanta	-6,353672	1,817573	-3,496	0,00047
Nilai Tukar Petani	0,001703	0,008680	0,196	0,84444
Gini Rasio Perdesaan	6,940471	1,434791	4,837	0,000001
Tenaga Kerja Sektor Pertanian	2,365211	1,230094	1,923	0,05451
Skala	60,5535	14,93		

Maka model persamaannya adalah:

$$\text{logit}(\hat{\mu}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\text{logit}(\hat{\mu}) = -6,353672 + 0,001703X_1 + 6,940471X_2 + 2,365211X_3$$

di mana:

- $\hat{\mu}$: Kemiskinan Perdesaan (%)
- X_1 : Nilai Tukar Petani
- X_2 : Gini Rasio Perdesaan
- X_3 : Tenaga Kerja Informal Sektor Pertanian

Selanjutnya akan dilakukan uji simultan. Adapun hipotesis yang digunakan yaitu:

H_0 : $\beta_1 = \beta_2 = \beta_3 = 0$; Variabel nilai tukar petani, gini rasio perdesaan, dan tenaga kerja informal sektor pertanian tidak berpengaruh terhadap kemiskinan perdesaan

H_1 : Minimal ada satu $\beta_j \neq 0$ dimana $j = 1, 2, 3$; Minimal terdapat satu variabel independen yang berpengaruh terhadap kemiskinan perdesaan

Dengan menggunakan software *R-Studio* didapat statistik uji sebagai berikut:

$$G = -2 \ln[L_0 - L_1] = -2 \ln[45,40614 - 59,43228] = 28,05229$$

Kriteria uji yang digunakan yaitu tolak H_0 jika $G > \chi^2_{(\alpha, k)}$. Karena, $G = 28,05229 > \chi^2_{(0,05;3)} = 7,814728$ maka H_0 ditolak. Sehingga, berdasarkan data yang ada dengan menggunakan taraf signifikan 5% dapat disimpulkan bahwa minimal terdapat satu variabel independen yang berpengaruh terhadap kemiskinan perdesaan.

Dikarenakan dari hasil uji simultan terdapat minimal satu variabel yang berpengaruh, maka selanjutnya dilakukan uji parsial dengan melalui uji Wald. Uji parsial ini dilakukan untuk melihat pengaruh masing-masing variabel independen terhadap variabel dependen. Adapun hipotesis yang digunakan yaitu:

H_0 : $\beta_j = 0$; Variabel bebas ke- j tidak berpengaruh terhadap kemiskinan perdesaan

$H_1 : \beta_j \neq 0$; Variabel bebas ke- j berpengaruh terhadap kemiskinan perdesaan

Dengan menggunakan software R-Studio didapat statistik uji seperti pada Tabel 1.

Kriteria uji yang digunakan yaitu tolak H_0 jika $|W| > Z(\alpha/2)$. Dari Tabel 1 didapat:

- Konstanta : $|W| = 3,496 > Z\left(\frac{0,05}{2}\right) = 1,96$, maka H_0 ditolak
- Nilai Tukar Petani $|W| = 0,196 < Z\left(\frac{0,05}{2}\right) = 1,96$, maka H_0 diterima
- Gini Rasio Perdesaan : $|W| = 4,837 > Z\left(\frac{0,05}{2}\right) = 1,96$, maka H_0 ditolak
- Tenaga Kerja Sektor Pertanian : $|W| = 1,923 < Z\left(\frac{0,05}{2}\right) = 1,96$, maka H_0 diterima

Berdasarkan data yang ada dengan menggunakan taraf signifikan 5% dapat disimpulkan bahwa di Indonesia tahun 2022, kemiskinan perdesaan hanya di pengaruhi oleh gini rasio perdesaan.

Mendeteksi Data Pencilan

Residu Pearson

Residu pearson memberikan informasi tentang seberapa jauh titik data berada dari model regresi dalam satuan standar deviasi. Obervasi ke- i dikatakan data pencilan ketika memiliki nilai mutlak dari residu pearson yang lebih dari 2. Berikut merupakan perhitungan residu pearson untuk data pertama yaitu Provinsi Aceh:

$$r_1 = \frac{y_1 - \hat{\mu}_1}{\sqrt{\widehat{var}(y_1)}} = \frac{0,1706 - 0,08599180}{\sqrt{0,0012768920}} = 2,36774729$$

Dari perhitungan tersebut didapat nilai residu pearson untuk Provinsi Aceh yaitu sebesar 2,36774729. Perhitungan dilakukan hingga observasi ke-tigapuluhtiga yaitu Provinsi Papua yang disajikan pada Tabel 2.

Residu Devians

Residu devians ini pada dasarnya adalah perbedaan antara fungsi *log-likelihood* dari model penuh dengan fungsi *log-likelihood* dari model reduksi. Obervasi ke- i dikatakan data pencilan ketika memiliki nilai mutlak dari residu devians yang lebih dari 2. Adapun perhitungan residu devians untuk data pertama yaitu Provinsi Aceh yaitu sebagai berikut:

$$r_1^d = \text{sign}(y_1 - \hat{\mu}_1) \left\{ 2 \left(\ell_1(\tilde{\mu}_1, \hat{\phi}) - \ell_1(\hat{\mu}_1, \hat{\phi}) \right) \right\}^{1/2} \\ = (2(2,102221 - (-0,008465409)))^{1/2} = 2,05459808$$

Dari perhitungan tersebut didapat nilai residu devians untuk Provinsi Aceh yaitu sebesar 2,05459808. Perhitungan dilakukan hingga observasi ke-tigapuluhtiga yaitu Provinsi Papua yang disajikan pada Tabel 2.

Nilai Leverage

Nilai *leverage* merupakan diagonal dari matriks HAT. Rumus matriks HAT didasarkan pada matriks bobot (W) yang merupakan matriks diagonal dari varians data respons. Obervasi ke- i dikatakan data pencilan ketika memiliki nilai *leverage* yang lebih dari $2p/n$ atau $2(4)/33 = 0,242424$. Berikut merupakan hasil perhitungan nilai *leverage*.

$$H = \begin{bmatrix} 0,057969237 & 0,046947267 & \dots & -0,045814476 \\ 0,046947267 & 0,062700261 & \dots & -0,036135930 \\ \vdots & \vdots & \vdots & \vdots \\ -0,045814476 & -0,036135930 & \dots & 0,265735558 \end{bmatrix}$$

Nilai *leverage* merupakan diagonal dari matriks H. Untuk data pertama yaitu Provinsi Aceh memiliki nilai leverage sebesar 0,057969237, untuk Sumatera Utara memiliki nilai leverage sebesar 0,0627003, dan seterusnya hingga data ke-tigapuluhtiga yaitu Provinsi Papua yang disajikan pada Tabel 2.

Tabel 2. Nilai Residu Pearson, Residu Devians, dan Nilai *Leverage*

No	Provinsi	Residu Pearson	Residu Devians	Nilai <i>Leverage</i>
1	Aceh	2,3677473	2,0545981	0,0579692
2	Sumatera Utara	-0,1173107	-0,18424	0,0627003
3	Sumatera Barat	-0,3767189	-0,0872893	0,0768912
4	Riau	-0,2335798	-0,2130494	0,2427726
5	Jambi	-0,9149761	-0,804792	0,1891768
6	Sumatera Selatan	0,1785142	0,2904686	0,1262338
7	Bengkulu	1,1103225	1,1380044	0,2689667
8	Lampung	0,5085864	0,617188	0,069928
9	Kep, Bangka Belitung	0,5151154	0,6910131	0,1265815
10	Kepulauan Riau	0,7338878	0,8377004	0,0840854
11	Jawa Barat	-0,5812319	-0,4207287	0,0668644
12	Jawa Tengah	-0,5897722	-0,4571895	0,0358873
13	DI Yogyakarta	-0,7074921	-0,6090445	0,0687109
14	Jawa Timur	-0,2014191	-0,1129086	0,0400125
15	Banten	-0,7264157	-0,5688302	0,0805335
16	Bali	-1,5014501	-1,6317381	0,0949365
17	Nusa Tenggara Barat	-0,9855325	-0,9312062	0,1200147
18	Nusa Tenggara Timur	1,5662751	1,5063084	0,0984974
19	Kalimantan Barat	0,0686045	0,1802358	0,1787482
20	Kalimantan Tengah	-0,5210476	-0,2117766	0,2212024
21	Kalimantan Selatan	-1,1876512	-1,1785076	0,0564453
22	Kalimantan Timur	-0,1566475	-0,1868521	0,1171064
23	Kalimantan Utara	0,0037765	0,0377264	0,0902311
24	Sulawesi Utara	-1,3663411	-1,4001252	0,0614378
25	Sulawesi Tengah	0,6366114	0,730903	0,0756321
26	Sulawesi Selatan	-0,7219987	-0,6113719	0,049232
27	Sulawesi Tenggara	-0,8142362	-0,7304008	0,065265
28	Gorontalo	0,5479465	0,6058462	0,1713081
29	Sulawesi Barat	-0,6880845	-0,5771345	0,2058023
30	Maluku	3,6263912	2,9155542	0,0905928
31	Maluku Utara	-1,1637335	-1,1429941	0,0927305
32	Papua Barat	0,5363202	0,5742881	0,3477678
33	Papua	1,2727396	1,2591315	0,2657356

Observasi ke-*i* dikatakan sebagai data pencilan ketika memiliki kriteria:

- Nilai mutlak residu pearson : $|r_i| > 2$
- Nilai mutlak residu devians : $|r_i^d| > 2$
- Nilai h_{ii} yang lebih besar daripada $2p/n$, di mana p adalah banyaknya parameter regresi.

Sehingga $h_{ii} > \frac{(2)(4)}{33}$ atau $h_{ii} > 0,242424$

Dari Tabel 4.3 dan Tabel 4.4 terlihat bahwa Provinsi Aceh dan Maluku memiliki nilai mutlak dari residu pearson dan residu devians yaitu $r_1 = 2,3677473$, $r_1^d = 2,0545981$, $r_{30} = 3,6263912$, $r_{30}^d = 2,9155542$, yang artinya lebih dari 2. Selain itu, nilai *leverage* yang menunjukkan lebih dari 0,242424 yaitu Provinsi Riau, Bengkulu, Papua Barat, dan Papua yang memiliki nilai leverage berturut-turut sebesar $h_{4,4} = 0,2427726$, $h_{7,7} = 0,2689667$, $h_{32,32} = 0,3477678$, dan $h_{33,33} = 0,2657356$. Sehingga dapat disimpulkan bahwa Provinsi Aceh, Riau,

Bengkulu, Maluku, Papua Barat, dan Papua dianggap sebagai pencilan.

D. Kesimpulan

Berdasarkan penelitian yang telah dilakukan mengenai deteksi pencilan pada kasus kemiskinan di Indonesia tahun 2022 dapat disimpulkan bahwa pada penelitian ini analisis regresi beta dilakukan untuk melihat pengaruh dari variabel nilai tukar petani, gini rasio perdesaan, dan tenaga kerja sektor pertanian terhadap kemiskinan perdesaan di Indonesia tahun 2022. Data kemiskinan perdesaan yang digunakan berbentuk persentase. Variabel respon yaitu kemiskinan perdesaan dengan bentuk persentase atau interval $(0,1)$ akan cocok dimodelkan dengan regresi beta. Di lain sisi, regresi beta berguna ketika distribusi variabel respon tidak mengikuti distribusi normal serta model linier tradisional tidak bisa digunakan karena varians tidak akan konstan dan distribusi dari respon yang tidak simetris.

Pada penerapannya, dalam rangka mengetahui ada atau tidaknya data pencilan pada model analisis pengaruh kemiskinan perdesaan di Indonesia tahun 2022, dilakukan terlebih dahulu pemodelan dengan menggunakan regresi beta. Hasilnya menunjukkan hanya gini rasio perdesaan yang berpengaruh terhadap kemiskinan perdesaan di Indonesia tahun 2022. Setelah mendapatkan model, selanjutnya dilakukan deteksi pencilan. Berdasarkan perhitungan melalui residu pearson, residu devians, dan nilai *leverage* menunjukkan Provinsi Aceh, Riau, Bengkulu, Maluku, Papua Barat, dan Papua dianggap sebagai pencilan. Sehingga, diperlukan metode analisis lain yang cocok untuk pemodelan data respon kemiskinan perdesaan yang berbentuk persentase dan terdapat adanya data pencilan.

Acknowledge

Selesainya penelitian ini tidak terlepas dari dukungan seluruh pihak yang terlibat. Penulis ingin mengucapkan terima kasih kepada Allah SWT yang telah memberikan kelancaran dan kemudahan kepada penulis dalam proses penyusunan penelitian ini. Besar terima kasih pun kepada orang tua dan keluarga yang telah mendukung penulis baik melalui doa maupun materi. Selain itu, terkhusus kepada bapak Dr. Nusar Hajarisman., M.Si. selaku pembimbing akademik yang telah memberikan bimbingan dan pengarahan kepada penulis. Dan terakhir kepada teman-teman yang telah memberikan dorongan dan semangat kepada penulis.

Daftar Pustaka

- [1] Espinheira, P. Í. L., Ferrari, S. L. P., & Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4), 407–419. <https://doi.org/10.1080/02664760701834931>
- [2] Hajarisman, N. (2012). Penaksiran Parameter Model Regresi Beta untuk Memodelkan Data Proporsi. *Statistika*, 12(1)
- [3] Dewi, E. T. K., Agoestanto, A., & Sunarmi, S. (2016). Metode Least Trimmed Square (LTS) dan MM-Estimation untuk Mengestimasi Parameter Regresi Ketika Terdapat Outlier. *Unnes Journal of Mathematics*, 5(1), 47–54.
- [4] Fajar Yusuf Permana. (2021, October 28). Analisis Pengaruh Dana Desa, Nilai Tukar Petani dan Upah Buruh Tani terhadap Kemiskinan Perdesaan di Kawasan Timur Indonesia Tahun 2015-2019.
- [5] Indra, I. (2023). Pengaruh Tenaga Kerja Sektor Pertanian Dan Pdrb Sektor Pertanian Terhadap Pertumbuhan Ekonomi Dan Kemiskinan Di Indonesia. *Jurnal Ekonomi Integra*, 13(1), 116–125.
- [6] Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815.
- [7] Fathurahman, M. (2017). Regresi logistik untuk pemodelan indeks pembangunan kesehatan masyarakat kabupaten/kota di Pulau Kalimantan. *Seminar Nasional Matematika Dan Aplikasinya*.
- [8] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*

- (3rd ed). 2013.
- [9] Sa'adah, U. (2018). Penerapan Cochrane-Orcutt Iterative Procedure untuk Mengatasi Pelanggaran Asumsi Non Autokorelasi pada Analisis Regresi Linier Berganda Menggunakan Software R. *Prosiding Seminar Nasional Matematika Dan Pendidikan Matematika*, 1(2), 325–333.
- [10] Nurdin, N. N., Raupong, R., & Islamiyati, A. (2014). Penggunaan regresi robust pada data yang mengandung pencilan dengan metode momen. *Jurnal Matematika, Statistika Dan Komputasi*, 10(2), 114–123.
- [11] Dima, A., Kleden, M. A., & Atti, A. (2023). Analisis Faktor-Faktor yang Mempengaruhi Prestasi Siswa Menggunakan Metode Structural Equation Modeling (SEM). *Statistika*, 23(2), 132–146. <https://doi.org/10.29313/statistika.v23i2.2642>
- [12] Hananti, H., Jaya, I. G. N. M., & Irlandia Ginanjar. (2023). Pemodelan Kasus Gizi Buruk Balita di Indonesia Menggunakan Panel Quantile Regression Model. *Statistika*, 23(2), 116–122. <https://doi.org/10.29313/statistika.v23i2.2025>
- [13] Satyahadewi, N., Aprizkiyandari, S., & Risky Oprasianti. (2023). Regresi Data Panel dalam Analisis Faktor-Faktor yang Mempengaruhi IPM di Kalimantan Barat. *Statistika*, 23(2), 123–131. <https://doi.org/10.29313/statistika.v23i2.2201>