ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 14, Nomor 2, Tahun 2025, Halaman 631 - 641
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ESTIMASI PROBIT BINER SEMIPARAMETRIK SPLINE TRUNCATED PADA STATUS PERSENTASE PENDUDUK MISKIN

Rizki Adisetya Tahwin[1*], Vita Ratnasari[2], I Nyoman Budiantara[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember

***e-mail** : rizkiadisetya02@gmail.com

DOI: 10.14710/j.gauss.14.2.631-641

Article Info:
Received: 2025-10-24
Accepted: 2025-12-11
Available Online: 2025-12-30

Keywords:
Poor Population; Probit; Semiparametric; Spline Truncated.

Abstract: Regression analysis is a statistical method used to model the relationship between predictor variables and response variables. To analyse categorical response variables, probit regression can be used. There are three probit models based on the type of response variable, namely binary probit, multinomial probit, and ordinal probit. Binary probit is a method used to analyse response variables with two categories. There are three types of binary probit models based on curve approaches, namely parametric, non-parametric, and semi-parametric. Semi-parametric regression was chosen because it combines parametric and non-parametric components. Conventional semi-parametric regression often cannot provide accurate estimates. Therefore, the use of truncated splines is relevant because they can handle the flexibility of the functions. This study aims to estimate a semi-parametric binary probit model with truncated splines using maximum likelihood estimation. The resulting likelihood function is not in closed form, requiring Newton-Raphson numerical iteration. The results show that the best model is obtained with one knot point, which has an accuracy of 84.21% and an AUC of 0.84, indicating that the model's prediction classification is verry good.

1. PENDAHULUAN

Pemilihan metode statistika yang tepat menjadi hal penting untuk memperoleh hasil yang akurat dalam suatu analisis. Metode statistika yang dapat digunakan dalam analisis adalah regresi. Analisis regresi adalah teknik statistik untuk mengetahui hubungan antara variabel prediktor (x) dengan variabel respon (y) (Montgomery, 2012). Terdapat dua jenis data dalam proses analisis yaitu data kuantitatif dan data kualitatif. Metode yang dapat digunakan untuk menganalisis data dengan variabel dependen berupa kategori adalah analisis regresi probit. Terdapat tiga model probit berdasarkan jenis variabel responnya yaitu probit biner, probit multinomial, dan probit ordinal. Probit multinomial digunakan untuk menganalisis variabel dependen yang terdiri atas lebih dari dua kategori dan tidak berurutan. Model probit ordinal adalah metode analisis untuk variabel respon kategorikal yang terurut (Güneri et al., 2022). Model probit biner merupakan metode yang digunakan untuk menganalisis variabel respon yang bersifat kualitatif dengan dua kategori (Sari & Ratnasari, 2020). Variabel respon dalam model ini memiliki dua kategori, yaitu 1 untuk menunjukkan adanya suatu karakteristik dan 0 untuk ketidakberadaannya.

Terdapat tiga jenis model probit biner berdasarkan pendekatan kurvanya, yaitu parametrik, nonparametrik, dan semiparametrik. Model probit biner parametrik merupakan metode dimana bentuk kurva regresi antara variabel respons dan variabel prediktor diketahui (Eubank, 1999). Probit biner nonparametrik merupakan pendekatan dimana bentuk kurva dari fungsi regresinya tidak diketahui. Model probit biner semiparametrik digunakan ketika terdapat bentuk kurva yang diketahui dan tidak diketahui (Adrianingsih & Dani, 2021).

Pendekatan semiparametrik mengurangi struktur yang diterapkan dibandingkan pendekatan parametrik, namun tetap lebih terstruktur daripada pendekatan nonparametrik, sehingga menawarkan keseimbangan antara asumsi distribusi yang lebih ketat dan fleksibilitas yang lebih besar (Horowitz & Savin, 2001). Beberapa penelitian terkait probit biner semiparametrik adalah penelitian oleh (Du et al., 2019) yang menganalisis tikus dan mencit yang terpapar *chloroprene*, dimana proses estimasinya menggunakan *sieve maximum likelihood estimation* (SMLE). Penelitian lain dilakukan oleh (Liu & Qin, 2018) yang menganalisis data *current-status* univariat dalam studi konversi pre-diabetes menjadi diabetes tipe 2, penelitian ini mengestimasi parameter menggunakan algoritma *expectation-maximization* (EM) dan *isotonic regression*.

Model semiparametrik probit biner telah mengalami perkembangan, namun belum ada pengembangan mengenai metode probit biner semiparametrik *spline truncated. Spline* merupakan potongan polinomial yang tersegmen dan kontinu yang menyesuaikan estimasi data dengan pola pergerakan data, karena titik knot yang menandakan perubahan pola data (Eubank, 1999). Oleh karena itu, pengembangan metode probit biner semiparametrik *spline truncated* dapat menjadi alternatif dalam analisis model probit semiparametrik.

Model probit biner semiparametrik *spline truncated* dapat diaplikasikan pada bidang ekonomi yang dapat membantu menganalisis berbagai permasalahan yang kompleks. Indonesia, sebagai negara berkembang, masih menghadapi tantangan besar dalam hal pertumbuhan ekonomi yang belum pesat dan tingginya angka kemiskinan. Indonesia masih berjuang untuk mengurangi persentase penduduk miskin. Persentase penduduk miskin adalah persentase penduduk dengan rata-rata pengeluaran per kapita per bulan di bawah garis kemiskinan. Persentase penduduk miskin Indonesia pada tahun 2024 tercatat sebesar 8,57%, meskipun mengalami penurunan sebesar 0,79% dibandingkan tahun 2023, namun penurunan ini belum cukup signifikan. Persentase penduduk miskin dimasing-masing Provinsi di Indonesia nilainya bervariasi ada yang masuk kedalam kategori persentase penduduk miskin tinggi atau rendah sesuai standar nasional persentase penduduk miskin Indonesia sehingga cocok dimodelkan dengan probit biner karena memiliki dua kategori.

Faktor-faktor yang dianggap memengaruhi nilai persentase penduduk miskin adalah tingkat pengangguran terbuka dan rata-rata lama sekolah. Tingkat pengangguran terbuka merupakan persentase jumlah pengangguran terbuka terhadap jumlah angkatan kerja. Tingkat pengangguran terbuka yang meningkat menyebabkan pendapatan masyarakat menurun dan meningkatkan persentase penduduk miskin. Berbeda dengan tingkat pengangguran terbuka, rata-rata lama sekolah menunjukkan bahwa semakin banyak penduduk yang memiliki pendidikan tinggi maka akan lebih banyak penduduk yang berperan sebagai penggerak ekonomi. Penelitian oleh (Juliana et al., 2023) menjelaskan bahwa peningkatan rata-rata lama sekolah, laju pertumbuhan penduduk, dan inflasi dapat menggerakkan perekonomian khususnya persentase penduduk miskin. Oleh karena itu, diperlukan analisis yang lebih mendalam dengan metode yang sesuai yaitu probit biner semiparametrik *spline truncated* yang dapat mengestimasi nilai masing-masing faktor sebagai variabel prediktor yang memengaruhi persentase penduduk miskin di Indonesia sehingga diperoleh hasil dan interpretasi mengenai pengaruh faktor-faktor tersebut terhadap persentase penduduk miskin sebagai variabel respon. Hasil penelitian ini diharapkan dapat menjadi pertimbangan dalam merumuskan kebijakan untuk mengurangi persentase penduduk miskin secara berkelanjutan.

2. TINJAUAN PUSTAKA

Model probit biner semiparametrik *spline truncated* merupakan model yang digunakan untuk menganalisis hubungan antara variabel respons biner dengan variabel prediktor yang

sebagian bentuk kurva regresinya diketahui dan tidak diketahui. Variabel respon pada penelitian ini adalah persentase penduduk miskin yang dikategorikan menjadi kategori persentase penduduk miskin tinggi atau kategori 1 dan persentase penduduk miskin rendah atau kategori 0. Data yang digunakan dalam penelitian ini terdiri dari variabel respons biner dan variabel prediktor dengan struktur data ditampilkan pada Tabel 1.

Tabel 1. Struktur Data Penelitian

No	Provinsi	y_i	x_{1i}	t_{1i}
1	Aceh	y_1	x_{11}	t_{11}
2	Sumatera Utara	y_2	x_{12}	t_{12}
⋮	⋮	⋮	⋮	⋮
38	Papua Pegunungan	y_{38}	$x_{1,38}$	$t_{1,38}$

Model probit biner parametrik adalah metode untuk memodelkan hubungan antara variabel prediktor dengan variabel respon berbentuk biner. Menurut (Epriliyanti & Ratnasari, 2020) variabel y berasal dari variabel respon laten atau y_i^*. Bentuk umum probit biner parametrik dituliskan pada persamaan (1).

$$y_i^* = \boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta} + \varepsilon_i \tag{1}$$

Model probit biner nonparametrik merupakan pendekatan dimana bentuk kurva regresinya tidak diketahui dengan model umum terdapat pada persamaan (2).

$$y_i^* = \sum_{v=1}^{s} f_v(t_{vi}) + \varepsilon_i \tag{2}$$

Probit biner nonparametrik pada persamaan (2) didekati dengan *spline truncated* sehingga mampu mengatasi perubahan pola data pada sub interval tertentu (Izzah & Budiantara, 2020). Pada penelitian ini digunakan *spline truncated* linear yang memiliki derajat model 1 dengan titik knot $(K_1, K_2, \dots, K_r)$, maka diperoleh model *spline truncated* pada persamaan (3).

$$y_i^* = \alpha_0 + \sum_{v=1}^{s}\left(\alpha_{v1}t_{vi} + \sum_{u=1}^{r}\alpha_{v(u+1)}(t_{vi} - K_{vu})_+\right) + \varepsilon_i \tag{3}$$

Model regresi data kategori yang didekati dengan *spline truncated* adalah sebagai berikut:

$$\Phi^{-1}\pi_i(t_i) = \alpha_0 + \sum_{v=1}^{s}\left(\alpha_{v1}t_{vi} + \sum_{u=1}^{r}\alpha_{v(u+1)}(t_{vi} - K_{vu})_+\right) \tag{4}$$

$$\Phi^{-1}\pi_i(t_i) = \boldsymbol{T}\boldsymbol{\alpha} \tag{5}$$

Titik knot merupakan bagian dari polinomial tersegmen yang terhubung kontinu yang menunjukkan perubahan pola data. Titik knot yang digunakan dalam penelitian ini maksimal sebanyak 3 knot. Pemilihan titik knot yang akan digunakan dilakukan dengan metode *Akaike Information Criterion* (AIC). Nilai AIC yang lebih rendah menunjukkan keseimbangan model yang lebih baik antara kecocokan data dan kompleksitas model (Goepp et al., 2018). Nilai AIC diperoleh dengan rumus sebagai berikut (Aikake, 1974):

$$AIC = 2b - 2L(\alpha) \tag{6}$$

Probit biner semiparametrik merupakan model yang digunakan ketika terdapat bentuk kurva yang diketahui dan tidak diketahui. Penggabungan kedua pendekatan tersebut membuat model semiparametrik mampu menangkap hubungan yang kompleks tanpa kehilangan interpretasi dan efisiensi estimasi. Bentuk umum model semiparametrik terdapat pada persamaan (7).

$$y_i^* = \boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta} + \sum_{v=1}^{s} f_v(t_{vi}) + \varepsilon_i \tag{7}$$

Fungsi nonparametrik pada persamaan (7) didekati dengan *spline truncated*. Pada penelitian ini digunakan *spline truncated* linear yang memiliki derajat model 1, dengan model dituliskan pada persamaan (8).

$$y_i^* = \boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta} + \sum_{v=1}^{s}\left(\alpha_{v1}t_{vi} + \sum_{u=1}^{r}\alpha_{v(u+1)}(t_{vi} - K_{vu})_+\right) + \varepsilon_i \tag{8}$$

Fungsi truncated didefinisikan pada persamaan (9).

$$(t_{vi} - K_{vu})_+ = \begin{cases} (t_{vi} - K_{vu}), & t_{vi} \geq K_{vu} \\ 0, & t_{vi} < K_{vu} \end{cases} \quad (9)$$

Model probit biner semiparametrik *spline truncated* adalah sebagai berikut:

$$\Phi^{-1}\pi_i(t_i) = \boldsymbol{x_i^T\beta} + \sum_{v=1}^{s}\left(\alpha_{v1}t_{vi} + \sum_{u=1}^{r}\alpha_{v(u+1)}(t_{vi} - K_{vu})_+\right) \quad (10)$$

$$\Phi^{-1}\pi_i(t_i) = \boldsymbol{x_i^T\beta} + \boldsymbol{T\alpha} \quad (11)$$

dengan,

$\boldsymbol{x_i^T} = (\mathbf{1} \quad \boldsymbol{x_{1i}} \quad \cdots \quad \boldsymbol{x_{pn}})$

$\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p)^T$

$\boldsymbol{T}$ adalah matriks dari variabel nonparametrik *spline truncated* yang didefinisikan sebagai berikut:

$$\boldsymbol{T} = \begin{bmatrix} 1 & t_{11} & (t_{11} - K_{11})_+ & \cdots & (t_{11} - K_{1r})_+ & \cdots & t_{s1} & (t_{s1} - K_{s1})_+ & \cdots & (t_{s1} - K_{sr})_+ \\ 1 & t_{12} & (t_{12} - K_{11})_+ & \cdots & (t_{12} - K_{1r})_+ & \cdots & t_{s2} & (t_{s2} - K_{s1})_+ & \cdots & (t_{s2} - K_{sr})_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{1n} & (t_{1n} - K_{11})_+ & \cdots & (t_{1n} - K_{1r})_+ & \cdots & t_{sn} & (t_{sn} - K_{s1})_+ & \cdots & (t_{sn} - K_{sr})_+ \end{bmatrix}$$

$\boldsymbol{\alpha}$ adalah vektor dari parameter komponen nonparametrik *spline truncated* yang didefinisikan sebagai berikut:

$$\boldsymbol{\alpha} = [\alpha_{11} \quad \alpha_{12} \quad \cdots \quad \alpha_{1(1+r)} \quad \cdots \quad \alpha_{s1} \quad \alpha_{s2} \quad \cdots \quad \alpha_{s(1+r)}]^T$$

Probabilitas untuk persentase penduduk miskin tinggi dinotasikan dengan $y_i = 1$ dan rendah dinotasikan dengan $y_i = 0$ terdapat pada persamaan (12) dan (13).

$$P(y_i = 0) = 1 - \Phi\left(\boldsymbol{x_i^T\beta} + \sum_{v=1}^{s}\left(\alpha_{v1}t_{vi} + \sum_{u=1}^{r}\alpha_{v(u+1)}(t_{vi} - K_{vu})_+\right)\right) \quad (12)$$

$$P(y_i = 1) = \Phi\left(\boldsymbol{x_i^T\beta} + \sum_{v=1}^{s}\left(\alpha_{v1}t_{vi} + \sum_{u=1}^{r}\alpha_{v(u+1)}(t_{vi} - K_{vu})_+\right)\right) \quad (13)$$

Estimasi merupakan proses untuk memperoleh nilai parameter pada suatu model. Proses estimasi model probit biner semiparametrik *spline truncated* dilakukan dengan *Maximum Likelihood Estimation* (MLE). Proses estimasi diawali dengan menyusun fungsi *likelihood* pada persamaan (14).

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n}[\Phi(\boldsymbol{x_i^T\beta} + \sum_{v=1}^{s} f_v(t_{vi}))]^{yi}[1 - \Phi(\boldsymbol{x_i^T\beta} + \sum_{v=1}^{s} f_v(t_{vi}))]^{1-yi} \quad (14)$$

Fungsi likelihood pada persamaan (14) kemudian di ln kan pada persamaan (15)

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[y_i ln\Phi(\boldsymbol{x_i^T\beta} + \sum_{v=1}^{s} f_v(t_v)) + (1 - y_i)ln(1 - \Phi(\boldsymbol{x_i^T\beta} + \sum_{v=1}^{s} f_v(t_{vi})))\right] \quad (15)$$

$\boldsymbol{\theta}$ adalah vektor koefisien parameter komponen semiparametrik yang didefinisikan sebagai berikut:

$$\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \boldsymbol{\alpha}]^T$$

$l(\boldsymbol{\theta})$ yang telah diperoleh kemudian diturunkan dan disamadengankan nol. Proses estimasi tidak *close form* maka diselesaikan dengan iterasi *newton raphson* dengan ketentuan iterasi dicukupkan ketika telah mencapai nilai konvergen.

Evaluasi model penting dilakukan untuk mengidentifikasi kemungkinan kesalahan klasifikasi yang terjadi. Ukuran keakuratan model dihitung dengan *Apparent Error Rate* (APER) yang mengukur sampel yang salah diklasifikasikan (Agresti, 2002). Pengukuran ketepatan klasifikasi dapat menggunakan *confusion matrix* untuk menilai performa model dengan membandingkan hasil klasifikasi yang diprediksi dengan hasil klasifikasi yang sebenarnya (Hidayat et al., 2024). Tabel *confusion matrix* ditampilkan pada tabel 2.

Tabel 2. *Confusion Matrix*

Aktual	Kelompok Prediksi		
	0	1	Total
Kategori 0	TN *(True Negative)*	FP *(False Positive)*	TN+FP
Kategori 1	FN *(False Negative)*	TP *(True Positive)*	FN+TP
Total	TN+FN	FP+TP	TN+FP+FN+TP

Rumus untuk menghitung nilai *Apparent Error Rate (APER)* adalah sebagai berikut (Ibrahim, 2024):

$$APER = \frac{FP+FN}{TP+TN+FP+FN} \times 100\% \quad (16)$$

Nilai APER yang diperoleh berdasarkan persamaan (16) dapat digunakan untuk menghitung nilai *accuracy* pada persamaan (17)

$$Accuracy\ (\%) = 1 - APER \quad (17)$$

Pengujian ketepatan klasifikasi lain yang dapat digunakan selain APER dan *accuracy* adalah *specificity* dan *sensitivity*. *Specificity* dapat menilai kemampuan model dalam mengidentifikasi kasus negatif sebenarnya dan *sensitivity* dapat menilai kemampuan model dalam mengidentifikasi kasus positif sebenarnya. Rumus untuk menghitung *sensitivity* dan *specificity* adalah sebagai berikut (Hidayat et al., 2024).

$$Sensitivity = \frac{\text{TP}}{\text{TP+FN}} \quad (18)$$

$$Specificity = \frac{\text{TN}}{\text{TN+FP}} \quad (19)$$

Metode lain yang digunakan dalam evaluasi model adalah *Area Under Curve* (AUC), dimana AUC memiliki nilai antara 0 sampai 1. Nilai AUC yang mendekati 1, berarti bahwa performa klasifikasi yang terbentuk semakin baik. Kriteria keakuratan klasifikasi pada AUC disajikan pada tabel berikut (Simundic & Bio-One, 2014).

Tabel 3. Klasifikasi Nilai AUC

Nilai AUC	Kategori
$0{,}90 < AUC < 1{,}00$	*Excellent Classification*
$0{,}80 < AUC < 0{,}90$	*Very Good Classification*
$0{,}70 < AUC < 0{,}80$	*Good Classification*
$0{,}60 < AUC < 0{,}70$	*Sufficient Classification*
$0{,}50 < AUC < 0{,}60$	*Bad Classification*
$AUC < 0{,}50$	*Test Not Useful*

Uji statistik yang digunakan dalam menilai ketepatan klasifikasi adalah *Press'Q*. *Press'Q* merupakan ukuran untuk mengetahui kestabilan klasifikasi atau sejauh mana kelompok-kelompok tersebut dapat dipisahkan menggunakan variabel yang ada. Uji hipotesis yang digunakan dalam *Press'Q* dituliskan sebagai berikut:

H_0: Hasil klasifikasi model tidak stabil /tidak konsisten

H_1: Hasil klasifikasi model stabil / konsisten

Statistik uji *Press'Q* yang digunakan terdapat pada persamaan (20).

$$Press'Q = \frac{(N-cG)^2}{N(G-1)} \quad (20)$$

Pengambilan keputusan adalah H_0 ditolak apabila *Press'Q* > $\chi^2_{(\alpha;N-G)}$ (Hair et al., 2019).

Efek marginal membantu mempermudah proses interpretasi model karena koefisien model probit tidak dapat diinterpretasikan secara langsung sebagai perubahan probabilitas. Efek marginal diperoleh dengan menurunkan persamaan probabilitas pada persamaan (13) sehingga diperoleh persamaan efek marginal sebagai berikut:

$$\frac{\partial P(Y=1)}{\partial x_i} = \beta_p \phi\left(\boldsymbol{x_i^T \beta} + \sum_{v=1}^{s} f_v(t_{vi})\right) \quad (21)$$

$$\frac{\partial P(Y=1)}{\partial t_i} = \alpha_s \phi\left(\boldsymbol{x_i^T \beta} + \sum_{v=1}^{s} f_v(t_{vi})\right) \quad (22)$$

3. METODE PENELITIAN

Data pada penelitian ini merupakan data yang diperoleh dari Badan Pusat Statistik (BPS). Penelitian ini menggunakan unit penelitian berupa 38 provinsi di Indonesia. Data yang digunakan dalam penelitian ini adalah sebagai berikut:

y : Persentase penduduk miskin
x_1 : Tingkat pengangguran terbuka
t_1 : Rata-rata lama sekolah

Persentase penduduk miskin dibedakan menjadi dua kategori yaitu 1 (tinggi) dan 0 (rendah). Langkah-langkah untuk menganalisis persentase penduduk miskin pada penelitan ini adalah sebagai berikut:

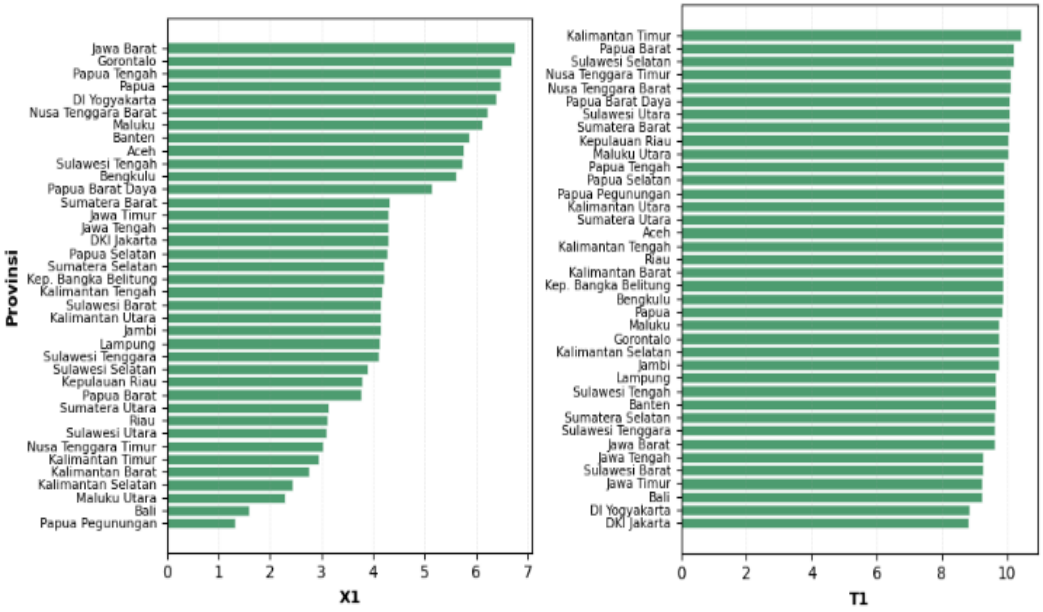
1. Mengumpulkan data persentase penduduk miskin disetiap provinsi di Indonesia dan mengkategorikannya secara biner yaitu 0 dan 1.
2. Melakukan eksplorasi dengan memeriksa hubungan antara masing-masing variabel prediktor dengan variabel respon.
3. Membuat analisis deskriptif pada variabel penelitian.
4. Memodelkan data persentase penduduk miskin di Indonesia menggunakan probit biner semiparametrik *spline truncated* yang diestimasi dengan *maximum likelihood estimation.*
5. Menyusun fungsi *likelihood* sesuai persamaan (14), kemudian di ln kan menjadi *ln-likehood* pada persamaan (15)
6. Menurunkan fungsi *ln-likelihood* dan menyamadenankan nol.
7. Karena tidak *close form* maka diselesaikan dengan iterasi *newton raphson* sampai konvergen.

$$|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}| < \varepsilon$$

8. Memilih titik knot optimal yang akan digunakan berdasarkan AIC terkecil sesuai rumus pada persamaan (6)
9. Memperoleh model terbaik data persentase penduduk miskin di Indonesia sesuai persamaan (10)
10. Menghitung evaluasi model melalui *Apparent Error Rate (*APER) pada persamaan (16), akurasi pada persamaan (17), sensitifitas pada persamaan (18), spesifisitas pada persamaan (19), AUC, dan *press-q* pada persamaan (20).
11. Melakukan interpretasi terhadap model terbaik yang diperoleh melalui efek marginal pada persamaan (21) dan (22).

4. HASIL DAN PEMBAHASAN

Statistika deskriptif adalah metode statistika yang berfungsi sebagai metode penyajian data secara sistematis dan konseptual yang dapat memberikan informasi mengenai variabel penelitian. Karakteristik data penelitian ditampilkan pada Gambar 1. Pada gambar 1 bagian (a) menggambarkan variasi tingkat pengangguran terbuka di 38 provinsi Indonesia. Jawa Barat memiliki nilai tingkat pengangguran terbuka tertinggi sebesar 6,75, sedangkan Papua Pegunungan memiliki nilai tingkat pengangguran terbuka terendah sebesar 1,32. Bagian (b) menunjukkan variasi rata-rata lama sekolah di 38 Provinsi di Indonesia. Kalimantan Timur memiliki rata-rata lama sekolah tertinggi sebesar 10,43, sedangkan DKI Jakarta mencatatkan rata-rata lama sekolah terendah sebesar 8,81. Proses deskripsi data bertujuan untuk memperoleh pemahaman yang lebih jelas mengenai gambaran umum data penelitian.



(a) (b)

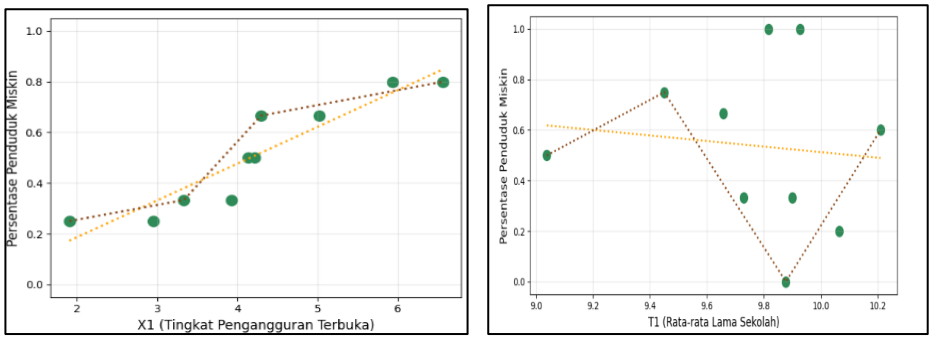
Gambar 1. Karakteristik Data Penelitian

Deskripsi data pada penelitian ini ditampilkan pada Tabel 4.

Tabel 4. Statistika Deskriptif Variabel Respon dan Prediktor

Variabel Prediktor	Kategori Respon	Mean	Median	Varians	Minimum	Maksimum
x_1	1	4,908	4,720	1,924	1,320	6,680
	0	3,713	3,840	1,538	1,590	6,750
t_1	1	9,735	9,815	0,120	8,850	10,210
	0	9,836	9,880	0,129	8,810	10,430

Tabel 4 menunjukkan statistika deskriptif variabel x_1 (tingkat pengangguran terbuka) dan t_1 (rata-rata lama sekolah). Pada variabel prediktor x_1 (tingkat pengangguran terbuka) kategori 1 memiliki rata-rata sebesar 4,908 dan varians sebesar 1,924 sehingga lebih tinggi dibandingkan kategori 0 dengan rata-rata sebesar 3,713 dan varians sebesar 1,538. Pada variabel prediktor t_1 (rata-rata lama sekolah) kategori 1 memiliki rata-rata sebesar 9,735 yang lebih rendah dibandingkan kategori 0 dengan rata-rata sebesar 9,836, sedangkan untuk varians kategori 1 sebesar 0,120 lebih rendah dibandingkan kategori 0 sebesar 0,129. Proses pemodelan pada model probit biner semiparametrik *spline truncated* terdapat data yang diklasifikasikan untuk pendekatan parametrik dan nonparametrik. Memeriksa kesesuain data dapat dilakukan melalui plot untuk masing-masing variabel penelitian yang terdapat pada Gambar (2).



(a) (b)

Gambar 2. Plot Variabel Penelitian

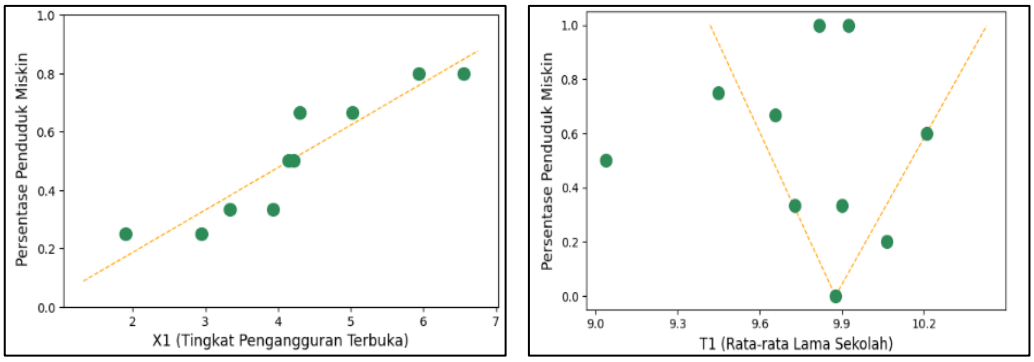
Pada Gambar 2 bagian (a) menunjukkan pola hubungan linear sehingga dipilih sebagai variabel parametrik. Pada bagian (b) menunjukkan pola hubungan nonlinear yang diawali dengan tren turun dan kemudian menunjukkan tren naik sehingga dipilih sebagai variabel nonparametrik dan cocok didekati dengan *spline truncated*.

Pemodelan probit biner semiparametrik *spline truncated* memerlukan penentuan titik knot untuk membangun model yang akurat dan interpretatif. Titik knot yang digunakan dibatasi sebanyak tiga knot dengan hasil perhitungannya ditampilkan pada tabel 5.

Tabel 5. Hasil Estimasi dan Kandidat Titik Knot

Prediktor x_1		Prediktor t_1		Jumlah Knot	K_{vu}	Nilai Knot	AIC
Parameter	Estimasi	Parameter	Estimasi				
β_0	1,648	α_{11}	-0,336	1	K_{11}	9,875	53,262
β_1	0,394	α_{12}	0,167				
β_0	2,619	α_{11}	-0,447	2	K_{11}	9,760	55,200
β_1	0,402	α_{12}	1,251		K_{12}	9,917	
		α_{13}	-1,511				
β_0	-1,007	α_{11}	-0,044	3	K_{11}	9,662	56,868
		α_{12}	-1,849		K_{12}	9,875	
β_1	0,386	α_{13}	5,436		K_{13}	10,005	
		α_{14}	-5,415				

Berdasarkan Tabel 5 diperoleh nilai AIC terendah sebesar 53,262 terdapat pada satu titik knot pada titik 9,875, sehingga titik knot ini yang akan digunakan untuk pemodelan probit biner semiparametrik *spline truncated*. Plot awal pada gambar 2 bagian (b) menunjukkan pola nonlinear yang dapat didekati dengan *spline truncated*, sehingga plot linear dan nonlinear dengan titik knot sebesar 9,875 ditampilkan pada Gambar (3).



(a) (b)

Gambar 3. Plot Linear dan Nonlinear Variabel Penelitian

Berdasarkan Gambar 3 bagian (a) menunjukkan bahwa plot membentuk pola linear yang konsisten dari kiri bawah ke kanan atas tanpa ada perubahan kemiringan. Plot pada gambar 3 bagian (b) menunjukkan adanya perubahan kemiringan yang cukup drastis dari atas ke bawah di sekitar titik knot yang merupakan ciri khas *spline truncated* yang mampu menangkap perubahan pola nonlinear dengan baik.

Pemodelan probit biner semiparametrik *spline truncated* memerlukan proses estimasi untuk mendapatkan nilai estimasi masing-masing parameter. Hasil estimasi pada tabel 5 digunakan untuk pemodelan. Model probit biner semiparametrik *spline truncated* pada data persentase penduduk miskin Indonesia terdapat pada persamaan (23).

$$\hat{y}_i^* = 1{,}648 + 0{,}394x_{1i} - 0{,}336t_{1i} + 0{,}167(t_{1i} - 9{,}875)_+ \quad (23)$$

Model probit biner semiparametrik *spline truncated* pada persamaan (23) digunakan untuk menghitung probabilitas persentase penduduk miskin. Nilai probabilitas digunakan

dalam evaluasi melalui *confusion matrix*. Proses perhitungan probabilitas dilakukan dengan mengambil contoh nilai pada x_{11} dan t_{11} sesuai persamaan (13) adalah sebagai berikut:

$$\hat{P}(y_i = 1) = \Phi(1{,}648 + 0{,}394x_{1i} - 0{,}336t_{1i} + 0{,}167(t_{1i} - 9{,}875)_+) = 0{,}723 \quad (24)$$

Model perlu dievaluasi untuk menilai peforma model dalam keakuratan klasifikasi data. Nilai probabilitas yang diperoleh dengan mengambil contoh x_{11} dan t_{11} pada persamaan (24) diklasifikasikan melalui *confusion matrix* yang ditampilkan pada tabel 6.

Tabel 6. Hasil *Confusion Matrix*

Aktual	Kelompok Prediksi		
	0	1	Total
Kategori 0	15	3	18
Kategori 1	3	17	20
Total	18	20	38

Berdasarkan tabel 6 diperoleh nilai APER sebesar 15,79% yang berarti bahwa sampel yang salah diklafisikasikan sebesar 15,79%. *Accuracy* yang diperoleh dalam analisis sebesar 84,21%, hal ini menunjukkan bahwa keakuratan pengklasifikasian sangat baik. *Sensitivity* pada evaluasi model diperoleh sebesar 85%, dan *specificity* sebesar 83,33% yang menunjukkan kemampuan model dalam mengidentifikasi kasus kategori 0 dan 1 sangat baik. Metode lain untuk menguji performa klasifikasi adalah AUC. Nilai AUC apabila semakin mendekati nilai 1 maka klasifikasi model sangat baik. Nilai AUC yang diperoleh adalah sebesar 0,84 yang termasuk kategori *verry good classification*. Evaluasi performa model dengan *Press'Q* dilakukan untuk menilai kestabilan klasifikasi model. Nilai *Press'Q* yang diperoleh sebesar 17,79, maka diperoleh bahwa nilai *Press'Q* > $\chi^2_{(\alpha;N-G)}$ sehingga menolak H_0 dan klasifikasi telah stabil atau konsisten.

Nilai efek marginal digunakan untuk mengetahui besarnya pengaruh masing-masing variabel prediktor. Efek marginal memudahkan interpretasi dalam pemodelan, diambil contoh untuk x_{11} dan t_{11} sehingga diperoleh efek marginalnya adalah sebagai berikut:

$$\frac{\partial \hat{P}(Y=1)}{\partial x_i} = 0{,}394 \times \phi(1{,}648 + 0{,}394x_{1i} - 0{,}336t_{1i} + 0{,}167(t_{1i} - 9{,}875)_+) = 0{,}1318 \quad (25)$$

$$\frac{\partial \hat{P}(Y=1)}{\partial t_i} = -0{,}169 \times \phi(1{,}648 + 0{,}394x_{1i} - 0{,}336t_{1i} + 0{,}167(t_{1i} - 9{,}875)_+) = -0{,}0566 \quad (26)$$

Setelah mendapatkan nilai dari perhitungan efek marginal, model dapat diinterpretasikan. Hasil yang diperoleh menunjukkan bahwa kenaikan satu satuan pada tingkat pengangguran terbuka akan meningkatkan persentase penduduk miskin sebesar 0,1318. Hal ini menunjukkan bahwa terdapat hubungan yang linear atau searah antara persentase penduduk miskin dengan tingkat pengangguran terbuka. Sementara itu, apabila rata-rata lama sekolah naik satu satuan maka akan menurunkan persentase penduduk miskin sebesar 0,0566. Hal ini menunjukkan bahwa terdapat hubungan yang nonlinear atau tidak searah antara persentase penduduk miskin dengan rata-rata lama sekolah. Hasil interpretasi pada efek marginal menunjukkan bahwa tingkat pengangguran terbuka di Indonesia harus diturunkan dengan menyediakan lapangan pekerjaan yang sesuai dengan kemampuan masing-masing individu sehingga dapat mendorong peningkatan perekonomian nasional. Berdasarkan hasil interpretasi efek marginal pada rata-rata lama sekolah menjelaskan bahwa rata-rata lama sekolah harus ditingkatkan sebagai potensi memperoleh pekerjaan yang strategis dengan peningkatan kualitas sumber daya manusia sebagai penggerak perekonomian nasional.

5. KESIMPULAN

Penelitian yang telah dilakukan dengan *maximum likelihood estimation* dan iterasi *newton raphson* memberikan hasil bahwa model terbaik probit biner semiparametrik *spline truncated* dengan satu titik knot adalah sebagai berikut:

$$\hat{y}_i^* = 1{,}648 + 0{,}394x_{1i} - 0{,}336t_{1i} + 0{,}167(t_{1i} - 9{,}875)_+$$

Model probit biner semiparametrik *spline truncated* memiliki nilai APER sebesar 15,79%, *accuracy* sebesar 84,21%, *sensitivity* sebesar 85%, dan *spesificity* sebesar 83,33%. Selain itu, nilai AUC yang diperoleh sebesar 0,84 yang termasuk kategori *verry good classification*. Nilai Press'Q yang diperoleh sebesar 17,79 sehingga diperoleh bahwa klasifikasi telah stabil atau konsisten. Model probit biner semiparametrk *spline truncated* dapat digunakan sebagai alternatif dalam menganalisis data kategori. Hasil penelitian ini diharapkan dapat berguna untuk pertimbangan pemerintah dalam membuat kebijakan dan strategi dalam mengurangi persentase penduduk miskin dengan meningkatkan lapangan pekerjaan dan kualitas sumber daya manusia melalui pendidikan.

DAFTAR PUSTAKA

Adrianingsih, N. Y., & Dani, A. T. R. (2021). Estimasi Model Regresi Semiparametrik Spline Truncated Menggunakan Maximum Likelihood Estimation (MLE). *Jambura Journal of Probability and Statistics*, *2*(2), 56–63.

Agresti, Alan. (2002). *Categorical data analysis*. Wiley-Interscience.

Aikake, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, *19*(6).

Du, M., Hu, T., & Sun, J. (2019). Semiparametric probit model for informative current status data. Statistics in Medicine, 38(12), 2219–2227.

Epriliyanti, Y. A., & Ratnasari, V. (2020). Pemodelan Faktor-faktor yang Mempengaruhi Keefektifan Sistem Pembelajaran Daring (SPADA) Menggunakan Regresi Probit Biner (Studi Kasus: Mahasiswa ITS Masa Pandemi COVID-19). *Inferensi*, *3*(2), 115–122.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*.

Goepp, V., Bouaziz, O., & Nuel, G. (2018). Spline Regression with Automatic Knot Selection. *Journal of Applied Statistics*, *45*(2), 212–229.

Güneri, Ö. İ., Durmuş, B., & İncekırık, A. (2022). Ordered Choice Models: Ordinal Logit and Ordinal Probit. *JIS Journal of Interdisciplinary Sciences*, *6*(2), 21–41.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivriate Data Analysis* (8th ed.).

Hidayat, F. M., Kusrini, & Yaqin, A. (2024). Dielektrika-Jurnal Ilmiah Kajian Teori dan Aplikasi Teknik Elektro Penerapan Algoritma Naïve Bayes Classifier Untuk Klasifikasi Status Gizi Stunting Pada Balita. *Jurnal Ilmiah Kajian Teori Dan Aplikasi Teknik Elektro* , *11*(2), 107–118.

Horowitz, J. L., & Savin, N. E. (2001). Binary Response Models: Logits, Probits and Semiparametrics. *Journal of Economic Perspectives*, *15*(4), 43–56.

Ibrahim, N. S. (2024). Analisis Diskriminan Linear Robust dengan Penduga Minimum Covariance Determinant (Studi Kasus: Indeks Kerentanan Pangan Menurut Kabupaten/Kota di Indonesia Tahun 2023). *Emerging Statistics and Data Science Journal*, *2*(2).

Izzah, N., & Budiantara, I. N. (2020). Pemodelan Faktor-faktor yang Mempengaruhi Tingkat Partisipasi Angkatan Kerja Perempuan di Jawa Barat Menggunakan Regresi Nonparametrik Spline Truncated. *Inferensi*, *3*(1), 21–27.

Juliana, S. F., Taaha, Y. R., & Guampe*, F. A. (2023). Laju Pertumbuhan Penduduk dan Inflasi Di Indonesia Tahun 2001-2021. JIM: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah, 8(2), 230–239

Liu, H., & Qin, J. (2018). Semiparametric probit models with univariate and bivariate current-status data. Biometrics, 74(1), 68–76. https://doi.org/10.1111/biom.12709

Montgomery, D. C. (2012). Introduction to Linear Regression Analysis

Sari, I. N. I., & Ratnasari, V. (2020). Pemodelan Regresi Logistik dan Probit Biner pada Faktor yang Memengaruhi Ketercapaian Target Unmet Need di Provinsi Jawa Barat. *Jurnal Sains Dan Seni ITS*, *9*(2), 200–207.

Simundic, A.-M., & Bio-One, G. (2014). Measures of Diagnostic Accuracy: Basic Definitions. *Journal EJIFCC*, *19*(4).