Imajiner: Jurnal Matematika dan Pendidikan Matematika
ISSN (Online): 2685-3892
Vol. 8, No. 2, Maret 2026, Hal. 511-517
Available Online at journal.upgris.ac.id/index.php/imajiner

Perbandingan Regresi Linear dan *Penalized Spline* dalam Analisis Indikator Pendidikan di Indonesia Tahun 2024

Sarmilah[1], Nonong Amalita[2]
[1,2]Universitas Negeri Padang
[1]sarmilah763@gmail.com

ABSTRAK

Penelitian ini bertujuan mengevaluasi kecukupan asumsi linearitas dalam memodelkan hubungan antara indikator sosial ekonomi dan rata-rata lama sekolah (RLS) kabupaten/kota di Indonesia tahun 2024. Secara khusus, penelitian ini memodelkan pengaruh kemiskinan, angka harapan hidup, dan pengeluaran per kapita terhadap RLS menggunakan regresi linear, menerapkan pendekatan *penalized spline* dalam kerangka *Generalized Additive Model* (GAM), serta membandingkan kinerja kedua pendekatan berdasarkan kriteria evaluasi statistik. Data yang digunakan mencakup 514 kabupaten/kota di Indonesia. Hasil analisis menunjukkan bahwa regresi linear mampu mengidentifikasi pengaruh signifikan kemiskinan dan pengeluaran per kapita terhadap RLS, namun asumsi linearitas belum sepenuhnya memadai dalam merepresentasikan struktur hubungan antarvariabel. Pendekatan *penalized spline* menghasilkan model dengan kemampuan penjelasan variasi yang lebih tinggi dan nilai AIC yang lebih rendah dibanding regresi linear. Efek parsial menunjukkan adanya pola nonlinear, termasuk kecenderungan penurunan RLS yang semakin tajam pada tingkat kemiskinan tinggi serta pola *diminishing return* pada pengeluaran per kapita. Temuan ini menunjukkan bahwa pendekatan semiparametrik lebih representatif dalam analisis indikator pendidikan regional yang bersifat heterogen.

Kata Kunci: *penalized spline*; regresi linear; *generalized additive model*; rata-rata lama sekolah; indikator pendidikan.

ABSTRACT

This study aims to evaluate the adequacy of the linearity assumption in modeling the relationship between socio-economic indicators and the mean years of schooling (MYS) of regencies/municipalities in Indonesia in 2024. Specifically, the study models the effects of poverty rate, life expectancy, and per capita expenditure on MYS using linear regression, applies a penalized spline approach within the framework of the Generalized Additive Model (GAM), and compares the performance of both approaches based on statistical evaluation criteria. The dataset consists of 514 regencies/municipalities across Indonesia. The results indicate that linear regression identifies significant effects of poverty and per capita expenditure on MYS; however, the linearity assumption does not fully capture the underlying relationship structure. The penalized spline model provides higher explanatory power and lower AIC values compared to the linear model. Partial effect analysis reveals nonlinear patterns, including a sharper decline in MYS at higher poverty levels and a diminishing return pattern for per capita expenditure. These findings suggest that a semiparametric approach is more appropriate for analyzing heterogeneous regional education indicators.

Keywords: *penalized spline; linear regression; generalized additive model; mean years of schooling; education indicators*

PENDAHULUAN

Rata-rata lama sekolah (RLS) merupakan indikator utama dalam pengukuran pembangunan manusia yang merepresentasikan akumulasi capaian pendidikan formal suatu wilayah. Dalam kerangka pembangunan ekonomi, pendidikan dipandang sebagai investasi modal manusia yang berkontribusi terhadap produktivitas dan pertumbuhan jangka panjang (Hanushek & Woessmann, 2020). Oleh karena itu, analisis faktor-faktor yang memengaruhi RLS menjadi penting baik secara substantif maupun metodologis.

Secara empiris, capaian pendidikan dipengaruhi oleh kondisi ekonomi masyarakat. Tingkat kemiskinan berimplikasi pada keterbatasan akses terhadap sumber daya pendidikan, sedangkan kondisi kesehatan yang direpresentasikan melalui angka harapan hidup (AHH) mencerminkan kualitas modal manusia yang lebih luas. Selain itu, pengeluaran per kapita sering digunakan sebagai indikator kemampuan ekonomi rumah tangga yang berkorelasi dengan investasi pendidikan (Todaro & Smith, 2020; World Bank Group, 2018). Hubungan antara variabel-variabel tersebut dan RLS umumnya dianalisis menggunakan regresi linear berganda karena kemudahannya dalam interpretasi parameter serta landasan teoritisnya yang kuat dalam ekonometrika (Wooldridge, 2020).

Namun demikian, pendekatan regresi linear mensyaratkan bahwa hubungan antara variabel prediktor dan variabel respon bersifat linear dan konstan pada seluruh rentang pengamatan. Dalam konteks data sosial regional yang heterogen seperti kabupaten/kota di Indonesia, asumsi tersebut sering kali tidak sepenuhnya terpenuhi. Literatur statistik modern menegaskan bahwa ketika bentuk hubungan fungsional tidak diketahui secara pasti, pemaksaan model parametrik berpotensi menghasilkan kesalahan spesifikasi (*model misspecification*) yang berdampak pada estimasi yang bias dan interpretasi yang kurang akurat (Hastie dkk., 2017).

Sebagai alternatif, pendekatan semiparametrik melalui *Generalized Additive Model* (GAM) memungkinkan setiap variabel prediktor dimodelkan dalam bentuk fungsi halus yang diestimasi dari data. *Penalized spline* sebagai salah satu teknik dalam GAM mengombinasikan fleksibilitas fungsi *spline* dengan mekanisme penalti terhadap kelengkungan, sehingga model tetap stabil dan terhindar dari *overfitting* (Wood, 2020). Pendekatan ini relevan untuk data sosial ekonomi yang memiliki kemungkinan pola nonlinear, efek ambang (*threshold effect*), maupun pola pelandaian (*diminishing return*) sebagaimana dijelaskan dalam teori pembangunan (Todaro & Smith, 2020).

Dengan demikian, permasalahan metodologis dalam penelitian ini bukan semata-mata pada identifikasi pengaruh variabel sosial ekonomi terhadap RLS, melainkan pada evaluasi apakah asumsi linearitas masih memadai untuk merepresentasikan struktur hubungan antarvariabel. Oleh karena itu, penelitian ini bertujuan untuk:

1. Memodelkan hubungan antara kemiskinan, AHH, dan pengeluaran per kapita terhadap RLS menggunakan regresi linear;
2. Menerapkan pendekatan *penalized spline* dalam kerangka *Generalized Additive Model*;
3. Membandingkan kinerja kedua pendekatan tersebut berdasarkan kriteria evaluasi statistik.

Pendekatan komparatif ini diharapkan memberikan kontribusi metodologis dalam analisis indikator pendidikan, khususnya dalam konteks pemilihan model yang sesuai untuk data sosial regional yang kompleks.

METODE PENELITIAN

Penelitian ini menggunakan data sekunder Badan Pusat Statistik (BPS) tahun 2024 yang mencakup 514 kabupaten/kota di Indonesia. Variabel respon adalah rata-rata lama sekolah (RLS), sedangkan variabel respon adalah rata-rata lama sekolah (RLS), sedangkan variabel prediktor meliputi persentase kemiskinan, angka harapan hidup (AHH), dan logaritma natural pengeluaran per kapita. Transformasi logaritma pada pengeluaran per kapita dilakukan untuk menstabilkan skala dan mengurangi potensi heterogenitas varians pada rentang nilai yang relatif lebar (Wooldridge, 2020).

Pendekatan analisis dilakukan melalui dua kerangka pemodelan, yaitu regresi linear parametrik dan *penalized spline* dalam kerangka *Generalized Additive Model* (GAM).

Perbandingan dilakukan untuk mengevaluasi kesesuaian asumsi linearitas terhadap struktur data.

Model Regresi Linear Parametrik

Secara umum, model regresi linear berganda dinyatakan sebagai:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \ , \qquad i = 1, 2, \dots, n \tag{1}$$

(Wooldridge 71, 2020)

dengan:

y_i adalah variabel respon ke- i,

x_{ij} adalah nilai prediktor ke- j pada observasi ke- i,

β_j adalah parameter regresi, dan

ε_i adalah galat acak yang diasumsikan berdistribusi normal.

Parameter β diestimasi menggunakan metode *Ordinary Least Squares* (OLS), yaitu dengan meminimalkan jumlah kuadrat residual.

Model *Penalized Spline* dalam Kerangka GAM

Untuk mengakomodasi kemungkinan hubungan nonlinear, digunakan pendekatan *Generalized Additive Model* (GAM). Secara umum, model aditif dinyatakan sebagai:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i \tag{2}$$

(Hastie dkk. 297, 2017)

dengan $f_j(\cdot)$ adalah fungsi halus yang tidak ditentukan bentuk parametriknya secara eksplisit dan diestimasi menggunakan basis *penalized spline* dengan estimasi *Restricted Maximum Likelihood* (REML).

Kriteria Evaluasi Model

Evaluasi model dalam penelitian ini dilakukan menggunakan beberapa ukuran performa statistik, yaitu *Adjusted R-Squared*, *Root Mean Square Error* (RMSE), *Akaike Information Criterion* (AIC), serta validasi silang. Pemilihan kriteria ini didasarkan pada literatur statistik dan pemodelan regresi modern.

Adjusted R-Squared

Koefisien determinasi R^2 mengukur proporsi variasi variabel respon yang dapat dijelaskan oleh model. Namun, R^2 cenderung meningkat seiring bertambahnya jumlah parameter, sehingga kurang tepat untuk membandingkan model dengan kompleksitas berbeda. Oleh karena itu digunakan *Adjusted R-Squared*, yang didefinisikan sebagai:

$$R^2_{Adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{3}$$

(Montgomery dkk. 333, 2012)

dengan n jumlah observasi dan p jumlah parameter.

Adjusted R^2 memberikan penalti terhadap penambahan parameter yang tidak memberikan kontribusi substansial pada model (Wooldridge, 2020). Dalam konteks perbandingan model parametrik dan semiparametrik, ukuran ini relevan untuk menilai kemampuan penjelasan variasi respon.

Root Mean Square Error (RMSE)

RMSE mengukur rata-rata kesalahan prediksi model dan didefinisikan sebagai:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

RMSE bersifat langsung terukur dalam satuan variabel respon dan sensitif terhadap kesalahan besar. Dalam literatur prediksi statistik dan *machine learning*, RMSE merupakan ukuran standar untuk membandingkan akurasi model regresi (Hastie dkk., 2017)

Akaike Information Criterion (AIC)

Kriteria AIC dirumuskan sebagai:

$$AIC = -2l(\hat{\theta}) + 2k \quad (5)$$

dengan k merupakan jumlah parameter efektif.

HASIL DAN PEMBAHASAN

Statistika Deskriptif

Tahap awal analisis dilakukan melalui pemeriksaan statistik deskriptif untuk memahami karakteristik data 514 kabupaten/kota di Indonesia tahun 2024. Ringkasan statistik disajikan pada Tabel 1.

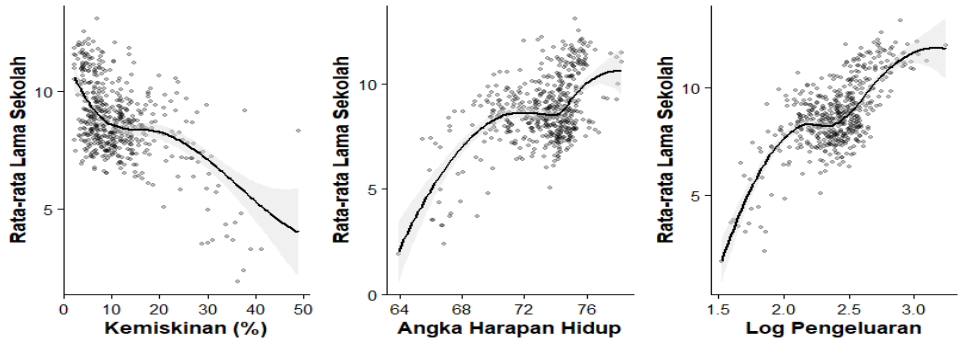
Tabel 1. Statistik Deskriptif Variabel Penelitian

Variabel	Minimum	Maksimum	Rata-rata
RLS (tahun)	1,920	13,100	8,737
AHH (tahun)	63,90	78,27	73,27
Kemiskinan (%)	2,23	48,93	11,25
Pengeluaran (juta)	4,597	25,573	11,439

Variasi RLS dan kemisikan yang cukup lebar menunjukkan heterogenitas antarwilayah. Secara teoritis, heterogenitas ini membuka kemungkinan hubungan yang tidak sepenuhnya linear, terutama pada rentang ekstrem (Hastie dkk., 2017).

Eksplorasi Awal

Untuk mengidentifikasi pola hubungan awal, dilakukan visualisasi *scatterplot* sebagaimana disajikan pada Gambar 1.



Gambar 1. Scatterplot dan Kurva Loess Hubungan (a) Kemiskinan terhadap RLS, (b) AHH terhadap RLS, (c) Pengeluaran terhadap RLS

Berdasarkan Gambar 1, hubungan antara kemiskinan dan RLS menunjukkan pola penurunan yang tidak konstan, dengan kemiringan yang lebih tajam pada tingkat kemiskinan tinggi. Hubungan pengeluaran terhadap RLS memperlihatkan kecenderungan pelandaian pada nilai tinggi, sedangkan AHH menunjukkan pola yang tidak sepenuhnya

linear. Gambar 1 ini mengindikasikan kemungkinan ketidaksesuaian asumsi linearitas dan mendukung penggunaan pendekatan yang lebih fleksibel.

Hasil Regresi Linear

Model regresi linear digunakan sebagai model pembanding awal. Hasil estimasi disajikan pada Tabel 2.

Tabel 2. Hasil Estimasi Regresi Linear

Variabel	Koefisien	Std. Error	*p-value*
Intersep	4,122	2,507	0,101
Kemiskinan (%)	-0.0326	0,0110	0,003
AHH (tahun)	-0,0666	0,0372	0,074
Pengeluaran (juta)	4,098	0,338	<0,001

Berdasarkan Tabel 2, terlihat bahwa kemiskinan berpengaruh negatif signifikan terhadap RLS, yang sejalan dengan laporan World Bank (2018) dan UNESCO (2024) mengenai dampak kemiskinan terhadap akses pendidikan. Pengeluaran per kapita berpengaruh positif signifikan terhadap RLS. Namun, AHH tidak signifikan pada taraf 5%.

Hasil *Penalized Spline*

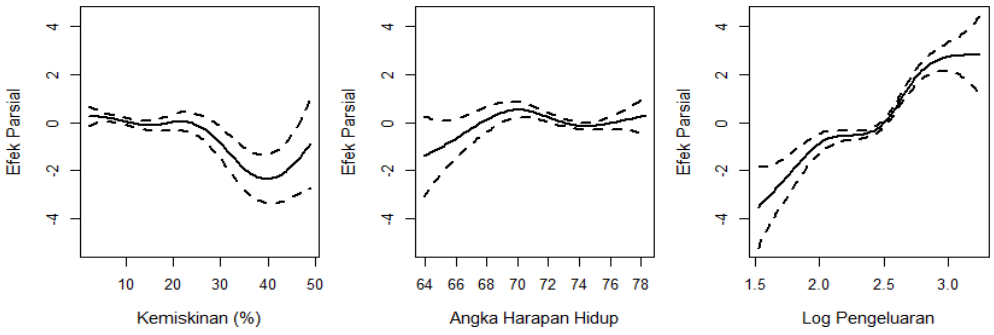
Model *penalized spline* dalam kerangka *Generalized Additive Model* (GAM) memberikan hasil yang berbeda. Ringkasan model *penalized spline* disajikan pada Tabel 3.

Tabel 3. Ringkasan Model *Penalized Spline*

Variabel	edf	F-statistik	*p-value*
s(Kemiskinan)	5,008	4,479	0,0002
s(AHH)	4,055	4,384	0,0008
s(Pengeluaran)	5,398	4,479	<0,001

Berdasarkan Tabel 3, seluruh fungsi halus signifikan dengan nilai *effective degrees of freedom* (edf) lebih besar dari 1. Menurut (Wood, 2020), nilai edf > 1 menunjukkan adanya kelengkungan fungsi, sehingga hubungan tidak bersifat linear sederhana.

Untuk memahami bagaimana variabel-variabel signifikan memengaruhi RLS, kurva spline parsial dari model GAM diplot dan ditunjukkan pada Gambar 2.



Gambar 2. Efek Parsial *Penalized Spline* untuk (a) Kemiskinan, (b) AHH, dan (c) Pengeluaran

Gambar 2 menunjukkan bahwa pengaruh kemiskinan terhadap RLS semakin tajam pada tingkat kemiskinan tinggi. Pola ini mengindikasikan adanya *threshold effect*, di mana dampak tambahan kemiskinan menjadi lebih besar setelah melewati tingkat tertentu. Temuan ini konsisten dengan laporan UNESCO (2024) mengenai risiko putus sekolah yang lebih tinggi pada wilayah dengan kemiskinan ekstrem.

Hubungan antara pengeluaran per kapita dan RLS menunjukkan pola *deminishing return*, yaitu peningkatan kesejahteraan pada wilayah berpendapatan rendah memberikan dampak pendidikan yang lebih besar dibanding pada wilayah berpendapatan tinggi (Todaro & Smith, 2020).

Sedangkan untuk AHH, pola nonlinear menunjukkan bahwa peningkatan kesehatan pada tingkat rendah berkorelasi lebih kuat dengan peningkatan RLS dibanding pada tingkat tinggi, sejalan dengan teori modal manusia (Hanushek & Woessmann, 2020)

Perbandingan Model

Perbandingan kinerja kedua model disajikan pada Tabel 4.

Tabel 4. Perbandingan Kinerja Model

Model	*Adjusted* R^2	RMSE	AIC
Regresi Linear	0,458	1,183	1641,545
Penalized Spline	0,572	1,040	1535,604

Berdasarkan Tabel 4, dapat dilihat terdapat penurunan AIC ketika menggunakan model *penalized spline*, menunjukkan efisiensi model *spline* yang lebih tinggi. Peningkatan *Adjusted* R^2 sebesar $\pm$11% menunjukkan kemampuan model dalam menangkap variasi yang sebelumnya tidak terwakili oleh model linear. Hal ini menunjukkan bahwa metode *penalized spline* lebih fleksibel dibanding regresi linear biasa dalam memodelkan hubungan antar variabel yang juga sejalan dengan penelitian Sarmilah dkk. (2025), di mana model *penalized spline* mampu menangkap pola non-linear yang muncul pada data, sehingga estimasi lebih akurat.

Kelebihan dan Kekurangan Metode

Regresi linear memiliki kelebihan dalam interpretasi parameter yang sederhana dan struktur model yang eksplisit. Namun, metode ini bergantung pada asumsi linearitas dan homogenitas efek pada seluruh rentang data. Dalam konteks data sosial yang heterogen, asumsi tersebut dapat membatasi representasi hubungan sebenarnya.

Penalized spline memiliki kelebihan dalam fleksibilitas bentuk fungsi dan kemampuan menangkap perubahan kemiringan lokal. Penggunaan penalti kelengkungan melalui REML membantu menghindari *overfitting*. Namun demikian, interpretasi model *spline* tidak sesederhana regresi linear karena koefisien tidak langsung merepresentasikan perubahan marginal konstan.

Sintesis

Berdasarkan pengujian statistik, evaluasi kinerja, dan visualisasi efek parsial, hubungan antara indikator sosial ekonomi dan RLS bersifat nonlinear. *Penalized spline* memberikan representasi yang lebih komprehensif dibanding regresi linear. Dengan demikian, untuk analisis indikator pendidikan regional yang heterogen, pendekatan semiparametrik lebih sesuai dibanding pendekatan parametrik linear klasik.

PENUTUP

Berdasarkan tujuan penelitian dan hasil analisis yang telah dilakukan, diperoleh kesimpulan sebagai berikut:

1. Pemodelan menggunakan regresi linear menunjukkan bahwa kemiskinan dan pengeluaran per kapita berpengaruh signifikan terhadap rata-rata lama sekolah (RLS), sedangkan angka harapan hidup (AHH) tidak signifikan pada taraf 5%. Model linear mampu menjelaskan 45,8% variasi RLS. Namun, hasil eksplorasi visual dan evaluasi statistik menunjukkan bahwa asumsi linearitas belum sepenuhnya memadai dalam merepresentasikan struktur hubungan antarvariabel.
2. Penerapan *penalized spline* dalam kerangka *Generalized Additive Model* (GAM) menghasilkan model dengan kinerja yang lebih baik, ditunjukkan oleh *Adjusted* R^2

sebesar 57,2% dan nilai AIC yang lebih rendah dibanding regresi linear. Nilai *effective degrees of freedom* (EDF) yang lebih besar dari 1 pada seluruh vairabel mengindikasikan adanya hubungan nonlinear yang signifikan antara kemiskinan, angka harapan hidup, serta pengeluaran per kapita terhadap RLS.

3. Perbandingan kinerja kedua pendekatan menunjukkan bahwa *penalized spline* lebih efektif dalam menangkap struktur hubungan yang kompleks pada data indicator pendidikan regional. Model semiparametrik mampu mengidentifikasi pola ambang pada kemiskinan dan kecenderungan *diminishing return* pada pengeluaran per kapita yang tidak dapat direpresentasikan secara memadai oleh regresi linear.

Dengan demikian, permasalahan metodologis yang diangkat dalam penelitian ini terjawab, yaitu bahwa asumsi linearitas tidak sepenuhnya memadai untuk menggambarkan hubungan indikator sosial ekonomi terhadap RLS pada data regional yang heterogen.

Berdasarkan hasil tersebut, pendekatan *penalized spline* direkomendasikan dalam analisis data sosial ekonomi yang memiliki potensi ketidaksesuaian asumsi linearitas. Sebagai prospek pengembangan, penelitian selanjutnya dapat memperluas cakupan data ke dimensi waktu (panel data) atau mempertimbangkan pendekatan spasial untuk mengakomodasi ketergantungan antarwilayah. Pengembangan metode semiparametrik lain juga berpotensi memperkaya analisis struktur hubungan indikator pendidikan.

REFERENSI

Hanushek, E. A., & Woessmann, L. (2020). Education, knowledge capital, and economic growth. *The Economics of Education*, 171–182.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning* (2nd ed.). Springer.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). John Wiley & Sons.

Sarmilah, Fitri, F., & Imran, M. (2025). Penalized Spline Regression Modeling on the Human and Cultural Development Index (IPMK) for 2022. *UNP Journal of Statistics and Data Science*, *3*, 473–481.

Todaro, M. P., & Smith, S. C. (2020). *Economic Development* (13th ed.). Pearson.

UNESCO. (2024). *Leadership in education*. United Nations Educational, Scientific and Cultural Organization.

Wood, S. N. (2020). Inference and computation with generalized additive models and their extensions. In *TEST* (Vol. 29, Issue 2). Springer.

Wooldridge, J. M. (2020). *Introductory Econometrics A Modern Approach* (7th ed.). Cengage Learning.

World Bank Group. (2018). *World Development Report: Learning to Realize Education's Promise*.