SKRIPSI

PEMODELAN SEMIPARAMETRIK DENGAN KOEFISIEN BERVARIASI PADA DATA LONGITUDINAL MENGGUNAKAN PENAKSIR *B-SPLINE*

Disusun dan diajukan oleh

NUR APRILIA DZULHIJJAH

H051171016

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2021

PEMODELAN SEMIPARAMETRIK DENGAN KOEFISIEN BERVARIASI PADA DATA LONGITUDINAL MENGGUNAKAN PENAKSIR *B-SPLINE*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

NUR APRILIA DZULHIJJAH

H051171016

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2021

.

LEMBAR PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama	: Nur Aprilia Dzulhijjah
NIM	: H051171016
Program Studi	: Statistika
Jenjang	: Sarjana (S1)

Menyatakan dengan ini bahwa karya tulis saya yang berjudul

PEMODELAN SEMIPARAMETRIK DENGAN KOEFISIEN BERVARIASI PADA DATA LONGITUDINAL MENGGUNAKAN PENAKSIR *B-SPLINE*

adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 27 Oktober 2021

NUR APRILIA DZULHIJJAH

NIM. H051171016

PEMODELAN SEMIPARAMETRIK DENGAN KOEFISIEN BERVARIASI PADA DATA LONGITUDINAL MENGGUNAKAN PENAKSIR *B-SPLINE*

Disetujui Oleh:

Pembimbing Utama,	Pembimbing Pertama,
Dr. Anna Islamiyati, S.Si., M.Si.	**Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.**
NIP. 19770808 200501 2 002	**NIP. 19740713 199903 2 001**

Ketua Departemen Statistika

Dr. Nurtiti Sunusi, S.Si., M.Si.

NIP. 19720117 199703 2 002

Pada Tanggal : 27 Oktober 2021

LEMBAR PENGESAHAN

PEMODELAN SEMIPARAMETRIK DENGAN KOEFISIEN BERVARIASI PADA DATA LONGITUDINAL MENGGUNAKAN PENAKSIR *B-SPLINE*

Disusun dan diajukan oleh

NUR APRILIA DZULHIJJAH

H051171016

Telah dipertahankan dihadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

pada tanggal 27 Oktober 2021

dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pertama,
Dr. Anna Islamiyati, S.Si., M.Si.	**Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.**
NIP. 19770808 200501 2 002	**NIP. 19740713 199903 2 001**

Ketua Departemen Statistika

Dr. Nurtiti Sunusi, S.Si., M.Si.

NIP. 19720117 199703 2 002

KATA PENGANTAR

Bismillahirrahmanirrahim

Assalamu'alaikum Warahmatullahi Wabarakatuh.

Alhamdulillahi robbil'alamin, Puji syukur kepada **Allah *Subhanahu Wa Ta'ala*** atas segala limpahan rahmat, nikmat, dan hidayah-Nya yang diberikan kepada penulis sehingga dapat menyelesaikan penulisan skripsi dengan judul **"Pemodelan Semiparametrik dengan Koefisien Bervariasi pada Data Longitudinal Menggunakan Penaksir *B-Spline*"** sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

Shalawat dan salam *InsyaAllah* senantiasa tercurah kepada **Rasulullah Muhammad *Shallallahu'alaihi Wasallam*** sebagai sebaik-baik teladan yang telah memberikan petunjuk cinta dan kebenaran dalam kehidupan.

Dalam penyelesaian skripsi ini, penulis telah melewati perjuangan panjang dan pengorbanan yang tidak sedikit dengan segala keterbatasan kemampuan dan pengetahuan. Namun berkat rahmat dan izin-Nya serta dukungan dari berbagai pihak yang turut membantu baik moril maupun materiel sehingga akhirnya tugas akhir ini *alhamdulillah* dapat terselesaikan. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang setinggi-tingginya dan penghargaan yang tak terhingga kepada Ayahanda **Syamsul Bahri** dan Ibunda tercinta **Widyawati** yang telah menjadi inspirasi, membesarkan dan mendidik penulis dengan sabar dan ikhlas, penuh cinta dan kasih sayang, dan tidak pernah ada habisnya selalu mendoakan yang terbaik untuk penulis, untuk kakak tersayang **Syahrul Awal** beserta istri **Sry Ariati** dan keluarga besar penulis yang selalu mendoakan, memberikan dukungan dan motivasi, serta menjadi penyemangat untuk segera menyelesaikan masa studi penulis.

Penghargaan yang tulus dan penuh hormat juga penulis ucapkan terima kasih yang sebesar-besarnya dan mendoakan semoga Allah *Subhanahu Wa Ta'ala* memberikan balasan yang terbaik kepada:

1. **Ibu Prof. Dr. Dwia Aries Tina Pulubuhu, MA,** selaku Rektor Universitas Hasanuddin beserta seluruh jajarannya.

2. **Bapak Dr. Khaeruddin, M.Sc.** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh jajarannya.
3. **Ibu Dr. Nurtiti Sunusi, S.Si., M.Si.** selaku Ketua Departemen Statistika dan **Bapak A. Kresna Jaya, S.Si., M.Si.** selaku Sekretaris Departemen Statistika, serta segenap dosen pengajar dan staf **Departemen Statistika** yang telah membekali ilmu, meluangkan waktu, tenaga dan pikiran, serta memberi kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika Program Studi Statistika.
4. **Ibu Dr. Anna Islamiyati, S.Si., M.Si.** selaku Pembimbing Utama penulis yang telah ikhlas meluangkan waktu dan pemikirannya untuk memberikan arahan, pengetahuan, motivasi dan bimbingan ditengah kesibukan beliau serta menjadi tempat berkeluh kesah untuk penulis dalam menyelesaikan tugas akhir ini.
5. **Ibu Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.** selaku Pembimbing Pertama penulis yang telah ikhlas meluangkan waktunya ditengah kesibukan untuk memberikan arahan bagi penulis.
6. **Ibu Dr. Erna Tri Herdiani, S.Si., M.Si.** selaku Dosen Penguji sekaligus Penasehat Akademik penulis dan **Bapak Drs. Raupong, M.Si.** selaku Dosen Penguji, terima kasih telah ikhlas meluangkan waktunya ditengah kesibukan untuk memberikan arahan berupa saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini.
7. Sahabat terbaik sejak di bangku sekolah, **Selfiana, Nur Ainun, Nurul Azzahra, Aisyah RM** dan **Gadis Lugina** yang sampai saat ini masih setia mendengarkan curhatan penulis dan senantiasa mendukung penulis dalam melewati berbagai hal selama menjalani proses pendidikan dibangku kuliah.
8. Spesial untuk saudari tercinta penulis **Fakhriyyah Dj. Junus** yang telah menjadi sekamar yang baik, senantiasa setia menemani dan mendukung penulis dalam setiap keadaan, menjadi tempat berbagi cerita dan berkeluh kesah bagi penulis sehingga waktu yang dilalui bersama selama perkuliahan *alhamdulillah* membuat penulis bahagia.
9. Spesial untuk sahabat tercinta penulis, **Riska Rasyid, Miftahul Jannah, Nurul Wahyuni, Nurul Annisa, Nurhidayatullah, Sakinah Oktoni, Munadiah Apriliani, Fitri** dan **Risnawati Azali** yang telah menjadi

saudari-saudari terbaik sejak awal perkuliahan dan senantiasa mendengarkan curhatan, memberikan dorongan, semangat, nasehat dan motivasi dalam setiap keadaan sehingga penulis bisa mendapatkan lebih banyak pelajaran hidup dan *alhamdulillah* tetap bahagia walau hidup dalam perantauan.

10. Teman-teman **Statistika 2017**, terima kasih atas kebersamaan, suka, dan duka selama menjalani pendidikan di Departemen Statistika. Penulis senang bisa mengenal kalian semua, terkhusus **Nurkamalia, Siti Ihza Arsella Kasim, Aqilah Salsabila Rahman, Zakiah Fitri, Izza Annisa Ramadhani, Eva Riyantie,** dan **Anisa Haura SFY**, terima kasih atas setiap cerita cita dan duka, motivasi, nasehat pembelajaran hidup sehingga memberi kesan perkuliahan yang *alhamdulillah* lebih berwarna.
11. Teman-teman **DISKRIT 2017**, terima kasih telah memberikan pelajaran yang berharga dan arti kebersamaan selama ini kepada penulis. Pengalaman yang berharga telah penulis dapatkan dari teman-teman selama berproses.
12. **Keluarga Mahasiswa FMIPA Unhas** terkhusus anggota keluarga **Himatika FMIPA Unhas** dan **Himastat FMIPA Unhas**, terima kasih atas ilmu yang mungkin tidak bisa didapatkan di proses perkuliahan dan telah menjadi keluarga selama penulis kuliah di Universitas Hasanuddin.
13. Kepada seluruh pihak yang tidak dapat penulis sebutkan satu persatu, terima kasih setinggi-tingginya untuk segala dukungan dan partisipasi yang diberikan kepada penulis semoga bernilai ibadah di sisi Allah *Subhanahu Wa Ta'ala.*

Penulis berharap skripsi ini dapat memberikan tambahan pengetahuan baru bagi para pembelajar statistika. Penulis menyadari bahwa dalam penulisan tugas akhir ini masih banyak terdapat kekurangan. Oleh karena itu, dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga dapat bermanfaat bagi pihak-pihak yang berkepentingan. *Aamiin Yaa Rabbal 'Aalamiin.*

Makassar, 27 Oktober 2021

Nur Aprilia Dzulhijjah

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertandatangan di bawah ini:

Nama	: Nur Aprilia Dzulhijjah
NIM	: H051171016
Program Studi	: Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Non eksklusif (*Non-exclusive Royalty- Free Right*)** atas tugas akhir saya yang berjudul:

"Pemodelan Semiparametrik dengan Koefisien Bervariasi pada Data Longitudinal Menggunakan Penaksir *B-Spline*"

beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.
Dibuat di Makassar pada tanggal 27 Oktober 2021

Yang menyatakan

Nur Aprilia Dzulhijjah

ABSTRAK

Model regresi semiparametrik dengan koefisien bervariasi digunakan untuk menjelaskan hubungan antara variabel respon dan variabel prediktor yang memiliki pola parametrik dan sebagian lainnya berpola nonparametrik. Dengan koefisien yang bervariasi, model diasumsikan memiliki koefisien regresi yang berubah atau bervariasi diakibatkan oleh adanya pengaruh variabel lain seperti waktu. Oleh karena itu, data yang digunakan pada penelitian ini adalah data longitudinal yang diterapkan pada data indeks pembangunan manusia di Provinsi Sulawesi Selatan tahun 2010-2019. Adapun variabel komponen parametriknya yaitu rata-rata lama sekolah, harapan lama sekolah dan pengeluaran per kapita, serta variabel komponen nonparametrik yaitu angka harapan hidup berdasarkan waktu pengamatan. Selanjutnya, koefisien yang bervariasi ditinjau berdasarkan variabel komponen nonparametrik sehingga didekati dengan penaksir *B-spline* yang dimodelkan pada orde 1-4 dengan jumlah titik knot 1-2. Hasil yang diperoleh adalah model regresi semiparametrik dengan koefisien bervariasi menggunakan penaksir *B-spline* yang optimal terdapat pada orde kuadratik dengan 2 titik knot yang memiliki nilai GCV minimum sebesar 454,16. Berdasarkan hasil tersebut, ditunjukkan bahwa seluruh komponen parametrik memiliki pengaruh positif yang artinya setiap peningkatan satu satuan per komponen dan yang lainnya dianggap konstan, dapat meningkatkan indeks pembangunan manusia masing-masing sebesar 2,09%, 0,84% dan 0,64%. Untuk komponen nonparametrik, angka harapan hidup mempengaruhi indeks pembangunan manusia cenderung bervariasi pada waktu pengamatan tertentu dan mengalami pola perubahan yang diperkirakan terjadi pada tahun 2012 dan 2017.

Kata kunci: *B-Spline*, Data Longitudinal, GCV, Indeks Pembangunan Manusia, Koefisien yang Bervariasi, Regresi Semiparametrik.

ABSTRACT

A semiparametric varying coefficient regression model is used to explain the relationship between response variables and predictor variables, that have parametric patterns and some of which do not. The model with varying coefficients assumes a regression coefficient that changes or varies due to the influence of other variables such as time. As a consequence, the data used in this study is longitudinal data applied to South Sulawesi Province's human development index data from 2010-2019. The parametric component variables are average length of schooling, expected years of schooling, and expenditure per capita, while the nonparametric component variable is life expectancy at the time of observation. Furthermore, the varying coefficients were analyzed to use nonparametric component variables and approached with the B-spline estimator modeled on the order of 1-4 with the number of knot points 1-2. The result is a semiparametric varying coefficient regression model using a B-spline estimator that is optimal in quadratic order with two knot points and a minimum GCV value of 454.16. Based on this analysis, it is proved that all parametric components have a positive effect, which means that each increase of one unit per component, while the others are held constant, can increase the human development index by 2.09 %, 0.84 %, and 0.64 %, respectively. For nonparametric components, life expectancy affecting the human development index tends to vary at a specific time of observation and experiences a pattern of changes that are expected to occur in 2012 and 2017.

Keywords: B-Spline, Longitudinal Data, GCV, Human Development Index, Varying Coefficient, Semiparametric Regression.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis regresi merupakan metode dalam ilmu statistika yang menganalisis pola hubungan antara variabel respon dan variabel prediktor. Secara umum terdapat tiga model pendekatan dalam analisis regresi, yaitu pendekatan parametrik, nonparametrik, dan semiparametrik (Wahba, 1990). Pendekatan semiparametrik merupakan kombinasi antara pendekatan parametrik dan nonparametrik (Budiantara, 2005). Apabila komponen parametriknya berpola linier, maka regresi semiparametrik disebut regresi linier parsial. Oleh karena itu, estimasi untuk model regresi diperoleh secara bersamaan dengan estimasi fungsi dan estimasi parameter dalam model regresi semiparametrik (Budiantara, 2011).

Model regresi semiparametrik telah dikembangkan menjadi model semiparametrik dengan koefisien bervariasi (*varying coefficient*), tujuannya untuk menjaga interpretabilitas model parametrik dan fleksibilitas model nonparametrik (Fan dkk., 2007). Dengan koefisien yang bervariasi, maka model diasumsikan memiliki koefisien regresi yang tidak bernilai konstan namun terdapat koefisien yang bervariasi diakibatkan oleh adanya pengaruh variabel lain seperti waktu (Hastie dan Tibshirani, 1993). Beberapa penelitian terkait regresi semiparametrik dengan koefisien bervariasi diantaranya, Zhang dkk., (2002) menggunakan regresi semiparametrik polinomial lokal koefisien bervariasi untuk memodelkan pengaruh faktor lingkungan terhadap polusi. Xia dkk., (2004) menggunakan regresi semiparametrik lokal linier koefisien bervariasi untuk memodelkan pengaruh kelembaban pada peredaran darah dan masalah pernapasan tidak langsung melalui nitrogen dioksida. Ahmad dkk., (2005) menggunakan regresi semiparametrik *spline* koefisien bervariasi dalam memodelkan data sensus industri di China pada sektor manufaktur makanan, minuman ringan dan rokok. Penelitian-penelitian tersebut menggunakan data yang berbentuk *cross section* yaitu data yang diamati pada satu titik waktu tertentu, sehingga tidak memiliki kemampuan untuk menjelaskan hubungan dari populasi yang diamati dalam periode waktu yang berbeda. Dengan

demikian, untuk mengatasi hal tersebut model regresi semiparametrik dengan koefisien bervariasi dapat digunakan pada data longitudinal (Nurdini, 2006).

Dalam mengestimasi parameter model regresi semiparametrik yang menjadi masalah adalah terdapat komponen nonparametrik berupa fungsi yang tidak diketahui bentuknya. Salah satu pendekatan nonparametrik yang sering digunakan adalah regresi *spline*. *Spline* merupakan potongan-potongan polinomial yang memiliki sifat tersegmen, sehingga memberikan sifat fleksibel yang lebih baik terhadap karakteristik suatu fungsi atau data. Hal ini terjadi karena dalam *spline* terdapat titik-titik knot, yaitu titik perpaduan bersama yang menunjukkan perubahan pola perilaku data (Eubank, 1988 *dalam* Sari, 2017). Beberapa penaksir *spline* yang telah dikembangkan oleh peneliti diantaranya *spline truncated* (Islamiyati, 2017), *spline smoothing* (Lestari dkk., 2012), *penalized spline* (Zia dkk., 2017; Islamiyati dkk., 2018) dan *B-spline* (Budiantara dkk., 2006).

Regresi semiparametrik dengan penaksir *B-spline* pada data longitudinal telah digunakan dalam beberapa kasus. Leng dkk., (2010) menganalisis faktor-faktor yang mempengaruhi kandungan sel CD4 pada pasien HIV menggunakan regresi semiparametrik *B-spline*. Kim (2010) menyelidiki tren tarif rawat inap untuk pasien gagal jantung kronis di Sistem Kesehatan Universitas Virginia menggunakan regresi semiparametrik *B-spline*. Elmi dkk., (2011) menggunakan regresi semiparametrik *B-spline* dalam memodelkan kurva tenaga kerja wanita yang mencoba melahirkan normal setelah operasi *caesar*. Yang dkk., (2020) membandingkan metode pengobatan kemoterapi intensif dan standar untuk penderita kanker payudara stadium 1 menggunakan regresi semiparametrik *B-spline*. Penelitian-penelitian tersebut tidak mempertimbangkan koefisien bervariasi yang dapat terjadi pada data longitudinal, yang bertujuan untuk memperoleh hubungan antara variabel respon dan variabel prediktor yang diasumsikan linier pada waktu tertentu namun koefisien regresi yang berubah berdasarkan waktu (Liang dkk., 2003; Nurmiati, 2011).

Data longitudinal dapat ditunjukkan pada data indeks pembangunan manusia (IPM) yang diamati pada setiap kabupaten/kota di Provinsi Sulawesi Selatan pada tahun 2010-2019. Menurut Badan Pusat Statistik (2019), IPM disusun berdasarkan tiga komponen dasar yaitu dimensi kesehatan, pendidikan dan standar hidup layak.

Dimensi kesehatan memiliki satu indikator pembentuk yaitu angka harapan hidup, dimensi pendidikan tersusun dari dua indikator yaitu rata-rata lama sekolah dan harapan lama sekolah, dan untuk dimensi standar hidup layak dihitung berdasarkan angka pengeluaran per kapita yang disesuaikan. Selanjutnya, pada tahun 2010-2016 status pembangunan manusia di Provinsi Sulawesi Selatan tergolong dalam kategori sedang, adapun pada tahun 2017-2019 terjadi peningkatan status pembangunan manusia menjadi kategori tinggi, sehingga dapat dikatakan bahwa IPM di Provinsi Sulawesi Selatan cenderung menunjukkan peningkatan setiap tahunnya. IPM di Provinsi Sulawesi Selatan tahun 2010-2019 memiliki karakteristik data yang dapat dimodelkan dengan pendekatan regresi semiparametrik, karena terdapat indikator yang berpola parametrik linier dan terdapat pula indikator yang berpola nonparametrik. Oleh karena itu, dalam penelitian ini penulis akan mengkaji tentang penaksir *B-spline* dalam mengestimasi model semiparametrik dengan koefisien bervariasi pada data longitudinal yang diaplikasikan pada data indeks pembangunan manusia di Provinsi Sulawesi Selatan tahun 2010-2019. Berdasarkan uraian tersebut, penulis mengajukan bahan skripsi dengan judul "Pemodelan Semiparametrik dengan Koefisien Bervariasi pada Data Longitudinal Menggunakan Penaksir *B-spline*".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka diperoleh rumusan masalah sebagai berikut:

1. Bagaimana estimasi model regresi semiparametrik dengan koefisien bervariasi pada data longitudinal menggunakan penaksir *B-spline* yang bersesuaian dengan data indeks pembangunan manusia di Provinsi Sulawesi Selatan tahun 2010-2019?
2. Bagaimana model data indeks pembangunan manusia di Provinsi Sulawesi Selatan tahun 2010-2019 berdasarkan regresi semiparametrik koefisien bervariasi menggunakan penaksir *B-spline*?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan adalah indeks pembangunan manusia berdasarkan 24 kabupaten/kota di Provinsi Sulawesi Selatan tahun 2010-2019 yang dianalisis secara simultan.
2. Pada penelitian ini, variabel komponen nonparametrik yang dimodelkan dengan koefisien yang bervariasi (*varying coefficient*).
3. Pemilihan titik knot optimal menggunkan metode *Generalized Cross Validation* (GCV).
4. Jumlah titik knot yang dilakukan pada penelitian ini adalah satu dan dua titik knot dengan orde yang dibatasi pada orde linier, kuadratik, kubik dan kuartik.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, tujuan penelitian ini adalah sebagai berikut:

1. Mendapatkan estimasi model regresi semiparametrik dengan koefisien bervariasi pada data longitudinal menggunakan penaksir *B-spline* yang bersesuaian dengan data indeks pembangunan manusia di Provinsi Sulawesi Selatan tahun 2010-2019.
2. Mendapatkan model data indeks pembangunan manusia di Provinsi Sulawesi Selatan tahun 2010-2019 berdasarkan regresi semiparametrik dengan koefisien bervariasi menggunakan penaksir *B-spline*.

1.5 Manfaat Penelitian

Berdasarkan rumusan masalah, manfaat penelitian ini adalah sebagai berikut:

1. Memberikan pengetahuan yang lebih khusus kepada penulis tentang pengembangan model regresi semiparametrik pada data longitudinal.
2. Memberikan informasi bagi pemerintah dalam rangka pengambilan kebijakan untuk meningkatkan indeks pembangunan manusia di Provinsi Sulawesi Selatan.

BAB II
TINJAUAN PUSTAKA

2.1 Pendekatan Regresi

2.1.1 Regresi Parametrik

Regresi parametrik merupakan suatu metode statsitika yang digunakan untuk mengetahui pola hubungan antara variabel respon dengan variabel prediktor yang diasumsikan telah diketahui bentuk kurva regresinya. Salah satu bentuk regresi parametrik dapat dinyatakan sebagai model regresi linier berganda yang secara umum dituliskan sebagai berikut (Maksum, 2019):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{kn} + \varepsilon_i, i = 1,2, \dots, n. \qquad (2.1)$$

dengan

y_i : variabel respon pengamatan ke-i

$x_{1i}, x_{2i}, \dots, x_{kn}$: variabel-variabel prediktor pengamatan ke-i

β_0 : intersep dari model

$\beta_1, \beta_2, \dots, \beta_k$: koefisien-koefisien regresi

ε_i : *error* acak pada pengamatan ke-i yang diasumsikan identik, independen, dan berdistribusi $N(0, \sigma^2)$

Persamaan (2.1) dapat diberikan dalam bentuk matriks sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (2.2)$$

Secara lengkap matriks dan vektor-vektor pada Persamaan (2.2) dapat ditulis:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & & x_{k2} \\ & \vdots & & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ dan } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

dengan $\boldsymbol{y}$ menyatakan vektor variabel respon, $\boldsymbol{X}$ menyatakan matriks variabel prediktor, $\boldsymbol{\beta}$ menyatakan vektor dari parameter model dan $\boldsymbol{\varepsilon}$ menyatakan vektor dari *error* (Ruppert dkk., 2003).

2.1.2 Regresi Nonparametrik

Regresi nonparametrik merupakan suatu metode statsitika yang digunakan untuk mengetahui hubungan antara variabel respon dan prediktor yang tidak diketahui bentuk fungsinya, hanya diasumsikan fungsi dipermulus (*smooth*) dalam arti termuat dalam ruang fungsi tertentu, sehingga regresi nonparametrik memiliki fleksibilitas yang tinggi. Secara umum, model regresi nonparametrik adalah sebagai berikut (Eubank, 1988 *dalam* Tupen dan Budiantara, 2011):

$$y_i = f(z_i) + \varepsilon_i, i = 1, 2, \dots, n. \tag{2.3}$$

dengan

y_i : variabel respon pengamatan ke-i

$f(z_i)$: fungsi regresi nonparametrik yang tidak diketahui

z_i : variabel prediktor pengamatan ke-i

ε_i : *error* acak pada pengamatan ke-i yang diasumsikan identik, independen, dan berdistribusi $N(0, \sigma^2)$

Persamaan (2.3) dapat dituliskan dalam bentuk matriks sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z}) + \boldsymbol{\varepsilon} \tag{2.4}$$

Matriks dan vektor pada Persamaan (2.4) dapat ditulis:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{f}(\boldsymbol{z}) = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_n) \end{bmatrix} \text{ dan } \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

2.1.3 Regresi Semiparametrik

Regresi semiparametrik merupakan gabungan dari regresi parametrik dan nonparametrik, sehingga estimasi model semiparametrik ekuivalen dengan estimasi parameter-parameter pada komponen parametrik dan estimasi fungsi pada komponen nonparametrik. Misalkan terdapat data berpasangan (y_i, x_i, z_i), yang hubungan antara y_i, x_i dan z_i diasumsikan mengikuti model regresi semiparametrik pada Persamaan (2.5) (Ruppert dkk., 2003):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + f(z_i) + \varepsilon_i, i = 1,2, \dots, n. \tag{2.5}$$

dengan

y_i : variabel respon pengamatan ke-i

x_i : variabel prediktor pengamatan ke-i komponen parametrik

β : parameter prediktor ke-i untuk komponen parametrik

$f(z_i)$: fungsi regresi nonparametrik yang tidak diketahui

z_i : variabel prediktor pengamatan ke-i komponen nonparametrik

ε_i : *error* acak pengamatan ke-i yang diasumsikan identik, independen, dan berdistribusi $N(0, \sigma^2)$

Persamaan (2.5) dapat ditulis dalam bentuk matriks sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{f}(\boldsymbol{z}) + \boldsymbol{\varepsilon} \tag{2.6}$$

dengan $\boldsymbol{y}$ adalah vektor variabel respon berukuran $n \times 1$, $\boldsymbol{X}$ adalah matriks variabel prediktor untuk komponen parametrik berukuran $n \times (k+1)$, $\boldsymbol{z}$ memuat variabel prediktor komponen nonparametrik, $\boldsymbol{\beta}$ adalah vektor parameter regresi berukuran $(k+1) \times 1$, $\boldsymbol{f}$ adalah vektor dari fungsi regresi yang tidak diketahui, $\boldsymbol{\varepsilon}$ adalah vektor *error* acak yang berdistribusi normal dengan mean nol dan variansi $\sigma^2\boldsymbol{I}$ (Salam, 2013).

2.2 *B-Spline* pada Regresi Nonparametrik

Berdasarkan Persamaan (2.3), jika f didekati dengan penaksir *B-spline* maka secara umum dapat ditulis dalam bentuk (Budiantara dkk., 2006):

$$f(z_i) = \sum_{l=1}^{m+K} \gamma_l \, B_{l-m,m}(z_i) \tag{2.7}$$

dengan $i = 1,2,\dots,n;\ l = 1,2,\dots,m+K$ dan $B_{l-m,m}(z_i)$ adalah basis *B-spline*.

Adapun cara membangun fungsi *B-spline* orde m dengan titik-titik $a < u_1 < u_2 < \cdots < u_K < b$ adalah dengan mendefinisikan titik knot tambahan sebanyak $2m$, yaitu $u_{-(m-1)}, u_{-(m-2)}, \dots, u_{-1}, u_0, u_1, u_2, \dots, u_K, u_{K+1}, \dots, u_{K+m}$ dengan $u_{-(m-1)} = u_{-(m-2)} = \cdots = u_{-1} = u_0 = a$ dan $u_{(K+1)} = u_{(K+2)} = \cdots = u_{K+m} = b$, dengan a diperoleh dari nilai minimum z dan b diperoleh dari nilai maksimum z.

Menurut (Eubank, 1999), basis fungsi *B-spline* pada orde m dengan titik-titik knot pada $u_1, u_2, \dots, u_K$ dapat didefinisikan secara rekursif pada Persamaan (2.8):

$$B_{l,m}(z_i) = \frac{z - u_l}{u_{l+m-1} - u_l} B_{l,m-1}(z_i) + \frac{u_{l+m} - z}{u_{l+m} - u_{l+1}} B_{l+1,m-1}(z_i) \tag{2.8}$$

dengan $l = -(m-1), \dots, K$, dan

$$B_{l,1}(z_i) = \begin{cases} 1, & \text{jika } u_l < z_i \leq u_{l+1} \\ 0, & \textit{untuk yang lainnya} \end{cases}$$

Untuk $m = 2$ memberikan fungsi *B-spline* linier, $m = 3$ memberikan fungsi *B-spline* kuadratik dan $m = 4$ memberikan fungsi *B-spline* kubik (Budiantara dkk., 2006). Adapun untuk mengestimasi γ pada Persamaan (2.7), dapat diuraikan terlebih dahulu sebagai berikut:

$$f(z_i) = \gamma_1 B_{1-m,m}(z_i) + \gamma_2 B_{2-m,m}(z_i) + \cdots + \gamma_{m+K} B_{(m+K)-m,m}(z_i)$$

$$f(z_i) = \gamma_1 B_{-(m-1),m}(z_i) + \gamma_2 B_{-(m-2),m}(z_i) + \cdots + \gamma_{m+K} B_{K,m}(z_i) \tag{2.9}$$

Jika Persamaan (2.9) disubtitusikan ke Persamaan (2.3) maka dapat ditulis:

$$y_i = \gamma_1 B_{-(m-1),m}(z_i) + \gamma_2 B_{-(m-2),m}(z_i) + \cdots + \gamma_{m+K} B_{K,m}(z_i) + \varepsilon_i \tag{2.10}$$

Berdasarkan uraian tersebut dapat ditulis dalam bentuk matriks sebagai berikut:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} B_{-(m-1),m}(z_1) & B_{-(m-2),m}(z_1) & \dots & B_{K,m}(z_1) \\ B_{-(m-1),m}(z_2) & B_{-(m-2),m}(z_2) & \dots & B_{K,m}(z_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_{-(m-1),m}(z_n) & B_{-(m-2),m}(z_n) & \dots & B_{K,m}(z_n) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{m+K} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.11}$$

Persamaan (2.11) dapat dituliskan dalam notasi matriks sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{\lambda})\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2.12}$$

dengan $\boldsymbol{B}(\boldsymbol{\lambda})$ adalah sebuah matriks berukuran $n \times (m + K)$ (Botella dan Shariff, 2003).

2.3 Model Koefisien Bervariasi

Model koefisien bervariasi (*varying coefficient model*) merupakan suatu model yang dibentuk untuk meningkatkan fleksibilitas model regresi linier, untuk menggantikan sebagian atau keseluruhan linier dan fungsi parametrik dari variabel prediktor dengan pemulusan fungsi nonparametrik. Model ini muncul ketika ingin

diketahui bagaimana koefisien regresi bisa berubah pada kelompok yang berbeda yang telah dikarakterisasikan melalui kovariat tertentu (Fan dan Zhang, 1999). Perbedaan antara model regresi linier klasik dengan model koefisien bervariasi adalah koefisien regresi tidak lagi bernilai konstan, melainkan koefisien yang bervariasi diakibatkan oleh adanya pengaruh variabel lain seperti waktu (Hastie dan Tibshirani, 1993).

Misalkan terdapat satu set observasi $\big(y(T), \big(x_1(T), \dots, x_k(T)\big), T\big)$ dengan T merupakan variabel waktu, $y(T)$ adalah variabel respon pada waktu T, dan $\big(x_1(T),, \dots, x_k(T)\big)$ merupakan vektor kovariat pada waktu T. Diasumsikan bahwa semua pengukuran bersifat independen untuk setiap subjek yang berbeda, namun pengukuran pada setiap titik waktu yang berbeda dalam setiap subjek dapat dikorelasikan. Untuk $t \in T$, maka bentuk model koefisien bervariasi adalah sebagai berikut (Hastie dan Tibshirani, 1993; Yunisa, 2020):

$$y(t) = \alpha_0(t) + x_1(t)\,\alpha_1(t) + \cdots + x_k(t)\,\alpha_k(t) + \varepsilon(t)$$

atau dalam bentuk matriks

$$\boldsymbol{y}(t) = \boldsymbol{X}(t)\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t) \tag{2.13}$$

dengan $\boldsymbol{y}(t)$ respon pada waktu t, $\boldsymbol{X}(t) = \big(x_1(t), \dots, x_k(t)\big)'$ adalah kovariat pada waktu t, $\boldsymbol{\alpha}(t) = \big(\alpha_0(t), \alpha_1(t), \dots, \alpha_k(t)\big)'$ adalah fungsi koefisien regresi yang tidak diketahui pada waktu t.

Jika diberikan suatu set sampel (y_{ij}, x_{ij}, t_{ij}) untuk $i = 1, \dots, n$, $j = 1, \dots, N_i$ dan t_{ij} adalah waktu ke-j untuk subjek ke-i dengan N_i menyatakan banyaknya pengukuran berulang dari subjek ke-i. Maka, bentuk model koefisien bervariasi secara nonparametrik adalah sebagai berikut:

$$y(t_{ij}) = \alpha_0(t_{ij}) + x_1(t_{ij})\,\alpha_1(t_{ij}) + \cdots + x_k(t_{ij})\,\alpha_k(t_{ij}) + \varepsilon(t_{ij}) \tag{2.14}$$

atau dalam bentuk matriks

$$\boldsymbol{y}(t_{ij}) = \boldsymbol{X}(t_{ij})\boldsymbol{\alpha}(t_{ij}) + \boldsymbol{\varepsilon}(t_{ij}) \tag{2.15}$$

Berdasarkan Persamaan (2.15), terlihat bahwa $\boldsymbol{\alpha}(t_{ij})$ merupakan intersep bervariasi yang mengukur hubungan langsung antara variabel t_{ij} dan variabel

respon secara nonparametrik. Variabel t_{ij} mengubah koefisien dari x_{ij} melalui fungsi parameter α. Ketergantungan α terhadap t_{ij} menunjukkan adanya interaksi khusus antara t_{ij} dan x_{ij}, yang berarti bahwa pengaruh dari x_{ij} terhadap y_{ij} tidak akan konstan melainkan akan bervariasi dengan mulus dengan variabel waktu (t_{ij}) (Durlauf dan Blume, 2010).

2.4 Model Semiparametrik Koefisien Bervariasi

Fan dkk., (2007) memberikan model semiparametrik dengan koefisien bervariasi sebagai berikut:

$$y(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{z}(t)\boldsymbol{\alpha}(t) + \varepsilon(t) \tag{2.16}$$

dengan $y(t)$ menyatakan variabel respon, $\boldsymbol{X}(t)$ menyatakan variabel prediktor komponen parametrik dan $\boldsymbol{z}(t)$ menyatakan matriks variabel prediktor komponen nonparametrik, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)'$ adalah parameter yang tidak diketahui, $\boldsymbol{\alpha}(t) = (\alpha_0(t), \alpha_1(t), \alpha_2(t), \dots, \alpha_k(t))'$ adalah parameter fungsi diperhalus (*smooth*) yang tidak diketahui, $\varepsilon(t)$ adalah *error* dan $E\{\varepsilon(t)|\boldsymbol{X}(t), \boldsymbol{z}(t)\} = 0$.

Adapun regresi semiparametrik dengan koefisien bervariasi dapat digunakan pada data longitudinal, yang bertujuan untuk memperoleh hubungan antara variabel respon dan variabel prediktor yang diasumsikan linier pada waktu tertentu namun koefisien regresi yang berubah berdasarkan waktu (Nurmiati, 2011). Misalkan sampel acak dari model (2.15) terdiri dari *n* subjek. Untuk subjek ke-i, $i = 1, \dots, n$, variabel respon $\big(y(t)\big)$ dan kovariat $\{x_i(t), z_i(t)\}$ dikumpulkan pada titik waktu $t = t_{ij}$, yang mana $j = 1, \dots, N_i$ dengan N_i adalah jumlah total pengamatan untuk subjek ke-i. Oleh karena itu, pada data longitudinal, model semiparametrik dengan koefisien bervariasi dapat dinyatakan sebagai berikut (Liang dkk., 2003):

$$y_{ij}(t_{ij}) = \boldsymbol{X}_{ij}(t_{ij})\boldsymbol{\beta} + z_{ij}(t_{ij})f(t_{ij}) + \varepsilon_{ij}(t_{ij}) \tag{2.17}$$

dengan α yang merupakan parameter fungsi diperhalus pada Persamaan (2.16) dapat dinyatakan sebagai f pada Persamaan (2.17) sebagai fungsi koefisien regresi yang tidak diketahui untuk komponen nonparametrik (Fan dkk., 2007).

Keterangan Persamaan (2.17):

y_{ij} : variabel respon pada pengamatan ke-*j* dari subjek ke-*i*

$\boldsymbol{X}_{ij}$: variabel prediktor komponen parametrik pada pengamatan ke-*j* dari subjek ke-*i*

z_{ij} : variabel prediktor komponen nonparametrik pada pengamatan ke-*j* dari subjek ke-*i*

t_{ij} : waktu pengamatan ke-*j* dari subjek ke-*i*

$\boldsymbol{\beta}$: parameter yang tidak diketahui untuk komponen parametrik

f : fungsi koefisien regresi yang tidak diketahui untuk komponen nonparametrik

ε_{ij} : *error* pada pengamatan ke-*j* dari subjek ke-*i*

2.5 Metode *Least Square*

Misalkan parameter $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ pada Persamaan (2.1) tidak diketahui, maka parameter-parameter tersebut perlu diestimasi. Estimasi parameter yang biasa digunakan adalah metode estimasi *least square*, yaitu dengan meminimumkan jumlah kuadrat *error*. Persamaannya dapat ditunjukan sebagai berikut:

$$\sum \varepsilon_i^2 = \sum (y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{kn})^2 \qquad (2.18)$$

dengan $\sum \varepsilon_i^2$ adalah jumlah kuadrat *error*, yang dalam notasi matriks dapat dituliskan sebagai berikut:

$$\boldsymbol{\varepsilon}_i' \boldsymbol{\varepsilon}_i = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n] \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 = \sum \varepsilon_i^2 \qquad (2.19)$$

Berdasarkan (2.19), maka diperoleh

$$\begin{aligned} \boldsymbol{J} = \boldsymbol{\varepsilon}_i' \boldsymbol{\varepsilon}_i &= (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \\ &= (\boldsymbol{y}' - \boldsymbol{\beta}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{2\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \end{aligned} \qquad (2.20)$$

Kemudian, $\boldsymbol{J}$ didiferensialkan terhadap $\boldsymbol{\beta}$ untuk meminimumkan jumlah kuadrat *error* dan hasilnya disamakan dengan nol sebagai berikut:

$$\left.\frac{\partial(\boldsymbol{J})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \frac{\partial\left(\boldsymbol{y}'\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{\partial\widehat{\boldsymbol{\beta}}}$$

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$$

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{2.21}$$

selanjutnya, untuk menyelesaikan Persamaan (2.21), kalikan kedua ruas dengan invers dari $(\boldsymbol{X}'\boldsymbol{X})$ sehingga diperoleh penaksir kuadrat terkecil dari $\widehat{\boldsymbol{\beta}}$ berbentuk

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\boldsymbol{I}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{2.22}$$

dengan $\widehat{\boldsymbol{\beta}}$ menyatakan vektor berukuran $(k+1)\times 1$ dari parameter yang akan diestimasi, k menyatakan banyaknya variabel prediktor yang digunakan, $\boldsymbol{X}$ menyatakan matriks variabel prediktor, $\boldsymbol{X}'$ menyatakan transpos matriks $\boldsymbol{X}$, dan $\boldsymbol{y}$ menyatakan vektor variabel respon.

2.6 Pemilihan Titik Knot Optimal

Pemilihan titik knot $u_1, u_2, \dots, u_K$ yang optimal sangat penting dalam penggunaan regresi *B-spline*. Titik knot merupakan titik perpaduan bersama yang menunjukkan perubahan pola perilaku data. Salah satu metode pemilihan titik knot yang optimal adalah *Generalized Cross Validation* (GCV). Model regresi *B-spline* yang sesuai berkaitan dengan titik knot optimal yang dapat diperoleh dari nilai GCV minimum. Adapun fungsi GCV didefinisikan pada Persamaan (2.23) (Eubank, 1988 *dalam* Sari, 2017):

$$GCV(u_1, u_2, \dots, u_K) = \frac{MSE(u_1, u_2, \dots, u_K)}{\left(\frac{1}{n}trace[\boldsymbol{I} - A(u_1, u_2, \dots, u_K)]\right)^2} \tag{2.23}$$

dengan

n : jumlah data

$\boldsymbol{I}$: matriks identitas

$u_1, \dots, u_K$: titik knot

MSE : $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(z_i)\right)^2$

$A(u_1, \dots, u_K)$: $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$

dengan

$$\boldsymbol{A} = \begin{bmatrix} 1 & z_1 & \dots & z_1^{m-1} & (z_1-u_1)^{m-1} & \dots & (z_1-u_K)^{m-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & \dots & z_n^{m-1} & (z_n-u_1)^{m-1} & \dots & (z_n-u_K)^{m-1} \end{bmatrix}$$

Nilai GCV digunakan karena aspek perhitungannya lebih sederhana dan cukup efisien. Selain itu, kriteria model regresi yang umumnya dipakai masih tetap dijadikan acuan pemilihan model regresi *B-spline* terbaik. Model regresi *B-Spline* terbaik adalah model yang mampu menjelaskan pola hubungan antara variabel respon dengan variabel prediktor (Yuniartika dkk., 2013).

2.7 Data Longitudinal

Data longitudinal merupakan penggabungan antara data *cross section* dan *time series* (Harlan, 2018). Ada beberapa keuntungan dari data longitudinal dibandingkan dengan data *cross section*. Pertama, untuk memperoleh kekuatan uji statistik yang sama, data longitudinal membutuhkan subjek yang lebih sedikit. Kedua, dengan jumlah subjek yang sama, hasil pengukuran *error* menghasilkan penaksir efek perlakuan yang lebih efisien dari data *cross section*. Ketiga, data longitudinal mampu menyediakan informasi tentang perubahan individu, sedangkan data *cross section* tidak (Laome, 2009).

Misalkan t_{ij} menyediakan pengamatan pada waktu ke-*j* dari subjek ke-*i* dan y_{ij} menyatakan variabel respon pada pada waktu ke-*j* dari subjek ke-*i* dan x_{ijr} menyatakan variabel prediktor ke-*r* yang diamati pada waktu ke-*j* dari subjek ke-*i*, maka data longitudinal dapat dinyatakan pada Persamaan (2.24) sebagai berikut:

$$\left\{\left(t_{ij}, y_{ij}, x_{ijr}\right), i = 1,2, \dots, n;\ j = 1,2, \dots, N_i;\ r = 1,2, \dots, k\right\} \qquad (2.24)$$

dengan n menyatakan banyaknya subjek, N_i menyatakan banyaknya waktu pengamatan dari subjek ke-*i* dan k menyatakan banyaknya variabel prediktor (Nurmiati, 2011). Selanjutnya, struktur data longitudinal dapat ditunjukkan pada Tabel 2.1.

Tabel 2.1. Struktur Data Longitudinal

Subjek (i)	Waktu Pengamatan ke-i (t_{ij})	y_{ij}	x_{ij1}	x_{ij2}	$\cdots$	x_{ijk}
1	t_{11}	y_{11}	x_{111}	x_{112}	$\cdots$	x_{11k}
	t_{12}	y_{12}	x_{121}	x_{122}	$\cdots$	x_{12k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	t_{1N_1}	y_{1N_1}	x_{1N_11}	x_{1N_12}	$\cdots$	x_{1N_1k}
2	t_{21}	y_{21}	x_{211}	x_{212}	$\cdots$	x_{21k}
	t_{22}	y_{22}	x_{221}	x_{222}	$\cdots$	x_{22k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	t_{2N_2}	y_{2N_2}	x_{2N_21}	x_{2N_22}	$\cdots$	x_{2N_2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
n	t_{n1}	y_{n1}	x_{n11}	x_{n12}	$\cdots$	x_{n1k}
	t_{n2}	y_{n2}	x_{n21}	x_{n22}	$\cdots$	x_{n2k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	t_{nN_n}	y_{nN_n}	x_{nN_n1}	x_{nN_n2}	$\cdots$	x_{nN_nk}

2.8 Indeks Pembangunan Manusia

Indeks pembangunan manusia (IPM) merupakan indikator penting untuk mengukur keberhasilan dalam upaya membangun kualitas hidup masyarakat, sehingga dapat menentukan peringkat atau level pembangunan suatu wilayah atau negara. IPM pertama kali diperkenalkan oleh *United Nations Development Programme* (UNDP) pada tahun 1990 dan dipublikasikan secara berkala dalam laporan tahunan *Human Development Report* (HDR). Terdapat tiga komponen dasar penyusun indeks pembangunan manusia, yaitu dimensi kesehatan, pendidikan, dan standar hidup layak. Setiap komponen dasar IPM distandarisasi dengan nilai minimum dan maksimum sebelum digunakan untuk menghitung IPM dengan rumus sebagai berikut (BPS, 2019):

1. Dimensi Kesehatan

Untuk mengukur dimensi kesehatan digunakan angka harapan hidup dengan rumus sebagai berikut:

$$I_{kesehatan} = \frac{AHH - AHH_{min}}{AHH_{maks} - AHH_{min}}$$

Angka harapan hidup (AHH) didefinisikan sebagai rata-rata perkiraan banyak tahun yang dapat ditempuh oleh seseorang sejak lahir. AHH mencerminkan derajat kesehatan suatu masyarakat. Adapun untuk menghitung indeks harapan hidup digunakan nilai maksimum dan minimum umur harapan hidup yang sesuai standar UNDP, yaitu 85 tahun untuk nilai maksimum dan 20 tahun untuk nilai minimum. Jika tingkat harapan hidup meningkat, maka ini juga mencerminkan membaiknya kondisi kesehatan penduduk secara umum dan tentunya juga merefleksikan membaiknya kondisi ekonomi penduduk.

2. Dimensi Pendidikan

Untuk mengukur dimensi pendidikan digunakan indikator rata-rata lama sekolah dan harapan lama sekolah yang merefleksikan dari kemampuan masyarakat untuk mengakses pendidikan, khususnya pendidikan berkualitas baik yang sangat diperlukan dalam kehidupan produktif masyarakat modern. Adapun rumus untuk menghitung dimensi pendidikan adalah sebagai berikut:

a. Indikator rata-rata lama sekolah

$$I_{RLS} = \frac{RLS - RLS_{min}}{RLS_{maks} - RLS_{min}}$$

b. Indikator harapan lama sekolah

$$I_{HLS} = \frac{HLS - HLS_{min}}{HLS_{maks} - HLS_{min}}$$

Sehingga dimensi pendidikan dapat dihitung dengan rumus:

$$I_{pendidikan} = \frac{I_{RLS} + I_{HLS}}{2}$$

Rata-rata lama sekolah (RLS) menggambarkan jumlah tahun yang digunakan oleh penduduk usia 25 tahun ke atas dalam menjalani pendidikan formal, sedangkan harapan lama sekolah merupakan lamanya sekolah (dalam tahun) yang diharapkan akan dirasakan oleh anak yang berumur 7 tahun. Perhitungan indeks pendidikan didasarkan pada rata-rata indeks RLS dan indeks HLS dengan bobot yang sama. Adapun dalam perhitungan indeks RLS dan HLS digunakan batasan nilai maksimum dan nilai minimum yang sama dengan standar UNDP yaitu untuk nilai maksimum dan minimum untuk RLS masing-masing sebesar 15 dan 0 tahun, sedangkan untuk HLS masing-masing sebesar 18 dan 0 tahun.

3. Dimensi Standar Hidup Layak

Untuk mengukur standar hidup layak digunakan indikator pengeluaran per kapita dengan rumus sebagai berikut:

$$I_{pengeluaran} = \frac{\ln(pengeluaran) - \ln(pengeluaran_{min})}{\ln(pengeluaran_{maks}) - \ln(pengeluaran_{min})}$$

Pengeluaran per kapita (PPP) adalah biaya yang dikeluarkan untuk konsumsi semua anggota rumah tangga selama sebulan dibagi dengan banyaknya anggota rumah tangga. Perubahan pendapatan seseorang akan berpengaruh pada pergeseran pola pengeluaran. Pola pengeluaran dapat dipakai sebagai salah satu alat untuk mengukur tingkat kesejahteraan penduduk. Data pengeluaran dapat mengungkap tentang pola konsumsi rumah tangga secara umum menggunakan indikator proporsi pengeluaran untuk makanan dan non makanan. Komposisi pengeluaran rumah tangga dapat dijadikan ukuran untuk menilai tingkat kesejahteraan ekonomi penduduk, makin rendah persentase pengeluaran untuk makanan terhadap total pengeluaran makin membaik tingkat kesejahteraan.

Berdasarkan uraian tersebut, IPM dapat dihitung dengan rumus sebagai berikut:

$$IPM = \sqrt[3]{I_{kesehatan} + I_{pendidikan} + I_{pengeluaran}}$$

Dalam upaya untuk membandingkan antar wilayah, dibentuklah klasifikasi IPM. Pengklasifikasian pembangunan manusia bertujuan untuk mengorganisasikan wilayah-wilayah menjadi kelompok-kelompok yang sama dalam hal pembangunan manusia. Menurut Badan Pusat Statistik, capaian IPM diklasifikasikan menjadi beberapa kategori yang disajikan dalam Tabel 2.2 berikut:

Tabel 2.2 Klasifikasi Indeks Pembangunan Manusia (IPM)

Klasifikasi	Capaian IPM
Sangat Tinggi	IPM ≥ 80
Tinggi	$70 \leq$ IPM < 80
Sedang	$60 \leq$ IPM < 70
Rendah	IPM < 60

(Sumber: BPS, 2019)