SKRIPSI

ESTIMASI MODEL REGRESI NONPARAMETRIK POISSON DIPERUMUM MENGGUNAKAN ESTIMATOR *SPLINE TRUNCATED*

Disusun dan diajukan oleh

NURUL WAHYUNI

H051171003

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2021

ESTIMASI MODEL REGRESI NONPARAMETRIK POISSON DIPERUMUM MENGGUNAKAN ESTIMATOR *SPLINE TRUNCATED*

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

NURUL WAHYUNI

H051171003

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

LEMBAR PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini :

Nama	: Nurul Wahyuni
NIM	: H051171003
Program Studi	: Statistika
Jenjang	: Sarjana (S1)

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Estimasi Model Regresi Nonparametrik Poisson Diperumum Menggunakan Estimator *Spline Truncated*

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain bahwa skripsi yang saya tulis ini benar merupakan hasil karya saya sendiri.

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, 11 Februari 2021

METERAI TEMPEL
6F9DBAHF437258943
6000
ENAM RIBU RUPIAH

<u>NURUL WAHYUNI</u>

NIM. H051171003

ESTIMASI MODEL REGRESI NONPARAMETRIK POISSON DIPERUMUM MENGGUNAKAN ESTIMATOR *SPLINE TRUNCATED*

Disetujui oleh:

Pembimbing Utama

Dr. Anna Islamiyati, S. Si., M. Si.
NIP. 19770808 200501 2 002

Pembimbing Pertama

Dra. Nasrah Sirajang, M.Si.
NIP. 19650519 1999303 2 002

Ketua Departemen Statistika,

Dr. Nurtiti Sunusi, S.Si., M.Si.
NIP. 19720117 199703 2 002

Pada tanggal: 11 Februari 2021

LEMBAR PENGESAHAN

ESTIMASI MODEL REGRESI NONPARAMETRIK POISSON DIPERUMUM MENGGUNAKAN ESTIMATOR *SPLINE TRUNCATED*

Disusun dan diajukan oleh

NURUL WAHYUNI

H051171003

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

pada tanggal 11 Februari 2021

dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama	**Pembimbing Pertama**
Dr. Anna Islamiyati, S. Si., M. Si.	**Dra. Nasrah Sirajang, M.Si.**
NIP. 19770808 200501 2 002	**NIP. 19650519 1999303 2 002**

Ketua Departemen Statistika,

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

NIP. 19720117 199703 2002

KATA PENGANTAR

Puji syukur kepada Allah *Subhanahu Wa Ta'ala* atas segala limpahan rahmat, nikmat dan hidayah-Nya, serta shalawat dan salam semoga senantiasa tercurahkan kepada Nabi yang paling dimuliakan, pemimpin orang-orang bertakwa yakni Rasulullah *Shallallahu Alaihi Wasallam* dan kepada para keluarga serta sahabat beliau yang senantiasa kita rindukan perjumpaan dengannya. Alhamdulillah, semua kemudahan yang penulis dapatkan tidak lepas dari pertolongan Allah dan doa dari orang-orang yang tulus, akhirnya skripsi dengan judul **"Estimasi Model Regresi Nonparametrik Poisson Diperumum Menggunakan Estimator *Spline Truncated*"** yang disusun sebagai salah satu syarat akademik untuk meraih gelar sarjana pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin ini dapat dirampungkan. Penulis berharap skripsi ini dapat memberikan tambahan pengetahuan baru bagi para pembelajar statistika.

Penulis menyadari bahwa penyelesaian skripsi ini dengan segala keterbatasan kemampuan dan pengetahuan dapat melewati segala hambatan dan masalah berkat bantuan dan dorongan dari berbagai pihak. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang tak terhingga serta penghargaan yang setinggi-tingginya untuk Ibunda **Rahmia, S. Ag., M. Pd.** dan kakak tersayang **Muhammad Nur Rahman, S. Pd.** yang tak kenal lelah mendoakan, memberikan dukungan, dan selalu melimpahkan cinta dan kasih sayangnya kepada penulis sehingga mereka menjadi motivasi terbesar penulis dalam menyelesaikan tugas akhir ini. Serta untuk keluarga besar penulis, terima kasih atas doa dan dukungannya selama ini.

Penghargaan yang tulus dan ucapan terima kasih dengan penuh keikhlasan juga penulis ucapkan kepada:

1. Ibu **Prof. Dr. Dwia Aries Tina Pulubuhu, M.A,** selaku Rektor Universitas Hasanuddin beserta seluruh jajarannya.
2. Bapak **Dr. Eng. Amiruddin,** 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, segenap dosen pengajar dan staf Departemen Statistika yang telah memberikan

ilmu dan pengetahuan serta bantuan-bantuan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika.

4. Ibu **Dr. Anna Islamiyati, S. Si., M. Si.** selaku dosen pembimbing utama yang dengan sabar, tulus dan ikhlas meluangkan begitu banyak waktu di tengah berbagai kesibukan dan prioritasnya untuk membimbing dan memberikan masukan serta motivasi kepada penulis mulai dari awal hingga selesainya penulisan tugas akhir ini dan Ibu **Dra. Nasrah Sirajang, M. Si.** selaku dosen pembimbing pertama yang telah meluangkan waktunya untuk memberikan masukan dan motivasi.
5. Bapak **Andi Kresna Jaya, S. Si., M. Si.** selaku dosen penguji sekaligus pembimbing akademik penulis dan Bapak **Dr. Nirwan Ilyas, M. Si.** selaku dosen penguji terima kasih telah meluangkan waktu dan memberikan segala saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini.
6. Teman-teman **Statistika 2017**, terkhusus **Novilia Jao, S. Si.** yang merupakan sosok guru bagi penulis di masa perkuliahan serta menjadi tempat bertukar pikiran selama proses penyelesaian tugas akhir, terima kasih atas ilmu, kebersamaan, suka dan duka dalam berjuang menjalani pendidikan di Departemen Statistika.
7. Spesial untuk sahabat tercinta penulis, **Riska Rasyid, Fakhriyyah Dj Junus, Nurul Annisa, Miftahul Jannah, Nurhidayatullah, Sakinah Oktoni, Nur Aprilia Dzulhijjah, Fitri, Munadiah Apriliani** dan **Risnawati Azali** yang telah menjadi sahabat terbaik sejak awal perkuliahan dan senantiasa mendengarkan curhatan, memberikan dorongan, semangat, dan motivasi dalam setiap keadaan sehingga penulis bisa mendapatkan lebih banyak pelajaran hidup. Serta sahabat terbaik sejak di bangku SMA, **Alifiya Hidayani, A.P.A. Pj., Lathifa Mardiyyah Kamaruddin SL, S. KG., Firmasari Zahlan Rohimahallah, Nazhiifa Bungawali, Laila Syukur, Intan Yu'tika Sani, Putri Dwi Wulandari**, dan **Excobar Arman** yang jarang bertemu tapi selalu memberikan dukungan dan motivasi kepada penulis.
8. Keluarga besar **DISKRIT 2017**, terima kasih untuk cerita sekaligus kenangan selama proses yang telah dilalui. Banyak pengalaman dan pelajaran berharga

yang penulis dapatkan dari kalian. Tetap **SATUKAN, ERATKAN, KUATKAN.**

9. Kakak-kakak, teman-teman, dan adik-adik anggota **Keluarga Mahasiswa FMIPA UNHAS,** terkhusus anggota keluarga **HIMASTAT FMIPA UNHAS** dan **HIMATIKA FMIPA UNHAS**, terima kasih atas ilmu yang mungkin tidak bisa didapatkan di proses perkuliahan. Penulis merasa bangga menjadi salah satu bagian dari keluarga ini.
10. Seluruh teman-teman **KKN Tematik Gelombang 104 Posko Kecamatan Tamalate 3**, terkhusus **Afiifah, Ulfah Ainun Humairaa, Safira Muliani,** dan **Nur Mahdiyah** yang telah menjadi sahabat sekaligus keluarga selama proses KKN berlangsung hingga saat ini. Terima kasih atas segala dukungan dan motivasi yang telah diberikan kepada penulis.
11. Kepada seluruh pihak yang tidak dapat penulis sebutkan satu persatu, terima kasih untuk segala dukungan dan partisipasi yang diberikan kepada penulis, semoga bernilai ibadah di sisi Allah *Subhanahu Wa Ta'ala*.

Penulis menyadari bahwa dalam penulisan skripsi ini masih banyak terdapat kekurangan. Oleh karena itu, dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga skripsi ini dapat bermanfaat bagi pihak-pihak yang berkepentingan terutama mahasiswa Statistika.

Makassar, 11 Februari 2021

Nurul Wahyuni

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

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

Nama	: Nurul Wahyuni
NIM	: H051171003
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 karya ilmiah saya yang berjudul:

"Estimasi Model Regresi Nonparametrik Poisson Diperumum Menggunakan Estimator *Spline Truncated*"

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 mempublikasikan 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 11 Februari 2021

Yang menyatakan,

Nurul Wahyuni

ABSTRAK

Regresi Poisson digunakan untuk menjelaskan hubungan antara variabel respon yang berupa data diskrit dan berdistribusi Poisson dengan variabel prediktor. Regresi Poisson mengasumsikan nilai varians dan rata-rata dari variabel respon mempunyai nilai yang sama (equidispersi). Akan tetapi, pada kenyataannya asumsi tersebut seringkali tidak terpenuhi karena ditemukan nilai varians lebih besar daripada nilai rata-ratanya yang disebut overdispersi. Masalah overdispersi dapat diatasi menggunakan model regresi Poisson diperumum. Namun, tidak semua data memiliki pola yang beraturan sehingga digunakan model regresi nonparametrik Poisson diperumum dengan estimator *spline truncated*. Pada penelitian ini, estimasi parameter dilakukan menggunakan metode *maximum likelihood estimation* dengan algoritma Fisher *scoring*. Selanjutnya, model regresi nonparametrik Poisson diperumum menggunakan estimator *spline truncated* diterapkan pada data jumlah kematian bayi di Provinsi Sulawesi Selatan tahun 2017. Pemilihan titik knot optimal menggunakan metode *Generalized Cross Validation* (GCV). Hasil yang diperoleh adalah model terbaik terbentuk pada model regresi nonparametrik Poisson diperumum menggunakan *spline* linier multiprediktor dengan satu titik knot.

Kata kunci: Jumlah Kematian Bayi, Overdispersi, Regresi Nonparametrik Poisson Diperumum, *Spline Truncated*, GCV.

ABSTRACT

Poisson regression is used to explain the relation between response variables in the form of discrete data and Poisson distributed with predictor variables. Poisson regression assumes the variance value and the mean of the response variables have the same value (equidispersion). However, in reality this assumption is often not fulfilled because it is found that the variance value is greater than the mean which is called overdispersion. The overdispersion problem can be solved by using a generalized Poisson regression model. However, not all data have an orderly pattern so that the generalized Poisson nonparametric regression model is used with the truncated spline estimator. In this study, parameter estimation was carried out by using the maximum likelihood estimation method with the Fisher scoring algorithm. Furthermore, the generalized Poisson nonparametric regression model with the spline truncated estimator is applied to the data on the number of infants mortality in South Sulawesi Province in 2017. The selection of optimal knot points uses the Generalized Cross Validation (GCV) method. The results showed that the best model is formed by the generelized Poisson nonparametric regression model using a multipredictor linear spline with one knot point.

Keywords: Number of Infants Mortality, Overdispersion, Generalized Poisson Nonparametric Regression, Spline Truncated, GCV.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis regresi adalah analisis yang digunakan untuk menjelaskan hubungan antara variabel respon dan variabel prediktor, yang mana umumnya variabel respon yang digunakan berupa data kontinu. Namun, dalam aplikasinya banyak ditemukan penelitian yang menggunakan variabel respon berupa data diskrit. Seperti penggunaan analisis regresi dalam menganalisis faktor-faktor yang mempengaruhi jumlah kematian bayi, yang mana data tersebut adalah data jumlahan sehingga tergolong ke dalam data diskrit. Long (1997) mengatakan apabila variabel respon berupa data diskrit atau data cacah dan dianalisis menggunakan regresi linier maka hasil estimasi yang diperoleh menjadi tidak efisien, tidak konsisten, dan berbias. Salah satu analisis yang dapat digunakan untuk menganalisis variabel respon yang berupa data diskrit adalah regresi Poisson. Regresi Poisson didapatkan dari data yang berdistribusi Poisson, yaitu suatu distribusi untuk peristiwa yang probabilitas kejadiannya kecil, yang mana kejadiannya tergantung pada interval waktu tertentu atau di suatu daerah tertentu dengan hasil pengamatan berupa variabel diskrit dan antar variabel saling independen (Walpole, 1992).

Regresi Poisson mengasumsikan nilai varians dan rata-rata dari variabel respon mempunyai nilai yang sama (equidispersi). Akan tetapi, pada kenyataannya asumsi tersebut banyak yang tidak terpenuhi pada data rill, misalnya ditemukan kondisi nilai varians lebih besar daripada nilai rata-ratanya (overdispersi). Penggunaan regresi Poisson pada data yang mengalami overdispersi kurang akurat, karena akan mengakibatkan *underestimate* pada *standard error* dan signifikansi dari parameter regresi menjadi *overstate* sehingga kesimpulan terhadap model tidak valid (Ismail dan Jemain, 2007).

Salah satu model regresi yang dapat digunakan untuk mengatasi masalah overdispersi adalah model regresi Poisson diperumum. Wang dan Famoye (1997) menggunakan Poisson diperumum dalam memodelkan kemakmuran rumah tangga yang melanggar asumsi equidispersi. Ismail dan Jemain (2005) membandingkan model regresi Poisson dan regresi Poisson diperumum pada data pengklasifikasian risiko di bidang asuransi. Putra, dkk (2013) mengaplikasikan regresi Poisson diperumum untuk mengatasi pelanggaran asumsi equidispersi pada data jumlah siswa SMA/SMK yang tidak lulus UN tahun 2012. Umami, dkk (2013) memodelkan angka kematian bayi di Jawa Tengah tahun 2007 menggunakan regresi Poisson diperumum. Semua penelitian tersebut menggunakan pendekatan regresi parametrik, yaitu pendekatan yang digunakan untuk bentuk kurva regresi diketahui. Namun, pada kenyataannya seringkali hubungan antara variabel prediktor dan variabel respon tidak mengikuti pola tertentu tetapi cenderung tidak beraturan sehingga pendekatan parametrik kurang tepat digunakan. Pada kondisi ini pendekatan nonparametrik menjadi alternatif pilihan.

Regresi nonparametrik merupakan pendekatan yang tidak mengasumsikan bentuk kurva regresi seperti pada regresi parametrik karena data diharapkan mencari sendiri estimasi kurvanya tanpa ada pengaruh faktor subjektivitas dari perancang penelitan (Eubank, 1988). Salah satu regresi nonparametrik yang sering digunakan adalah regresi *spline truncated*. *Spline truncated* dalam regresi nonparametrik mempunyai sifat fleksibel dan mempunyai kemampuan mengestimasi perilaku data yang cenderung berbeda pada interval yang berlainan (Eubank, 1999 dan Budiantara, 2005). Pendekatan *spline truncated* mempunyai keunggulan dalam mengatasi pola data yang menunjukkan naik atau turun yang tajam dengan bantuan titik-titik knot, serta kurva yang dihasilkan relatif mulus (Hardle, 1990). Hurley, dkk (2005) membandingkan model regresi polinomial biasa dengan model regresi *spline truncated* pada data simulasi dengan struktur data yang berbeda (linier, kuadratik, kubik, dan pola data lainnya). Bintariningrum dan Budiantara (2014) memodelkan angka kelahiran kasar menggunakan metode regresi nonparametrik *spline truncated*. Dukalang (2017) memodelkan jumlah kematian bayi dengan menggunakan pendekatan regresi *spline* adaptif multivariat. Sitanggang, dkk (2018) menggunakan regresi nonparametrik *spline truncated* untuk mengetahui faktor-faktor yang mempengaruhi persentase peserta aktif KB suntik.

Dalam penelitian ini, penulis akan mengkaji lebih lanjut hasil penelitian yang telah dilakukan oleh Jao (2020) pada data jumlah kematian bayi di Provinsi Sulawesi Selatan. Pada penelitian tersebut, peneliti tidak memperhatikan asumsi equidispersi. Oleh karena itu, dalam penelitian ini asumsi equidispersi akan diperhatikan dengan menggunakan regresi Poisson diperumum sebagai perkembangan dari regresi Poisson dalam mengatasi masalah pelanggaran equidispersi dengan menggunakan estimator *spline truncated*. Berdasarkan uraian tersebut, penulis akan melakukan penelitian dengan judul **"Estimasi Model Regresi Nonparametrik Poisson Diperumum Menggunakan Estimator *Spline Truncated*"**.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka permasalahan yang akan dibahas dalam penelitian ini adalah:

1. Bagaimana estimasi parameter model regresi nonparametrik Poisson diperumum menggunakan estimator *spline truncated* yang bersesuaian dengan data jumlah kematian bayi di Provinsi Sulawesi Selatan?
2. Bagaimana penerapan model regresi nonparametrik Poisson diperumum menggunakan estimator *spline truncated* pada data jumlah kematian bayi di Provinsi Sulawesi Selatan?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan adalah jumlah kematian bayi di Provinsi Sulawesi Selatan tahun 2017 yang diperinci berdasarkan 24 kabupaten/kota.
2. Pemilihan titik knot optimal menggunakan metode *Generalized Cross Validation* (GCV).
3. Jumlah titik knot yang dicobakan adalah 1-2 titik knot dengan orde dibatasi pada penggunaan 1-2 orde.
4. Pada penelitian ini diasumsikan tidak terjadi korelasi antara variabel prediktor.

1.4 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah yang telah diuraikan, tujuan dari penelitian ini adalah:

1. Memperoleh estimasi parameter model regresi nonparametrik Poisson diperumum menggunakan estimator *spline truncated* yang bersesuaian dengan data jumlah kematian bayi di Provinsi Sulawesi Selatan.
2. Memperoleh model regresi nonparametrik Poisson diperumum menggunakan estimator *spline truncated* pada data jumlah kematian bayi di Provinsi Sulawesi Selatan.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan wawasan keilmuan yang lebih khusus kepada penulis tentang pemodelan regresi nonparametrik untuk data diskrit yang mengalami masalah overdispersi.
2. Memberikan informasi kepada pemerintah untuk menetapkan kebijakan dalam rangka mengatasi jumlah kematian bayi di Provinsi Sulawesi Selatan.

BAB II
TINJAUAN PUSTAKA

2.1 *Generalized Linear Model*

Generalized Linear Model (GLM) merupakan perluasan dari model regresi ketika variabel respon mengikuti distribusi keluarga eksponensial yaitu normal, binomial, Poisson, geometrik, binomial negatif, eksponensial, dan gamma. McCullagh dan Neldar (1989) menyatakan ada tiga komponen utama dalam GLM, yaitu:

1. Komponen acak, yaitu suatu komponen yang mengidentifikasi distribusi dari variabel respon (Y) berasal dari keluarga eksponensial. Menurut Dobson (2002), suatu variabel acak Y merupakan anggota keluarga distribusi eksponensial jika dapat dinyatakan dalam fungsi peluang berikut:

$$\begin{aligned} f(y) &= s(y)t(\theta)e^{a(y)b(\theta)} \\ &= \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right] \end{aligned} \tag{2.1}$$

dengan $a(.)$, $b(.)$, $s(.)$, $t(.)$ adalah fungsi yang diketahui, $s(y) = \exp d(y)$, $t(\theta) = \exp c(\theta)$ dan θ adalah parameter. Distribusi Poisson memiliki fungsi peluang yang dapat dinyatakan sebagai berikut:

$$\begin{aligned} f(y|\lambda) &= \frac{e^{-\lambda}\lambda^{y}}{y!} \\ &= \exp\left\{\ln\left(\frac{e^{-\lambda}\lambda^{y}}{y!}\right)\right\} \\ &= \exp\left\{\ln\left(e^{-\lambda}\lambda^{y}\right) - \ln(y!)\right\} \\ &= \exp\left\{\ln\left(\lambda^{y}e^{-\lambda}\right) - \ln(y!)\right\} \\ &= \exp\{(y\ln\lambda - \lambda) - \ln y!\} \end{aligned}$$

dengan $a(y) = y$ $\qquad c(\theta) = -\lambda$

$b(\theta) = \ln\lambda$ $\qquad d(y) = -\ln y!$

Distribusi Poisson merupakan anggota dari distribusi keluarga eksponensial. Hal ini dibuktikan dari fungsi massa peluang distribusi Poisson memenuhi Persamaan (2.1).

2. Komponen sistematik, yaitu hubungan dari sebuah vektor prediktor linier $[\eta_1, \eta_2, \dots, \eta_n]$ untuk menjelaskan variabel-variabel yang berhubungan dalam sebuah model linier.

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \ , \ i = 1,2, \dots, n \qquad (2.2)$$

atau dalam notasi matriks dapat ditulis

$$\boldsymbol{\eta} = \boldsymbol{X\beta}$$

dengan $\boldsymbol{\eta}$ adalah vektor berukuran $n \times 1$ dari observasi, $\boldsymbol{X}$ adalah matriks berukuran $n \times c$ yang elemennya terdiri dari variabel prediktor, $\boldsymbol{\beta}$ adalah vektor berukuran $c \times 1$ yang elemennya terdiri dari parameter regresi, dengan $c = p + 1$ (Agresti, 2002).

3. Fungsi penghubung (*link function*), yaitu fungsi yang menghubungkan ekspektasi dari variabel respon (Y) dengan variabel-variabel prediktor melalui persamaan linier. Misalkan $\lambda_i = E(Y_i), i = 1,2, \dots, n$, fungsi untuk menghubungkan λ_i dengan η_i adalah $g(.)$ sehingga $g(\lambda_i) = \eta_i$. Fungsi penghubung $g(.)$ menghubungkan $E(Y_i)$ dengan variabel prediktor, yaitu:

$$g(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \ , \ i = 1,2, \dots, n \qquad (2.3)$$

2.2 Regresi Poisson

Regresi Poisson didapatkan dari data yang berdistribusi Poisson, yaitu suatu distribusi untuk peristiwa yang probabilitas kejadiannya kecil, yang mana kejadiannya tergantung pada interval waktu tertentu atau di suatu daerah tertentu dengan hasil pengamatan berupa variabel diskrit dan antar variabel saling independen (Walpole, 1992). Menurut Cameron dan Trivedi (1998), suatu variabel acak Y yang bertipe diskrit akan mengikuti distribusi Poisson dengan fungsi massa peluang sebagai berikut:

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, y = 0,1,2, \dots ; \ \lambda > 0 \qquad (2.4)$$

dan $E(Y) = \lambda, Var(Y) = \lambda$.

Untuk mengetahui suatu data berdistribusi Poisson atau tidak, dapat dilakukan menggunakan uji *Kolmogorov-Smirnov* dengan statistik uji sebagai berikut:

$$D_{hitung} = maks|F_n(Y) - F_0(Y)| \qquad (2.5)$$

dengan:

$F_0(Y)$ = fungsi distribusi kumulatif yang dihipotesiskan

$F_n(Y)$ = fungsi distribusi kumulatif yang diamati

Kriteria pengujiannya yaitu tolak H_0 apabila $D_{hitung} > D_{tabel}\ (\alpha)$ atau nilai signifikansi $< \alpha$.

Regresi Poisson adalah salah satu metode yang digunakan untuk memodelkan hubungan fungsional antara variabel respon yang diasumsikan berdistribusi Poisson dengan parameter λ dan variabel prediktor. Misalkan $y_1, y_2, \dots, y_n$ adalah n pengamatan yang saling bebas dan bersesuaian dengan variabel bebas $x_1, x_2, \dots, x_n$ dengan $x_i = \left[1, x_{1i}, x_{2i}, \dots, x_{pi}\right]^T$, maka regresi Poisson dapat dinyatakan sebagai berikut:

$$E(Y_i) = \lambda_i = x_i^T \beta \tag{2.6}$$

dengan $\beta = \left[\beta_0, \beta_1, \dots, \beta_p\right]^T$.

Regresi Poisson menggunakan GLM agar modelnya dapat digunakan dalam data pengamatan, yang mana variabel responnya tidak mengharuskan berdistribusi normal. Dalam GLM, terdapat sebuah fungsi $g(.)$ yang menghubungkan rata-rata dari variabel respon dengan variabel prediktor, yaitu

$$g(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\,, \quad i = 1,2, \dots, n$$

$$g(\lambda_i) = \boldsymbol{x}_i^{\boldsymbol{T}} \boldsymbol{\beta}$$

Selanjutnya hubungan antara rata-rata dari variabel respon dan variabel prediktor dapat dinyatakan sebagai berikut:

$$\lambda_i = g^{-1}(\lambda_i) = g^{-1}\left(\boldsymbol{x}_i^{\boldsymbol{T}} \boldsymbol{\beta}\right)$$

Model regresi Poisson menggunakan fungsi penghubung logaritma natural, karena rata-rata dari variabel responnya akan berbentuk fungsi eksponensial dan menjamin bahwa nilai variabel yang diestimasi dari variabel responnya akan bernilai non-negatif. Rata-rata dari variabel respon dan variabel prediktor yang dihubungkan dengan fungsi penghubung tersebut berbentuk sebagai berikut:

$$g(\lambda_i) = \ln \lambda_i = \boldsymbol{x}_i^{\boldsymbol{T}} \boldsymbol{\beta}$$

$$\lambda_i = \exp\left(\boldsymbol{x}_i^{\boldsymbol{T}} \boldsymbol{\beta}\right) \tag{2.7}$$

Terdapat asumsi yang harus dipenuhi pada model regresi Poisson yaitu variabel respon harus berbentuk diskrit dan equidispersi yaitu nilai varians dan nilai rata-rata dari variabel respon Y_i memiliki nilai yang sama, sehingga dapat dinyatakan sebagai berikut:

$$E(Y_i) = Var(Y_i) = \lambda_i = \exp\left(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta}\right)$$

(Cameron dan Trivedi, 1998).

Berdasarkan Persamaan (2.7), fungsi massa peluang distribusi Poisson dapat dituliskan sebagai berikut:

$$f(y_i|\boldsymbol{\beta}) = \frac{e^{-\lambda_i(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})}\lambda_i\left(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta}\right)^{y_i}}{y_i!}$$

dengan $\lambda_i\left(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta}\right)$ adalah rata-rata dari distribusi Poisson dan vektor $\boldsymbol{\beta}$ menunjukkan parameter regresi yang tidak diketahui. Oleh karena itu, model regresi Poisson dengan fungsi penghubung logaritma natural dapat dituliskan sebagai berikut:

$$y_i = \lambda_i + \varepsilon_i\ ,\ \ i = 1,2,\dots,n$$

$$y_i = \exp\left(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta}\right) + \varepsilon_i\ ,\ \ i = 1,2,\dots,n \tag{2.8}$$

dengan y_i adalah variabel respon pada pengamatan ke-*i*, $\boldsymbol{x}_i^{\boldsymbol{T}}$ adalah vektor yang berukuran $1 \times c$ yang elemennya terdiri dari variabel prediktor, $\boldsymbol{\beta}$ adalah vektor yang berukuran $c \times 1$ yang elemennya terdiri dari parameter regresi, dan ε_i adalah *error* acak pada pengamatan ke-*i* (Cahyandari, 2014).

Pendugaan parameter regresi Poisson dilakukan dengan menggunakan penduga kemungkinan maksimum (*maximum likelihood estimation*). Fungsi logaritma natural *likelihood* dari distribusi Poisson adalah

$$\ln L(\boldsymbol{\beta}|y_i) = \sum_{i=1}^{n}\left[y_i\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta} - \exp\left(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta}\right) - \ln(y_i!)\right] \tag{2.9}$$

Dengan memaksimumkan model pada Persamaan (2.9), akan diperoleh penaksir *maximum likelihood*, yaitu melalui turunan pertama terhadap $\boldsymbol{\beta}$ yang disamakan dengan nol pada Persamaan (2.10) dan diselesaikan dengan suatu metode iterasi.

$$\frac{\partial \ln L(\boldsymbol{\beta}|y_i)}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n}\boldsymbol{x}_i.\exp\left(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta}\right) + \sum_{i=1}^{n} y_i\boldsymbol{x}_i = 0$$

$$\frac{\partial \ln L(\boldsymbol{\beta}|y_i)}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n}\boldsymbol{x}_i.\lambda_i + \sum_{i=1}^{n} y_i\boldsymbol{x}_i = 0$$

$$\frac{\partial \ln L(\boldsymbol{\beta}|y_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\boldsymbol{x}_i\,(y_i - \lambda_i) = 0 \tag{2.10}$$

dengan $\lambda_i = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$. Persamaan (2.10) dapat dituliskan dalam bentuk matriks sebagai berikut:

$$\frac{\partial \ln L(\boldsymbol{\beta}|y_i)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{y} - \boldsymbol{\lambda}) = 0 \tag{2.11}$$

2.3 Overdispersi

Menurut Cameron dan Trivedi (1998) dalam model regresi Poisson terdapat asumsi yang harus dipenuhi, yaitu variabel respon harus berdistribusi Poisson. Karakteristik distribusi Poisson adalah equidispersi yaitu nilai variansi sama dengan nilai rata-rata. Namun, seringkali terjadi pelanggaran asumsi tersebut karena nilai variansi lebih besar dari nilai rata-rata atau disebut overdispersi. Jika terjadi overdispersi pada data, maka model regresi Poisson kurang akurat dalam pemodelan data karena mengakibatkan *underestimate* pada *standard error* dan signifikansi dari parameter regresi menjadi *overstate* sehingga kesimpulan terhadap model tidak valid (Ismail dan Jemain, 2007).

Cameron dan Trivedi (1998) mengatakan ada atau tidaknya overdispersi dapat dilihat dari nilai *deviance* yang dibagi dengan derajat bebasnya. Nilai *deviance* adalah nilai logaritma natural dari uji rasio *likelihood*, yaitu membandingkan *current model*-nya dan *saturated model*-nya yang dituliskan sebagai berikut (McCullagh dan Nelder, 1989):

$$\phi = \frac{D}{db} \tag{2.12}$$

$$D = 2 \ln \left[\frac{L(y_i, y_i)}{L\left(\hat{\lambda}_i, y_i\right)} \right]$$

$$D = 2 \sum_{i=1}^{n} y_i \ln \left(\frac{y_i}{\hat{\lambda}_i} \right) \tag{2.13}$$

dengan, $db = n - c$, n merupakan banyaknya pengamatan, $c = p + 1$ merupakan banyaknya parameter, dan D adalah nilai *deviance*. Jika nilai statistik uji overdispersi (ϕ) lebih dari 1 maka dapat disimpulkan data mengalami overdispersi.

2.4 Regresi Poisson Diperumum

Regresi Poisson tidak tepat digunakan dalam memodelkan data yang mengalami overdispersi atau underdispersi. Oleh karena itu, perlu dilakukan pendekatan dengan model regresi yang lebih tepat. Consul dan Jain (1973) mengusulkan suatu distribusi lain sebagai alternatif dari distribusi Poisson yang disebut distribusi Poisson diperumum. Keunggulan distribusi ini adalah dapat digunakan pada data diskrit yang mengalami overdispersi, underdispersi ataupun equidispersi.

Famoye, dkk (2004) mendefinisikan fungsi peluang distribusi Poisson diperumum sebagai berikut:

$$f(y|\lambda,\varphi) = \left(\frac{\lambda}{1+\varphi\lambda}\right)^y \frac{(1+\varphi y)^{y-1}}{y!} \exp\left(\frac{-\lambda(1+\varphi y)}{1+\varphi\lambda}\right), y = 0,1,2,\ldots \qquad (2.14)$$

Distribusi Poisson diperumum memuat parameter dispersi yaitu φ. Ketika $\varphi = 0$ maka Persamaan (2.14) akan tereduksi menjadi fungsi distribusi Poisson. Hal ini menunjukkan bahwa distribusi Poisson diperumum merupakan perluasan dari distribusi Poisson. Ketika $\varphi > 0$ maka terjadi overdispersi dan ketika $\varphi < 0$ maka terjadi underdispersi (Famoye, 1993). Rata-rata dan variansi dari distribusi Poisson diperumum adalah sebagai berikut (Famoye dkk., 2004):

$$E(Y) = \lambda \text{ dan } Var(Y) = \lambda(1+\varphi\lambda)^2 \qquad (2.15)$$

Model regresi Poisson diperumum dapat dituliskan sebagai berikut (Ismail dan Jemain, 2007):

$$y_i = \lambda_i + \varepsilon_i\,,\ \ i = 1,2,\ldots,n \qquad (2.16)$$

dengan $\lambda_i = \exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})$.

Estimasi parameter pada model regresi Poisson diperumum dapat dilakukan dengan menggunakan metode *Maximum Likelihood Estimation* (MLE), yaitu dengan memaksimumkan model fungsi logaritma natural *likelihood*. Fungsi logaritma natural *likelihood* dari regresi Poisson diperumum dapat dituliskan sebagai berikut:

$$ln\, L(\boldsymbol{\beta},\varphi|y_i) = \sum_{i=1}^{n}\left[y_i\, ln\left(\frac{\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\varphi\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})}\right) + (y_i - 1)\, ln(1+\varphi y_i) - ln(y_i!) + \left(\frac{-\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})(1+\varphi y_i)}{1+\varphi\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})}\right)\right] \qquad (2.17)$$

Dengan memaksimumkan model logaritma natural *likelihood* pada Persamaan (2.17) diperoleh penaksir *maximum likelihood* untuk $\boldsymbol{\beta}$ dan φ sebagai berikut:

$$\frac{\partial \ln L(\boldsymbol{\beta},\varphi|y_i)}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n} y_i\, \boldsymbol{x}_i^{\boldsymbol{T}} - \sum_{i=1}^{n}\frac{y_i\varphi\boldsymbol{x}_i^{\boldsymbol{T}}\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\varphi\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})} + \sum_{i=1}^{n}(1+\varphi y_i)\frac{-x_i^T\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})[1+\varphi\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})]+[\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})]^2\varphi x_i^T}{[1+\varphi\exp(\boldsymbol{x}_i^{\boldsymbol{T}}\boldsymbol{\beta})]^2} = 0 \qquad (2.18)$$

dan

$$\frac{\partial \ln L(\boldsymbol{\beta},\varphi|y_i)}{\partial \varphi} = -\sum_{i=1}^{n}\frac{y_i \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\varphi\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} + \sum_{i=1}^{n}\frac{y_i(y_i-1)}{1+\varphi y_i} + \sum_{i=1}^{n}\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})-y_i)}{(1+\varphi\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))^2} = 0 \quad (2.19)$$

Persamaan (2.18) dan (2.19) memperlihatkan bahwa turunan pertama dari model fungsi logaritma natural *likelihood* pada estimasi parameter menggunakan metode *maximum likelihood estimation* berbentuk implisit dan sulit diatasi secara analitik, sehingga diperlukan suatu metode iterasi untuk memperoleh estimasi parameter model regresi Poisson diperumum.

2.5 Regresi *Spline Truncated*

Regresi nonparametrik merupakan suatu metode statistika yang digunakan untuk mengetahui hubungan antara variabel respon dan prediktor yang tidak diketahui bentuk fungsinya, hanya diasumsikan fungsi mulus (*smooth*) dalam arti termuat dalam suatu ruang fungsi tertentu. Berdasarkan pendekatan nonparametrik, data diharapkan mencari sendiri bentuk estimasinya dengan demikian pendekatan nonparametrik memiliki fleksibilitas yang tinggi (Eubank, 1988). Secara umum bentuk model regresi nonparametrik dapat ditulis sebagai berikut (Eubank, 1988 dan Wahba, 1990):

$$y_i = f(x_i) + \varepsilon_i\ , i = 1,2,\dots,n \quad (2.20)$$

dengan:

y_i = variabel respon pada pengamatan ke-*i*

$f(x_i)$ = fungsi yang tidak diketahui bentuknya

x_i = variabel prediktor pada pengamatan ke-*i*

ε_i = *error* acak pada pengamatan ke-*i*

Salah satu regresi nonparametrik yang sering digunakan adalah regresi *spline truncated*. Eubank (1988) menyatakan *spline truncated* merupakan potongan (*piecewise*) polinomial yang memiliki sifat tersegmen kontinu sehingga efektif menjelaskan karakteristik lokal dari fungsi data. Pendekatan *spline truncated* mempunyai keunggulan dalam mengatasi pola data yang menunjukkan naik atau turun yang tajam dengan bantuan titik-titik knot, serta kurva yang dihasilkan relatif mulus. Titik knot merupakan titik perpaduan bersama yang menunjukkan pola perilaku fungsi *spline* pada selang yang berbeda. Titik knot diambil pada selang $a < K_1 < K_2 < \cdots < K_r < b$, yang mana a adalah nilai minimum dan b adalah nilai maksimum dari suatu data (Hardle, 1990).

Fungsi *spline truncated* berorde m dengan titik-titik knot $K_1, K_2, \dots, K_r$ didefinisikan sebagai sembarang fungsi f yang disajikan dalam bentuk (Eubank, 1988):

$$f(x_i) = \sum_{l=0}^{m} \beta_l x_i^l + \sum_{k=1}^{r} \beta_{m+k}(x_i - K_k)_+^m \tag{2.21}$$

dengan fungsi *truncated*

$$(x_i - K_k)_+^m = \begin{cases} (x_i - K_k)^m & , x_i \geq K_k \\ 0 & , x_i < K_k \end{cases}$$

dan $(x_i - K_k)_+^m$ adalah fungsi potongan polinomial tersegmen yang kontinu.

Jika Persamaan (2.21) disubtitusikan ke dalam Persamaan (2.20) maka akan diperoleh model regresi nonparametrik *spline* keluarga polinomial *truncated* orde ke-m dan r titik knot dengan satu variabel prediktor sebagai berikut:

$$y_i = \sum_{l=0}^{m} \beta_l x_i^l + \sum_{k=1}^{r} \beta_{m+k}(x_i - K_k)_+^m + \varepsilon_i \tag{2.22}$$

dengan β adalah parameter, m adalah derajat polinomial, K_k adalah titik knot ke-k dengan $k = 1,2,\dots,r$, dan ε_i adalah *error* acak pada pengamatan ke-i dengan $i = 1,2,\dots,n$. Sedangkan bentuk umum regresi *spline truncated* multiprediktor berorde m dengan r titik knot dapat diberikan pada persamaan berikut:

$$y_i = \sum_{j=1}^{p} f(x_{ji}) + \varepsilon_{ji} \tag{2.23}$$

dengan,

$$f(x_{ji}) = \beta_{j0} + \sum_{l=1}^{m} \beta_{jl} x_{ji}^l + \sum_{k=1}^{r} \beta_{j(m+k)}\left(x_{ji} - K_{jk}\right)_+^m, j = 1,2,\dots,p \tag{2.24}$$

Secara umum model regresi *spline truncated* orde ke-m dan r titik knot dengan multiprediktor dapat ditulis sebagai berikut:

$$y_i = \beta_0 + \sum_{j=1}^{p} \sum_{l=1}^{m} \beta_{jl} x_{ji}^l + \sum_{j=1}^{p} \sum_{k=1}^{r} \beta_{j(m+k)}\left(x_{ji} - K_{jk}\right)_+^m + \varepsilon_{ji} \tag{2.25}$$

dengan, $\beta_0 = \sum_{j=1}^{p} \beta_{j0}$

$$\left(x_{ji} - K_{jk}\right)_+^m = \begin{cases} \left(x_{ji} - K_{jk}\right)^m & , x_{ji} \geq K_{jk} \\ 0 & , x_{ji} < K_{jk} \end{cases}$$

Model regresi nonparametrik pada Persamaan (2.20) dapat dituliskan dalam bentuk matriks sebagai berikut:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Apabila model regresi nonparametrik *spline truncated* pada Persamaan (2.22) disajikan dalam bentuk matriks, diperoleh:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m & (x_1 - K_1)_+^m & (x_1 - K_2)_+^m & \cdots & (x_1 - K_r)_+^m \\ 1 & x_2 & \cdots & x_2^m & (x_2 - K_1)_+^m & (x_2 - K_2)_+^m & \cdots & (x_2 - K_r)_+^m \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^m & (x_n - K_1)_+^m & (x_n - K_2)_+^m & \cdots & (x_n - K_r)_+^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \\ \beta_{m+1} \\ \beta_{m+2} \\ \vdots \\ \beta_{m+r} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

atau dapat ditulis dalam notasi matriks, menjadi:

$$\boldsymbol{y} = \boldsymbol{X}[K_1, K_2, \dots, K_r]\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.26}$$

Model regresi *spline truncated* multiprediktor pada Persamaan (2.25) dalam notasi matriks dapat dituliskan sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X_j}\left[K_{j1}, K_{j2}, \dots, K_{pr}\right]\boldsymbol{\beta_j} + \boldsymbol{\varepsilon} \tag{2.27}$$

Estimasi parameter $\boldsymbol{\beta_j}$ dapat diperoleh dengan menggunakan metode kuadrat terkecil, yaitu dengan meminimumkan jumlah kuadrat *error* sebagai berikut:

$$\min_{\boldsymbol{\beta} \in R^{p(m+r)}} \{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}\} = \min_{\boldsymbol{\beta} \in R^{p(m+r)}} \left\{ \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}^T \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \right\}$$

$$\min_{\boldsymbol{\beta} \in R^{p(m+r)}} \{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}\} = \min_{\boldsymbol{\beta} \in R^{p(m+r)}} \left\{ \left(\boldsymbol{y} - \boldsymbol{X_j}\left[K_{j1}, K_{j2}, \dots, K_{pr}\right]\boldsymbol{\beta_j}\right)^T \left(\boldsymbol{y} - \boldsymbol{X_j}\left[K_{j1}, K_{j2}, \dots, K_{pr}\right]\boldsymbol{\beta_j}\right) \right\}$$

Misalkan $K_{jk} = \left[K_{j1}, K_{j2}, \dots, K_{pr}\right]$ maka penyajian matriks diberikan sebagai berikut:

$$\begin{aligned} \sum_{i=1}^{n} \varepsilon_i^2 &= \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \\ &= \left(\boldsymbol{y} - \boldsymbol{X_j}\left[K_{j1}, K_{j2}, \dots, K_{pr}\right]\boldsymbol{\beta_j}\right)^T \left(\boldsymbol{y} - \boldsymbol{X_j}\left[K_{j1}, K_{j2}, \dots, K_{pr}\right]\boldsymbol{\beta_j}\right) \\ &= \left(\boldsymbol{y} - \boldsymbol{X_j}\left[K_{jk}\right]\boldsymbol{\beta_j}\right)^T \left(\boldsymbol{y} - \boldsymbol{X_j}\left[K_{jk}\right]\boldsymbol{\beta_j}\right) \\ &= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}_j^T \boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{y} + \boldsymbol{\beta}_j^T \boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{X_j}\left[K_{jk}\right]\boldsymbol{\beta_j} \end{aligned} \tag{2.28}$$

Selanjutnya, Persamaan (2.28) diturunkan terhadap vektor $\boldsymbol{\beta}_j^T$ dan disamakan dengan nol, maka diperoleh:

$$\frac{\partial \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}_j^T} = \frac{\partial\left(\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}_j^T \boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{y} + \boldsymbol{\beta}_j^T \boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{X_j}\left[K_{jk}\right]\boldsymbol{\beta_j}\right)}{\partial \boldsymbol{\beta}_j^T}$$

$$-2\boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{y} + 2\boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{X_j}\left[K_{jk}\right]\boldsymbol{\beta_j} = 0$$

$$\boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{X_j}\left[K_{jk}\right]\boldsymbol{\beta_j} = \boldsymbol{X}_j^T \left[K_{jk}\right]\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} = \left(\boldsymbol{X}_j^T[K_{jk}]\boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\right)^{-1}\boldsymbol{X}_j^T[K_{jk}]\boldsymbol{y} \tag{2.29}$$

dengan:

$\boldsymbol{X}_{\boldsymbol{j}}[K_{jk}] = \boldsymbol{X}_{\boldsymbol{j}}[K_{j1}, K_{j2}, \dots, K_{pr}]$

Akibatnya estimasi untuk kurva regresi $f(x_{ji})$ diberikan oleh (Hidayat, 2017):

$$\widehat{\boldsymbol{y}}[K_{jk}](x_{ji}) = \widehat{\boldsymbol{f}}[K_{jk}](x_{ji}) = \boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$$

$$\widehat{\boldsymbol{f}}[K_{jk}](x_{ji}) = \boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\left(\boldsymbol{X}_j^T[K_{jk}]\boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\right)^{-1}\boldsymbol{X}_j^T[K_{jk}]\boldsymbol{y}$$

$$\widehat{\boldsymbol{f}}[K_{jk}](x_{ji}) = \boldsymbol{A}[K_{jk}]\boldsymbol{y} \tag{2.30}$$

dengan, $\boldsymbol{A}[K_{jk}] = \boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\left(\boldsymbol{X}_j^T[K_{jk}]\boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\right)^{-1}\boldsymbol{X}_j^T[K_{jk}]$ dan $\boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]$ adalah matriks dari model yang bergantung pada titik knot dan $[K_{jk}] = [K_{j1}, K_{j2}, \dots, K_{pr}]$ adalah titik knot.

2.6 Pemilihan Titik Knot Optimal

Salah satu langkah penting dalam pemodelan regresi nonparametrik *spline truncated* adalah memilih titik knot yang optimal. Jika titik knot besar maka estimasi dari fungsi yang didapat akan semakin mulus namun kemampuan memetakan datanya kurang baik. Sebaliknya, jika titik knot terlalu kecil maka estimasi fungsi yang didapat akan lebih kasar (*undersmoothing*). Oleh karena itu, diperlukan titik knot yang optimal (Nurcahayani dkk., 2019). Budiantara, dkk (2009) mengatakan salah satu metode yang sering digunakan dalam memilih titik knot optimal adalah *Generalized Cross Validation* (GCV). Titik knot optimal diperoleh dengan melihat nilai GCV dari masing-masing orde dan titik knot kemudian dipilih dari nilai $GCV[K_{jk}]$ yang paling minimum. Menurut Wu dan Zang (2006), metode GCV dapat dituliskan sebagai berikut:

$$GCV[K_{jk}] = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\left(n^{-1}tr\left[\boldsymbol{I} - \boldsymbol{A}[K_{jk}]\right]\right)^2} \tag{2.31}$$

dengan, $\boldsymbol{A}[K_{jk}] = \boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\left(\boldsymbol{X}_j^T[K_{jk}]\boldsymbol{X}_{\boldsymbol{j}}[K_{jk}]\right)^{-1}\boldsymbol{X}_j^T[K_{jk}]$.

2.7 Metode *Maximum Likelihood Estimation*

Metode *Maximum Likelihood Estimation* (MLE) dapat digunakan untuk mengestimasi parameter suatu model yang diketahui fungsi peluangnya. Misalkan

$Y_1, Y_2, \ldots, Y_n$ adalah sampel acak dari populasi dengan fungsi peluang $f(y|\boldsymbol{\theta})$ yang bergantung pada $\boldsymbol{\theta} = \boldsymbol{\beta}, \varphi$ dengan $\boldsymbol{\theta}$ adalah parameter yang tidak diketahui. Karena $Y_1, Y_2, \ldots, Y_n$ saling bebas, maka fungsi peluang bersama dari $Y_1, Y_2, \ldots, Y_n$ dapat dinyatakan sebagai berikut:

$$f(y_1, y_2, \ldots, y_n|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta}) \cdot f(y_2|\boldsymbol{\theta}) \cdot \ldots \cdot f(y_n|\boldsymbol{\theta})$$

Fungsi *likelihood* didefinisikan sebagai fungsi peluang bersama dari $Y_1, Y_2, \ldots, Y_n$ yang dianggap sebagai fungsi dari $\boldsymbol{\theta}$ yang dituliskan sebagai $L(\boldsymbol{\theta}|y)$, yaitu:

$$L(\boldsymbol{\theta}|y) = f(y_1, y_2, \ldots, y_n|\boldsymbol{\theta})$$

$$L(\boldsymbol{\theta}|y) = f(y_1|\boldsymbol{\theta}) \cdot f(y_2|\boldsymbol{\theta}) \cdot \ldots \cdot f(y_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) \qquad (2.32)$$

Estimator *maximum likelihood* $\widehat{\boldsymbol{\theta}}$ adalah nilai $\boldsymbol{\theta}$ yang memaksimumkan fungsi *likelihood* $L(\boldsymbol{\theta}|y)$. Namun, lebih mudah bekerja dengan logaritma natural dari fungsi *likelihood*, yaitu $l(\boldsymbol{\theta}|y) = \ln L(\boldsymbol{\theta}|y)$. Karena fungsi logaritma natural adalah fungsi yang monoton naik, maka nilai yang memaksimumkan fungsi $l(\boldsymbol{\theta}|y)$ sama dengan memaksimumkan fungsi $L(\boldsymbol{\theta}|y)$ sehingga dapat dituliskan sebagai berikut:

$$l(\boldsymbol{\theta}|y) = \ln L(\boldsymbol{\theta}|y)$$

$$l(\boldsymbol{\theta}|y) = \ln\left[\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})\right] = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}) \qquad (2.33)$$

Untuk memperoleh nilai $\boldsymbol{\theta}$ yang memaksimumkan fungsi $l(\boldsymbol{\theta}|y)$, maka $l(\boldsymbol{\theta}|y)$ diturunkan terhadap $\boldsymbol{\theta}$ dan kemudian menyamakannya dengan nol seperti pada Persamaan (2.34) berikut (Hogg dkk., 2013).

$$l'(\boldsymbol{\theta}|y) = \frac{\partial l(\boldsymbol{\theta}|y)}{\partial \boldsymbol{\theta}} = 0 \qquad (2.34)$$

2.8 Algoritma Fisher *Scoring*

Algoritma adalah prosedur komputasi yang mana mengambil sebuah nilai atau menentukan nilai sebagai *input* dan menghasilkan beberapa nilai sebagai *output* (Cormen dkk., 2011). Algoritma Fishe*r scoring* adalah salah satu bentuk pengembangan dari algoritma *Newton Raphson* dengan mengganti matriks Hessian dengan matriks informasi. Rumus perulangan Fisher *scoring* adalah

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} + I\left(\boldsymbol{\theta}^{(t)}\right)^{-1} l'\left(\boldsymbol{\theta}^{(t)}\right) \qquad (2.35)$$

dengan $I^{(t)}$ adalah taksiran ke-t dari matriks informasi yang diamati. Matriks informasi adalah negatif dari nilai ekspektasi matriks turunan kedua fungsi logaritma natural *likelihood* yaitu $I = -E\left(\frac{\partial^2 \boldsymbol{l}(\boldsymbol{\theta}|y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)$ yang berukuran $(p+1) \times (p+1)$.

Adapun langkah-langkah dalam algoritma Fisher *scoring* adalah sebagai berikut:

a. Menentukan nilai awal $\widehat{\boldsymbol{\theta}}^{(0)}$.

b. Menentukan $l'\left(\boldsymbol{\theta}^{(t)}\right)$ dan $I\left(\boldsymbol{\theta}^{(t)}\right)$.

c. Menghitung estimator parameter untuk $t = 0,1,2, \ldots$ dengan menggunakan Persamaan (2.35).

d. Mengulangi iterasi sampai diperoleh nilai yang konvergen, yaitu $\left|\boldsymbol{\theta}^{(t+1)} - \widehat{\boldsymbol{\theta}}^{(t)}\right| \leq \epsilon$ dengan ϵ adalah konstanta positif yang ditentukan.

2.9 Jumlah Kematian Bayi

Jumlah kematian bayi adalah jumlah kematian yang terjadi saat bayi lahir hingga bayi belum berusia tepat satu tahun. Banyak faktor yang dikaitkan dengan kematian bayi. Secara garis besar, dari sisi penyebabnya, kematian bayi ada dua macam yaitu endogen dan eksogen. Kematian bayi endogen atau yang biasa disebut dengan kematian neonatal adalah kematian bayi yang terjadi pada bulan pertama setelah dilahirkan, dan umumnya disebabkan oleh faktor-faktor yang dibawa anak sejak lahir, yang diperoleh dari orang tuanya pada saat konsepsi atau didapat selama kehamilan. Sedangkan kematian bayi eksogen atau kematian post neonatal adalah kematian bayi yang terjadi setelah usia satu bulan sampai menjelang usia satu tahun yang disebabkan oleh faktor-faktor yang berkaitan dengan pengaruh lingkungan luar (Dinkes, 2018).

Jumlah kematian bayi merupakan indikator yang mengukur derajat kesehatan serta indikator yang menilai tingkat kesejahteraan suatu bangsa. Keadaan gizi merupakan salah satu penyebab dasar kematian bayi dan anak. Gizi buruk seringkali disertai penyakit seperti TB, ISPA, diare, dan lain-lain. Dalam Profil Kesehatan Indonesia dijelaskan bahwa penyebab kematian bayi yang terbanyak disebabkan oleh pertumbuhan janin yang lambat, kekurangan gizi pada janin, kelahiran prematur dan berat badan lahir rendah (Dinkes, 2018).