PERBANDINGAN METODE *CROSS VALIDATION* DAN *GENERALIZED CROSS VALIDATION* DALAM REGRESI NONPARAMETRIK BIRESPON SPLINE

Luh Putu Safitri Pratiwi, S.Si.,M.Si.
STMIK STIKOM Bali
e-mail: safitri.pratiwi@yahoo.com

Abstract

Regression analysis is one of the most popular methods in statistics to explain causal relationships between one predictor variables to one response variable. In general, modeling can be done using regression analysis. The regression curve can be assumed by the parametric regression approach and the nonparametric regression approach. However, not all data acquired follows a certain pattern so that this type of data uses a nonparametric regression approach. The nonparametric regression approach is not related to the assumption of the regression curve form as it is to the parametric regression, and more flexible. There are several techniques performed for estimation in nonparametric regression ie Spline. Some cases in the regression analysis found many problems that can not be solved by simple regression analysis of one response because if using two response variables in the research, it must be seen the value of correlation between variables. As a result, regression issues must be solved by the birespon regression model. This study aims to describe the IMR and malnutrition status of children under five and to get the Spline model in the best birespon nonparametric regression through the relationship between the suspected variables by using Cross Validation (CV) and Generalized Cross Validation (GCV) methods. The results obtained are the best model that is suitable for health by using CV method, obtained the minimum CV value located on Spline model linear one knot that is equal to 77.37831 with MSE of 76.75449.

Keywords: *Nonparametric Regression*, GCV, CV, *Spline*, *Birespon*.

Abstrak

Analisis regresi merupakan salah satu metode yang sangat populer dalam statistika untuk menjelaskan hubungan sebab akibat antara satu/beberapa variabel prediktor terhadap satu variabel respon. Pada umumnya, pemodelan yang dapat dilakukan dengan menggunakan analisis regresi. Kurva regresi dapat diduga dengan pendekatan regresi parametrik dan pendekatan regresi nonparametrik. Namun, tidak semua data yang diperoleh mengikuti pola tertentu sehingga jenis data ini menggunakan pendekatan regresi nonparametrik. Pendekatan regresi nonparametrik tidak terkait dengan asumsi bentuk kurva regresi seperti halnya pada regresi parametrik, dan lebih fleksibel. Ada beberapa teknik yang dilakukan untuk estimasi dalam regresi nonparametrik yaitu Spline. Beberapa kasus dalam analisis regresi banyak dijumpai permasalahan yang tidak dapat diselesaikan dengan analisis regresi sederhana satu respon karena jika menggunakan dua variabel respon pada penelitian, maka harus dilihat nilai korelasi antar variabel. Akibatnya, persoalan regresi harus diselesaikan dengan model regresi birespon. Penelitian ini bertujuan untuk mendeskripsikan AKB dan status gizi buruk balita dan mendapatkan model Spline dalam regresi nonparametrik birespon terbaik melalui hubungan antara variabel yang diduga berpengaruh dengan menggunakan metode *Cross Validation* (CV) dan Generalized Cross Validation (GCV). Hasil yang didapat yaitu model terbaik yang sesuai untuk derajat kesehatan di Indonesia pada tahun 2015 yaitu dengan menggunakan metode CV dengan nilai CV minimum yang terletak pada model Spline linier satu knot yakni sebesar 77.37831 dengan MSE sebesar 76.75449.

Kata kunci: Regresi Nonparametrik, GCV, CV, Spline, Birespon.

I. PENDAHULUAN

Upaya untuk meningkatkan derajat kesehatan tidak telepas dari indikator yang terlibat. Indikator tersebut pada umumnya tercermin pada kondisi angka kematian, angka kesakitan dan status gizi. Derajat kesehatan masyarakat digambarkan melalui Angka Kematian Bayi (AKB), Angka Kematian Balita (AKABA), Angka Kematian Ibu (AKI), angka morbiditas beberapa penyakit, dan status gizi balita [1]. Angka kematian bayi (AKB) dan status gizi balita merupakan indikator yang paling menggambarkan tingkat kesejahteraan masyarakat [2].

Berdasarkan data [1], AKB di Provinsi Bali dari tahun 2006 sampai dengan tahun 2015 menunjukan trend yang fluktuatif, meski sudah lebih rendah dari angka kematian bayi secara nasional,. AKB di Provinsi Bali tetap lebih rendah dibandingkan dengan target Renstra Dinkes Provinsi Bali yaitu 15 per 1.000 kelahiran hidup di tahun 2014 dan target MDG's tahun 2015 yaitu 23 per 1.000 kelahiran hidup. AKB terendah dicapai oleh Kota Denpasar sebesar 0,6 per 1000 kelahiran hidup, sedangkan AKB tertinggi dicapai oleh Kabupaten Karangasem sebesar 10,6 per 1000 kelahiran hidup. Jumlah kasus balita gizi buruk di Bali tahun 2015 sebesar 80,52% mengalami peningkatan dibandingkan tahun 2014 sebesar 79,92%, akan tetapi capaian ini masih dibawah target Renstra Dinkes Provinsi Bali sebesar 81%. Kabupaten yang belum mencapai target yaitu Buleleng (67,12%) dan Bangli (73,52%). Sementara kabupaten lainnya sudah memenuhi target. Capaian tertinggi dicapai oleh kabupaten Klungkung (86,1%) dan Badung (86,02%). Masih ada 19% balita yang belum terpantau status gizinya, hal inilah yang memungkinkan terjadi masalah-masalah kesehatan. Masih terdapatnya Kabupaten/kota yang memiliki AKB dan status gizi Buruk balita tinggi tentunya tidak terlepas dari faktor-faktor yang mempengaruhi, yang secara tepat dapat dilakukan dengan pemodelan terhadap dua indikator tersebut.

Pada umumnya, pemodelan yang dapat dilakukan dengan menggunakan analisis regresi. Kurva regresi dapat diduga dengan pendekatan regresi parametrik dan pendekatan regresi nonparametrik. Pendekatan regresi parametrik digunakan jika pola data mengikuti bentuk pola tertentu misalnya linier, kuadratik, dan kubik [3]. Namun, tidak semua data yang diperoleh mengikuti pola tertentu sehingga jenis data ini menggunakan pendekatan regresi nonparametrik. Pendekatan regresi nonparametrik tidak terkait dengan asumsi bentuk kurva regresi seperti halnya pada regresi parametrik, dan lebih fleksibel. Ada beberapa teknik yang dilakukan untuk estimasi dalam regresi nonparametrik yaitu Spline. Regresi Spline mempunyai interpretasi statistik dan interpretasi visual yang sangat khusus dan sangat baik [4], sehingga memiliki keistimewaan dibandingkan regresi lainnya.

Pemilihan parameter penghalus optimal dalam regresi Spline pada hakikatnya merupakan pemilihan lokasi titik knot [5]. Budiantara [6] menyebutkan bentuk estimator Spline sangat dipengaruhi oleh nilai parameter penghalus, jika nilai parameter penghalus sangat kecil maka akan memberikan estimator kurva regresi yang sangat kasar. Sebaliknya, jika nilai parameter penghalus sangat besar maka akan dihasilkan estimator kurva regresi nonparametrik yang sangat mulus sehingga perlu dipilih parameter penghalus yang optimal agar diperoleh estimator yang paling sesuai untuk data. Beberapa metode untuk memilih parameter penghalus yaitu *Unbiased Risk* (UBR) [7], *Cross Validation* (CV) [8] dan [9] memberikan suatu metode *Generalized Cross Validation* (GCV). Dalam beberapa kasus, CV memiliki kaitan yang erat dengan GCV. GCV merupakan modifikasi dari CV yang didapat dengan meminimumkan fungsi CV, sehingga menarik untuk dilakukan suatu pembandingan antara metode CV maupun GCV [10].

Beberapa kasus dalam analisis regresi banyak dijumpai permasalahan yang tidak dapat diselesaikan dengan analisis regresi sederhana satu respon karena jika menggunakan dua variabel respon pada penelitian, maka harus dilihat nilai korelasi antar variabel. Apabila variabel respon dianalisis secara parsial atau satu-satu korelasi antara variabel respon yang akan diteliti tidak akan menghasilkan model yang optimal. Akibatnya, persoalan regresi harus diselesaikan dengan model regresi birespon.

[6] dalam penelitiannya menyebutkan bahwa untuk memilih model Spline terbaik dalam regresi nonparametrik dapat digunakan fungsi prediksi yaitu uji CV, uji GCV, uji GML maupun uji UBR. Penelitian mengenai metode CV dan metode GCV pernah dilakukan oleh [11], diaplikasikan dengan menggunakan data fertilitas di Provinsi Jawa Timur.

Berdasarkan hasil-hasil penelitian yang telah

diuraikan di atas, maka penelitian ini melihat karakteristik AKB dan status gizi dan memodelkan kedua indikator tersebut dengan menggunakan pendekatan regresi nonparametrik birespon Spline dengan metode CV dan GCV.

A. Faktor – Faktor yang Mempengaruhi Angka Kematian Bayi dan Angka Status Gizi Buruk

Mortalitas atau kematian adalah keadaan hilangnya semua tanda-tanda kehidupan secara permanen yang dapat terjadi setiap saat pada siapa saja setelah kelahiran hidup. Mortalitas dapat diukur dari Angka Kematian Bayi (AKB), yang mengukur banyaknya bayi yang meninggal sebelum mencapai usia 1 tahun yang dinyatakan dalam 1000 kelahiran hidup pada tahun yang sama [12]. Secara garis besar, penyebabnya kematian bayi ada dua yaitu endogen dan eksogen. Kematian bayi endogen atau yang dikenal atau yang umum 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, diperoleh dari orang tuanya selama dalam kandungan [2]. Sedangkan kematian bayi eksogen atau kematian post neo-natal, adalah kematian bayi yang terjadi setelah satu bulan sampai menjelang usia satu tahun yang disebabkan oleh faktor-faktor yang berhubungan dengan pengaruh lingkungan sekitar [2].

Indikator kedua dalam derajat kesehatan yaitu status gizi balita yang merupakan ukuran kondisi tubuh seseorang yang dapat dilihat dari makanan yang dikonsumsi dan penggunaan zat-zat gizi di dalam tubuh. Status gizi dibedakan atas gizi buruk, gizi kurang, gizi baik, dan gizi lebih. Status gizi masyarakat dapat dilihat dari indikator banyaknya balita dengan gizi buruk. Gizi buruk merupakan status kondisi seseorang yang kekurangan nutrisi, atau nutrisinya di bawah standar rata-rata [13]. Menurut [14] faktor-faktor yang mempengaruhi status gizi ialah tingkat pendapatan keluarga, tingkat pengetahuan ibu, tingkat pendidikan ibu, tingkat pekerjaan ibu, dan tingkat asupan makanan.

B. Penelitian yang Terkait

Beberapa penelitian tentang Spline dalam regresi nonparametrik birespon yaitu:

1. [15] dengan judul faktor-faktor yang mempengaruhi derajat kesehatan dengan menggunakan regresi multivariat (studi kasus : derajat kesehatan kabupaten dan kota di Provinsi Sumatera Barat). Hasil penelitian yang di dapat ialah faktor-faktor yang mempengaruhi angka kematian bayi dan persentase gizi buruk balita adalah persentase penduduk dengan akses sanitasi yang layak dan persentase berat bayi lahir rendah (BBLR)
2. [16] dengan judul analisis faktor – faktor yang mempengaruhi persentase penduduk miskin dan pengeluaran perkapita makanan di Jawa Timur dengan metode regresi nonparametrik birespon Spline. Hasil penelitian yang di dapat ialah model regresi nonparametrik birespon Spline terbaik adalah model Spline linier dengan satu titik knot
3. [17] dengan judul pendekatan regresi nonparametrik birespon untuk pemodelan determinan tingkat pendidikan di Pulau Papua, model terbaik yang diperoleh dengan GCV dan MSE yang terkecil adalah model regresi nonparametrik birespon Spline dengan 1 titik knot.
4. [11] dengan judul metode *Cross Validation* dan *Generalized Cross Validation* dalam regresi nonparametrik Spline studi kasus data fertilitas di Jawa Timur hasil yang diperoleh estimasi kurva regresi nonparametrik Spline dengan metode GCV memberikan hasil yang lebih baik dibandingkan menggunakan metode CV dengan memperhatikan kriteria kebaikan model serta pengujian asumsi residual model.

C. Regresi Parametrik

Regresi parametrik merupakan metode yang digunakan untuk mengetahui pola hubungan antara variabel respon dan variabel bebas, yang diketahui bentuk kurva regresinya. Model persamaan regresi sebagai berikut.

$$y_i = f(x_i) + \varepsilon_i \quad ; i = 1,2,\cdots,n \tag{1}$$

dengan :

y_i merupakan respon ke-i, $f(x_i)$ merupakan kurva regresi, dan ε_i merupakan *error* yang diasumsikan identik, independen, dan berdistribusi normal

Model regresi parametrik linear dengan variabel prediktor $x_1, x_2, \ldots x_m$ secara umum dapat dituliskan pada persamaan berikut.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \tag{2}$$
$$i = 1,2,\cdots,n$$

D. Regresi Nonparametrik Spline

Model regresi nonparametrik Spline secara umum dapat disajikan sebagai berikut [18].

$$y_i = f(t_i) + £_i \quad ; i = 1,2,\cdots,n \tag{3}$$

dengan $f(t_i)$ merupakan fungsi Spline berorde *p* dengan titik knot $k_1, k_2, \cdots, k_r$ dan $£_i$ adalah *error* yang berdistribusi normal independen dengan mean nol dan varians $\pm^2$. Titik knot merupakan titik perpaduan bersama yang memperlihatkan terjadinya perubahan pola perilaku dari fungsi Spline pada interval-interval yang berbeda.

Apabila persamaan (3) disubstitusikan kedalam persamaan (2) maka diperoleh persamaan regresi nonparametrik Spline sebagai berikut.

$$f(t_i) = \sum_{j=0}^{p} ¥_j t_i^j + \sum_{l=1}^{r} ¥_{p+l} (t_i - k_l)_+^p \tag{4}$$

$(t_i - k_l)_+^p$ merupakan fungsi *truncated* (potongan) yang dapat dijabarkan sebagai berikut.

$$(t_i - k_l)_+^p = \begin{cases} (t_i - k_l)^p & , t_i \geq k_l \\ 0 & , t_i < k_l \end{cases} \tag{5}$$

Bila Persamaan (5) disubstitusikan ke persamaan (3) akan menghasilkan model regresi nonparametrik Spline sebagai berikut.

$$y_i = \sum_{j=0}^{p} ¥_j t_i^j + \sum_{l=1}^{r} ¥_{p+l} (t_i - k_l)_+^p + £_i \quad ; i = 1,2,\cdots,n \tag{6}$$

Estimasi regresi nonparametrik Spline *truncated* dapat diperoleh dengan menggunakan metode *Maximum Likelihood Estimation* (MLE). Apabila pada Persamaan (6) diasumsikan *error* berdistribusi normal y_i juga berdistribusi normal dengan *mean* $f(t_i)$ dan varians $\pm^2$. Sehingga diperoleh fungsi *Likelihood* sebagai berikut.

$$L(y, f) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{\exp\left(-\frac{1}{2\pm^2}(y_i - f(t_i))^2\right)}{\left(2®\pm^2\right)} = \left(2®\pm^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\pm^2}(y_i - f(t_i))^2\right) \tag{7}$$

Estimasi titik untuk fungsi f didapatkan dengan memaksimumkan fungsi *Likelihood* $L(y, f)$ yang dapat dijabarkan sebagai berikut.

$$\max_f \{L(y, f)\} = \max_{¥£R^{p+r}} \left\{ \left(2®\pm^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\pm^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} ¥_i t_i^j + - \sum_{l=1}^{r} ¥_{p+l} (t_i - k_l)_+^p \right)^2 \right) \right\} \tag{8}$$

Kemudian menerapkan transformasi logaritma sehingga menghasilkan persamaan sebagai berikut

$$\log L(y, ¥) = -\frac{n}{2}(2®\pm^2) - \frac{1}{2\pm^2} \sum_{i=1}^{n} \left(y_j - \sum_{j=0}^{p} ¥_j t_i^j + - \sum_{l=1}^{r} ¥_{p+l} (t_i - k_l)_+^p \right)^2 \tag{9}$$

dengan $\mathbf{k} = (k_1, k_2, \cdots, k_r)'$.

Bila Persamaan (9) diturunkan secara parsial terhadap $\boldsymbol{\gamma}$ dan disamakan dengan nol pada sisi kanan akan mendapatkan hasil yang dijabarkan sebagai berikut. (2) (2.4)

$$\frac{\partial(\log(y, ¥))}{\partial \boldsymbol{\gamma}} = 0 \tag{10}$$

Estimasi *Likelihood* untuk $\boldsymbol{\gamma}(\mathbf{k})$ didapatkan dengan melakukan penjabaran kembali pada persamaan (10). Hasil estimasi parameter $\hat{\boldsymbol{\gamma}}(\mathbf{k})$ dapat dilihat pada persamaan berikut.

$$\hat{\boldsymbol{\gamma}}(\mathbf{k}) = \mathbf{X}(\mathbf{k})\left(\mathbf{X}(\mathbf{k})'\mathbf{X}(\mathbf{k})\right)^{-1}\mathbf{X}(\mathbf{k})'\mathbf{y} \tag{11}$$

E. Regresi Nonparametrik Birespon Spline

Menurut [19] dalam analisis regresi Spline jika terdapat satu variabel respon dan satu variabel prediktor maka regresi ini disebut dengan regresi Spline univariabel. Sebaliknya, apabila terdapat satu variabel respon dan lebih dari satu variabel prediktor, maka regresi tersebut dinamakan regresi Spline multivariabel. Sedangkan regresi birespon didefinisikan sebagai salah satu model regresi yang memiliki variabel respon lebih dari satu dan diantara variabel-variabel tersebut terdapat korelasi atau hubungan yang kuat [20]. Jika regresi birespon memiliki bentuk kurva regresi yang tidak diketahui bentuk polanya, maka pendekatan yang digunakan adalah regresi nonparametrik birespon. Model untuk regresi nonparametrik birespon Spline dapat dituliskan sebagai berikut.

$$y_{1i} = \sum_{j=1}^{p} f(x_{ji}) + £_{1ii}$$
$$y_{2i} = \sum_{j=1}^{p} g(x_{ji}) + £_{2ii} \tag{12}$$

dengan fungsi f dan g adalah kurva regresi yang tidak diketahui bentuknya dan dihampiri dengan fungsi Spline sebagai berikut.

$$f(x_{ji}) = \sum_{h=1}^{p} a_{hj} x_{ji}^{h} + \sum_{l=1}^{m} {}_{lj} \left(t_{ji} - k_{lj}\right)_{+}^{p} \qquad (13)$$

dan

$$g(x_{ji}) = \sum_{h=1}^{p} \text{a}_{hj} x_{ji}^{h} + \sum_{l=1}^{m} \text{¤}_{lj} \left(t_{ji} - k_{lj}\right)_{+}^{p}$$

dengan a_{hj} dan ${}_{lj}$ merupakan untuk parameter variabel respon pertama sedangkan a_{hj} dan ${}_{lj}$ merupakan parameter variabel respon kedua.

Estimasi parameter bisa dicari dengan melakukan optimasi *Weighted Least Square* (WLS). Penentuan matrik pembobot $\boldsymbol{W}$ dalam kasus ini yaitu dengan perhitungan nilai varian kovarian dari respon pertama dan respon kedua. Adapun penyelesaian optimasi parameter dengan WLS didapat persamaan sebagai berikut.

$$\min\{\text{£}\,\text{Æ}\} = \left\{(y - X\ )' W (y - X\ )\right\}\text{p} \qquad (14)$$

persamaan 14 dapat diselesaikan dengan penurunan secara parsial dan memisalkan fungsi sebagai berikut.

$$\cdot(\ ) = (y - X\ )' W (y - X\ )$$

Selanjutnya persamaan yang diperoleh diturunkan terhadap sebagai berikut.

$$\frac{\partial \cdot(\ )}{\partial} \qquad (15)$$

Setelah dilakukan penurunan terhadap , hasil penurunan disamakan dengan nol. Sehingga bentuk estimasi model Spline dalam regresi nonparametrik birespon akan diperoleh sebagai berikut.

$$\hat{y} = X\ \hat{}$$
$$= X(X'WX)^{-1} X'Wy$$
$$= A(k)y$$

dengan

$$A(k) = X\left(X'WX\right)^{-1} X'W \qquad (16)$$

F. Korelasi antara Variabel - Variabel Respon

Sebelum merakukan pemodelan, terlebih dahulu harus diketahui besar hubungan atau korelasi antar variabel-variabel tersebut. Untuk mengetahui nilai korelasinya dapat digunakan koefisien korelasi Pearson yang secara umum memiliki persamaan sebagai berikut.

$$r\left(y_{1,} y_{2}\right) = \frac{\text{cov}\left(y_{1,} y_{2}\right)}{\left\{\text{var}\left(y_{1,}\right)\text{var}\left(y_{2}\right)\right\}^{\frac{1}{2}}} \qquad (17)$$

atau dapat dituliskan dengan rumus sebagai berikut:

$$r\left(y_{1,} y_{2}\right) = \frac{\text{cov}\left(y_{1,} y_{2}\right)}{\left\{\text{var}\left(y_{1,}\right)\text{var}\left(y_{2}\right)\right\}^{\frac{1}{2}}} \qquad (18)$$

dengan

$$\text{cov}\left(y_{1,} y_{2}\right) = \sum_{i=1}^{n} \left(y_{1i} - \bar{y}_{1}\right)\left(y_{2i} - \bar{y}_{2}\right)$$

$$\text{var}\left(y_{1}\right) = \sum_{i=1}^{n} \left(y_{1i} - \bar{y}_{1}\right)^{2}; \text{var}\left(y_{1}\right) = \sum_{i=1}^{n} \left(y_{2i} - \bar{y}_{2}\right)^{2}$$

. Nilai koefisien korelasi yang dihasilkan berdasarkan perhitungan dengan korelasi Pearson berkisar antara -1 sampai dengan 1. Apabila nilai koefisien korelasi mendekati -1 atau 1 maka hubungan atau korelasi antara variabel-variabel respon semakin kuat, sedangkan jika nilai koefisien korelasi mendekati 0 maka hubungan atau korelasi antara variabel-variabel respon semakin lemah [21].

G. Pemilihan Titik knot optimal

Untuk memperoleh Spline terbaik bergantung pada pemilihan titik-titik knot. Dalam Spline, titik knot merupakan perpaduan bersama antara perubahan fungsi pada interval yang berlainan. Pemilihan titik knot optimal dalam regresi Spline nonparametrik pada model-model koefisien bervariasi, tidak berbeda jauh dengan pemilihan titik knot pada regresi Spline nonparametrik pada umumnya.

Untuk tujuan memilih parameter optimal ini, telah dikembangan beberapa metode dalam regresi nonparametrik untuk data *cross section*, seperti [8] memberikan metode *Cross Validation* (CV), [7] memberikan metode *Unbiased Risk* (UBR), dan [9] memberikan suatu metode *Generalized Cross Validation* (GCV).

H. *Cross Validation* (CV)

Dasar dari metode CV adalah untuk memilih nilai $\mathbf{k}$ (knot) yang minimumkan $\text{CV}(\mathbf{k})$. Langkah awal dalam metode CV adalah memperhatikan bahwa nilai bergantung secara linear pada data dengan persamaan

$$\hat{y} = \mathbf{A}(\mathbf{k})y$$

dengan

$$A(k) = X\left(X'WX\right)^{-1} X'W$$

Hasil yang diperoleh dari pengembangan perhitungan nilai CV diberikan sebagai berikut.

$$\text{CV}(\mathbf{k}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{y}_i}{1 - \mathbf{A}_{ii}(\mathbf{k})}\right]^{2} \qquad (19)$$

Metode CV mengasumsikan bahwa fungsi diperoleh tanpa pengamatan ke-i dari data. Selanjutnya, estimator diperoleh dari suatu subsampel berukuran yang diambil dari data asli. Apabila penghapusan pengamatan ke-*i* ini dilakukan secara berulang-ulang, maka akan diperoleh suatu persamaan CV [10]. Metode CV biasa disebut sebagai metode hapus-satu, yaitu suatu metode yang bertujuan untuk meminimumkan jumlah kuadrat dari error prediksi untuk variabel respon, dimana prediktor untuk respon tersebut didasarkan pada estimator yang menggunakan seluruh data kecuali data [22].

I. Generalized Cross Validation (GCV)

Metode untuk memilih titik knot optimal salah satunya adalah dengan metode GCV [6]. GCV merupakan suatu bentuk modifikasi dari CV. Nilai GCV kemudian diperoleh, dengan menjumlahkan residual-residual kuadrat yang telah terkoreksi dengan kuadrat dari faktor-faktor ini. Karena faktor yang diperoleh bernilai sama untuk setiap *i*, maka diperoleh.

$$GCV(k) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\left[1 - n^{-1}trace\mathbf{A}(\mathbf{k})\right]^2} = \frac{MSE(k)}{\left[n^{-1}trace(\mathbf{I} - \mathbf{A}(\mathbf{k}))\right]^2} \quad (20)$$

dengan,

$$A(k) = X(k)(X'(k)WX(k))^{-1}X'(k)W$$

J. Kriteria Pemilihan Model Terbaik

Tujuan analisis regresi adalah mendapatkan model terbaik yang mampu menjelaskan hubungan antara variabel prediktor dan variabel respon berdasarkan kriteria tertentu. Kriteria yang sering digunakan adalah pemilihan model terbaik *Mean Square Error* (MSE). Nilai MSE adalah nilai taksiran dari varians residual, sehingga model regresi terbaik adalah model dengan MSE minimum. Koefisien determinasi adalah nilai dari proporsi keragaman total disekitar nilai tengah yang dijelaskan dari model regresi [21].

II. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Dinas Kesehatan Provinsi Bali tahun 2015 dan data dari Survei Sosial Ekonomi Nasional (SUSENAS) Provinsi Bali tahun 2015.

B. Variabel Penelitian

Tabel 1. Variabel Penelitian

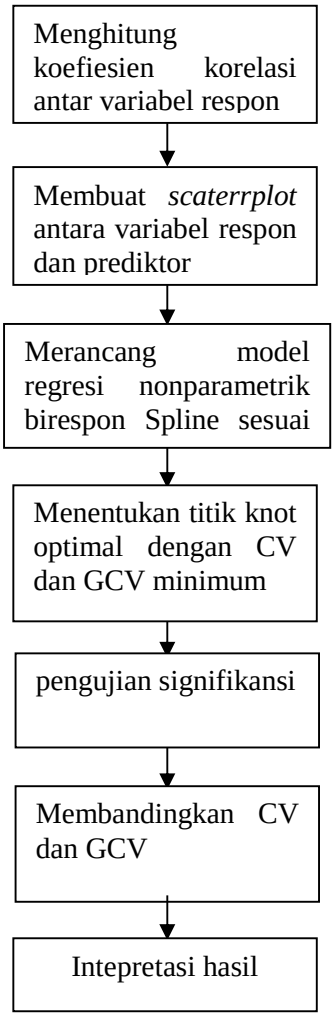
Variabel	Keterangan Variabel
Y_1	Angka Kematian Bayi (AKB)
Y_2	angka gizi buruk balita
X_1	persentase bayi yang tidak diberi ASI
X_2	Persentase bayi berat badan lahir rendah

C. Metode Penelitian

Metode penelitian yang digunakan untuk mencapai setiap tujuan penelitian dijabarkan sebagai berikut :

1. Mendeskripsikan AKB dan status gizi balita di Bali serta faktor – faktor yang diduga mempengaruhinya dengan membuat plot antara variabel prediktor dengan variabel respon
2. Melihat korelasi antara variabel respon yaitu AKB dan status gizi.
3. Memodelkan data dengan model regresi nonparametrik birespon Spline dengan satu titik knot, dua titik knot, dan tiga titik knot
4. Menghitung nilai CV dan GCV untuk masing-masing model regresi Spline.
5. Menentukan titik knot dan orde knot optimal berdasarkan nilai CV dan GCV minimum.
6. Melakukan pengujian signifikansi parameter yang dihasilkan dari estimasi model regresi nonparametrik Spline dengan metode CV dan GCV minimum.
7. Melakukan diagnostik residual yang dihasilkan dari estimasi model regresi nonparametrik Spline dengan metode CV dan GCV minimum.
8. Membandingkan nilai MSE estimasi model regresi nonparametrik Spline dengan titik knot optimal menggunakan metode CV dan GCV.
9. Menginterpretasikan hasil analisis dan mengambil kesimpulan.

Tahapan-tahapan di atas dilakukan secara otomatis oleh komputer dengan menggunakan *software* Statistika. Untuk mempermudah pemahaman alur analisis, maka langkah-langkah analisis dibuat dalam bentuk diagram alir seperti berikut:



Gambar 1. Tahapan Implementasi Model dan *Software* Terhadap Kasus

III. HASIL DAN PEMBAHASAN

A. Aplikasi Model Regresi Nonparametrik Birespon Spline pada data AKB dan Angka gizi buruk di Provinsi Bali

Sebelum memodelkan Derajat kesehatan di Provinsi Bali maka perlu dilihat deskripsi statistik dari data untuk masing masing variabel seperti pada table berikut ini. Statistik deskriptif yang ditampilkan digunakan dalam program terutama untuk inisialisasi titik knot.

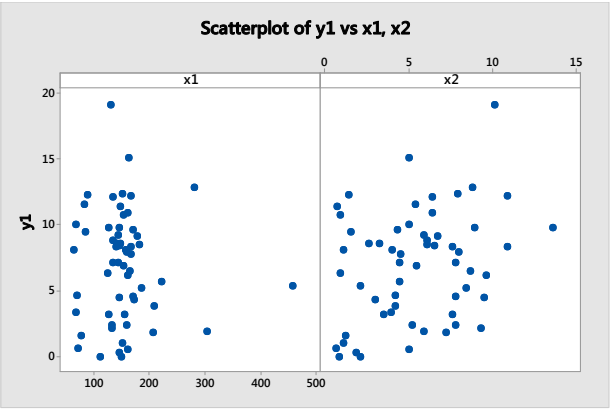
Tabel 2 Karakteristik AKB dan Angka Gizi Buruk Balita Faktor yang diduga Mempengaruhi

Variabel	Rata-rata	Varians i	Minimum	Maksimum
y_1	6.9	17.7	0.0	19.1
y_2	1.8	6.6	0.0	14.0
x_1	152.1	3774.1	63.9	458.4
x_2	5.4	9.9	0.6	13.6

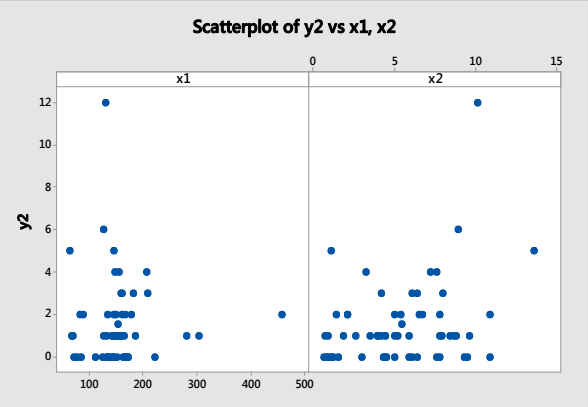
Tabel 2. menunjukkan karakteristik AKB dan angka gizi buruk balita yakni nilai rata – rata, variansi, minimum, dan maksimum. Berdasarkan Tabel 1 terlihat bahwa rata – rata AKB di Provinsi Bali yang tersebar pada 57 Kecamatan tiap 1000 penduduk adalah sebanyak 6 atau 7 dengan keragaman sebesar 17.7. Terlihat dari nilai minimum dan maksimum pada Tabel 1. mengindikasikan bahwa AKB di Provinsi Bali tertinggi sebesar 19.1% yaitu terletak di Kecamatan Kubu dan terendah sebesar 0% terletak di Kecamatan Denpasar Barat dan Kecamatan Kuta. Sementara itu, AKB di Provinsi Bali adalah sebesar 1.8% dengan keragaman sebesar 6.6.

Angka gizi buruk balita di Provinsi Bali tertinggi sebesar 14.0% terletak di Kecamatan Kubu dan terendah sebesar 0% terletak di Kecamatan Ubud, Tegallalang, Gianyar, Melaya, Mendoyo, Susut, Banjarangkan, Tabanan, Selemadeg, Selemadeg timur, Pupuan, Marga Petang, Kuta, Kuta Selatan, kuta Utara, Kubutambahan, Seririt, dan Manggis.

Penelitian ini bertujuan untuk membandingkan metode CV dan metode GCV untuk memilih titik knot optimal dalam regresi Spline birespon. Untuk melihat pola hubungan antar variabel dapat dilihat dari grafik scatter plot. Hasil sebaran data atau *scatter plot* untuk masing masing variabel respon dan variabel prediktor sebagai berikut:



Gambar 2. *Scatterplot* antara y_1 dengan x_1 dan x_2



Gambar 3. *Scatterplot* antara y_2 dengan x_1 dan x_2

Dilihat dari kedua gambar tersebut maka pemodelan yang tepat adalah memodelkan dengan regresi nonparametrik birespon dengan estimator yang digunakan adalah Spline. Dari variabel tersebut akan dibuat model dari model Spline linier dengan jumlah titik knot satu dan dua. Hasilnya kemudian dibandingkan nilai CV dan GCV terkecil diantara model yang terbentuk. Pemilihan titik knot pada model dilakukan secara otomatis oleh program komputer

B. Regresi Nonparametrik Birespon Spline pada Data AKB dan Angka status gizi Buruk Balita

Dalam memodelkan data dengan nonparametrik birespon Spline, langkah awal yang dilakukan adalah menentukan titik knot optimal yang berkaitan dengan nilai CV dan GCV terkecil. Tabel-tabel berikut menunjukkan nilai CV dan nilai GCV pada data AKB dan Angka Gizi Buruk Balita Tahun 2015 dengan titik knot satu dan dua diperoleh dengan menggunakan bantuan software Matlab.

Bentuk umum birespon Spline linier dua variabel prediktor dengan k titik knot adalah:

$$\hat{y}_1 = \sim_0^{(1)} + \sim_{11}^{(1)} x_1 + \ {}_{11}^{(1)}(x_1 - K_{1_1}^{(1)})_+ + \ldots + \ {}_{k1}^{(1)}(x_1 - K_{k_1}^{(1)})_+$$
$$+ \sim_{12}^{(1)} x_1 + \ {}_{12}^{(1)}(x_2 - K_{1_2}^{(1)})_+ + \ldots + \ {}_{k2}^{(1)}(x_2 - K_{k_2}^{(1)})_+$$
$$\hat{y}_2 = \sim_0^{(2)} + \sim_{11}^{(2)} x_1 + \ {}_{11}^{(2)}(x_1 - K_{1_1}^{(2)})_+ + \ldots + \ {}_{k1}^{(2)}(x_{1j} - K_{k_1}^{(2)})_{+1}$$
$$+ \sim_{12}^{(2)} x_1 + \ {}_{12}^{(2)}(x_2 - K_{1_2}^{(2)})_+ + \ldots + \ {}_{k2}^{(2)}(x_2 - K_{k_2}^{(2)})_+$$

C. Pemilihan Titik Knot Optimal dengan metode GCV

Pada hasil pengolahan data derajat kesehatan di Indonesia dengan dua variabel prediktor yaitu AKB dan angka status gizi buruk balita dan menggunakan satu dan dua titik knot untuk masing-masing variabel prediktor didapatkan nilai GCV minimum untuk masing masing model sebagai berikut:

Tabel 3. Nilai GCV Minimum dan MSE Masing-masing Titik Knot

Variabel Prediktor	GCV Minimum	MSE
1 Titik knot	78.7909	77.4146
2 Titik knot	82.3998	80.9605

Tabel 3 menunjukkan bahwa nilai GCV minimum dihasilkan pada saat menggunakan knot satu yakni sebesar 78.7909 dengan MSE sebesar 77.4146. Sehingga, titik knot optimal yaitu terletak pada GCV satu titik knot . Nilai titik knot untuk masing masing variabel prediktor pada masing-masing respon seperti pada tabel berikut ini.

Tabel 4. Nilai Titik Knot untuk masing-masing variabel

Model	Variable rediktor	Titik knot Respon 1	Titik knot Respon 2
Linier 1 Knot	x_1	176.61	176.61
	x_2	0.60	0.60
Linier 2 Knot	x_1	63.90 195.40	195.40 326.90
	x_2	0.60 4.93	0.60 4.93

Hasil estimasi parameter dari model terbaik pada regresi Spline linier dengan satu titik knot adalah sebagai berikut:

Tabel 5. Estimasi Parameter Model Spline Linier Multirespon dengan 1 titik knot

Estimasi	Respon 1	Respon 2
$\sim_0$	0.00045	0.00072
$\sim_1$	0.04816	0.08016
$_{+1}$	-0.01067	-0.02090
$\sim_2$	0.00292	0.00536
$_2$	0.00264	0.00493

Sehingga estimasi model Spline linier birespon dengan satu titik knot dapat ditulis keadalam bentuk persamaan sebagai berikut:

Untuk variabel respon pertama yaitu AKB sebagai berikut.

$$\hat{y}_1 = 0.00045 + 0.04816x_1 - 0.01067(x_1 - 176.61)_+ + 0.00292x_2 + 0.00264(x_2 - 0.60)_+$$

Untuk variabel respon kedua angka gizi buruk balita sebagai berikut.

$$\hat{y}_2 = 0.00072 + 0.08016x_1 - 0.02090(x_1 - 176.61)_+ + 0.00536x_2 + 0.00493(x_2 - 0.60)_+$$

D. Pemilihan Titik Knot Optimal dengan Metode CV

Pada hasil pengolahan data derajat kesehatan di Indonesia dengan dua variabel prediktor yaitu AKB dan angka status gizi buruk balita dan menggunakan satu dan dua titik knot untuk masing-masing variabel prediktor didapatkan nilai CV minimum untuk masing masing model sebagai berikut:

Tabel 6. Nilai CV Minimum dan MSE Masing-masing Titik Knot

Variabel Prediktor	CV Minimum	MSE
1 Titik knot	77.37831	76.75449
2 Titik knot	80.62915	80.96054

Tabel 6 menunjukkan bahwa nilai CV minimum dihasilkan pada saat menggunakan knot satu yakni sebesar 77.37831 dengan MSE sebesar 76.75449. Sehingga, titik knot optimal yaitu terletak pada CV satu titik knot.

Nilai titik knot untuk masing masing variabel prediktor pada masing-masing respon seperti pada tabel berikut ini.

Tabel 7. Nilai Titik Knot untuk masing masing variabel

model	Variable rediktor	Titik knot Respon 1	Titik knot Respon 2
Linier 1 Knot	x_1	162.53	162.53
	x_2	0.60	0.60
Linier 2 Knot	x_1	63.90 195.40	195.40 326.90
	x_2	0.60 4.93	0.60 4.93

Hasil estimasi parameter dari model terbaik pada regresi Spline linier dengan satu titik knot adalah sebagai berikut:

Tabel 8. Estimasi Parameter Model Spline Linier Birespon dengan 1 titik knot

Estimasi	Respon 1	Respon 2
$\tilde{}_0$	0.00046	0.00072
$\tilde{}_1$	0.04825	0.08030
$_1$	-0.01160	-0.02003
$\tilde{}_2$	0.00292	0.00537
$_2$	0.00265	0.00494

Sehingga estimasi model Spline linier birespon dengan 1 titik knot dapat ditulis keadalam bentuk persamaan sebagai berikut:

Untuk variabel respon pertama yaitu AKB sebagai berikut.

$$\hat{y}_1 = 0.00046 + 0.04825x_1 - 0.01160(x_1 - 162.53)_+ + 0.00292x_2 + 0.00265(x_2 - 0.60)_+$$

Untuk variabel respon kedua angka gizi buruk balita sebagai berikut.

$$\hat{y}_2 = 0.00072 + 0.08030x_1 - 0.02003(x_1 - 162.53)_+ + 0.00537x_2 + 0.00494(x_2 - 0.60)_+$$

D. Perbandingan Titik Knot Optimal dengan metode GCV dan CV

Tabel 9. Nilai CV Minimum dan MSE Masing-masing Titik Knot

Variabel Prediktor	GCV Minimum	MSE
1 Titik knot	78.7909	77.4146
2 Titik knot	82.3998	80.9605
Variabel Prediktor	CV Minimum	MSE
1 Titik knot	77.37831	76.75449
2 Titik knot	80.62915	80.96054

Dari nilai CV dan GCV minimum tersebut maka model terbaik yang sesuai untuk derajat kesehatan di Indonesia pada tahun 2015 yaitu dengan menggunakan metode CV dengan model linier satu titik knot.

Dari model dengan menggunakan metode CV maka model derajat kesehatan di Indonesia dapat diintepretasikan kedalam regresi nonparametrik birespon yaitu:

1. Model yang terbaik yang menjelaskan derajat kesehatan yaitu dengan variabel respon AKB dan angka gizi buruk balita di Indonesia adalah model Spline linier dengan satu titik knot.

2. Pada respon AKB dan respon angka gizi buruk balita, perubahan pola perilaku data pada variabel persentase bayi diberikan ASI terjadi pada titik 162.53, dimana jika nilai variabel tersebut dibawah 162.53persen maka persentase AKB dan angka gizi buruk balita memiliki pola yang berbeda dengan setelah persentase bayi diberikan ASI bernilai 162.53 persen dan lebih.
3. Pada respon AKB dan respon angka gizi buruk balita, perubahan pola perilaku data pada variabel BBLR terjadi pada titik 0.60, yaitu pola persentase AKB dan angka gizi buruk balita berubah setelah titik 0.60.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil dan pembahasan dapat diambil kesimpulan sebagai berikut:

1. Model yang terbaik yang dapat menggambarkan tingkat derajat kesehatan di Indonesia yaitu AKB dan angka gizi buruk balita adalah dengan model Spline linier satu titik knot menggunakan metode CV. Model yang terbentuk yaitu:

Untuk variabel respon pertama yaitu AKB sebagai berikut.

$$\hat{y}_1 = 0.00046 + 0.04825x_1 - 0.01160(x_1 - 162.53)_+ + 0.00292x_2 + 0.00265(x_2 - 0.60)_+$$

Untuk variabel respon kedua angka gizi buruk balita sebagai berikut.

$$\hat{y}_2 = 0.00072 + 0.08030x_1 - 0.02003(x_1 - 162.53)_+ + 0.00537x_2 + 0.00494(x_2 - 0.60)_+$$

Dengan, nilai CV : 77.37831dan nilai MSE sebesar: 76.75449.

B. Saran

Penelitian ini terbatas pada penggunaan regresi birespon Spline linier. Untuk penelitian selanjutnya dapat dikembangkan dengan regresi kuadratik dan kubik. Saran yang dapat diberikan dalam penelitian ini adalah perlu dilakukan penelitian lebih lanjut untuk menentukan uji statistik dari model yang terbentuk misalnya uji hipotesis dan interval konfidensi serta dalam memodelkan derajat kesehatan di Indonesia, penulis hanya menggunakan dua variabel respon dan dua variabel prediktor, dengan jumlah knot terbanyak hanya dua titik knot sedangkan masih banyak respon dan prediktor yang mungkin mempengaruhi derajat kesehatan di Indonesia

V. DAFTAR PUSTAKA

[1] Dinkes Bali (Dinas Kesehatan Provinsi Bali). (2015). *Profil Kesehatan Provinsi Bali Tahun 2015.* Bali: Pemerintah Provinsi Bali.

[2] Depkes RI (Departemen Kesehatan RI). (2003). Indikator Indonesia Sehat 2010 dan Pedoman Penetapan Indikator Provinsi Sehat dan Kabupaten/Kota Sehat: Keputusan Menteri Kesehatan Nomor 1202/Menkes/SK/VIII/2003. Departemen Kesehatan RI, Jakarta.

[3] Budiantara, I. N. (2007). *Pendugaan Model Fertilitas Wanita di Indonesia dengan Menggunakan Regresi Spline*, Laporan Akhir Pelaksanaan Penelitian Studi Kajian Wanita Tahun Anggaran 2007, LPPM, Intitut Teknologi Sepuluh Nopember, Surabaya

[4] Budiantara, I. N. (2009). Spline dalam Regresi Nonparametrik dan Semiparametrik: Sebuah Pemodelan Statistika Masa Kini dan Masa Mendatang, ITS Press, Institut Teknologi Sepuluh Nopember Surabaya.

[5] Tripena, Agustini. (2011). *Penentuan Model Regresi Spline Terbaik*, Prosiding Seminar Nasional Statistika Universitas Diponegoro 2011, Semarang.

[6] Budiantara, I. N. (2000). Metode U, GML, CV dan GCV dalam Regresi Nonparametrik Spline, *Majalah Ilmiah Himpunan Matematika Indonesia (MIHMI)*, 6, 41-45.

[7] Wang, Y. (1998). "Spline Smoothing Models with Correlated Error". Journal of the Royal Statistical Sociaty, Series B, 5r0, 341-348.

[8] Craven, P., and Wahba, G. (1979). "Smoothing Noisy Data With Spline Functions". Numerische Mathematik, 31, 377-403

[9] Wahba, G. (1990). "Spline Model for Observational Data}". SIAM, XII, Philadelphia.

[10] Eubank, R. L. (1988). "Spline Smoothing and Nonparametric Regression". New York: Marcel Dekker.

[11] Fitriyani, Nurul. (2014). Metode Cross Validation Dan Generalized Cross Validation Dalam Regresi Nonparametrik Spline Studi Kasus Data Fertilitas Di Jawa Timur. Tesis, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember (ITS), Surabaya

[12] Kemenkes Sumut (Kementerian Kesehatan Pemerintahan Provinsi Sumatera Barat). (2014). *Profil Kesehatan Provinsi Sumatera Barat 2014*. Sumatera Barat: Kementerian Kesehatan Pemerintahan Provinsi Sumatera Barat.
[13] Dinkes Jateng (Dinas Kesehatan Provinsi Jawa Tengah). (2015). *Profil Kesehatan Provinsi Jawa Tengah Tahun 2015*. Jawa Tengah: Pemerintah Provinsi Jawa Tengah.
[14] Santoso, Soegoeng dan Anne Lies Ranti. (2009). Kesehatan dan Gizi, Jakarta, Rineka Cipta
[15] Aulia, Annisya (2017). Faktor-faktor yang Mempengaruhi Derajat Kesehatan dengan Menggunakan Regresi Multivariat (Studi kasus : Derajat Kesehatan Kabupaten dan Kota di Provinsi Sumatera Barat). Diploma thesis, Universitas Andalas.
[16] Wulandari, I. D. A. M. I. (2014). Analisis Faktor – Faktor Yang Mempengaruhi Persentase Penduduk Miskin Dan Pengeluaran Perkapita Makanan Di Jawa Timur Dengan Metode Regresi Nonparametrik Birespon Spline. Tugas Akhir, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember (ITS), Surabaya.
[17] Setyawan, N.A.D. (2011). *Pendekatan Regresi Nonparametrik Birespon Spline untuk Pemodelan Determinan Tingkat Pendidikan di Pulau Papua, Thesis*, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya.
[18] Eubank, R. L. (1999). "Nonparametric Regression and Spline Smoothing Second Edition". New York: Marcel Dekker.
[19] Budiantara, I. N. (2004). *Model Spline Multivariabel dalam Regresi Nonparametrik*, Makalah Seminar Nasional Matematika, Jurusan Matematika ITS Surabaya.
[20] Similia,T. Dan Tikka, J. (2007). Input Selection and Shrinkage in Multiresponse Linear Regression, *Preprint Submitted to Elsevier*.
[21] Draper, N.R. and Smith, H. (1998). "Applied Regression Analysis,Three Edition". John Wiley and sons, Inc. New York.
[22] Andrews, Y. Ng. (1991). "Preventing "Overfitting" of Cross-Validation Data". School of Computer Science, Carnegie Mellon University, Pittsburgh PA.